CEH contract report
to the
Department of the Environment, Food and Rural Affairs

CS2000 Module 9
DATA INTEGRATION FOR LOCALISED RESULTS AND SUPPORT FOR INDICATORS OF COUNTRYSIDE CHARACTER AND QUALITY
Reports A4 – February 2004

R.A. Hill & G.M. Smith,

Centre for Ecology and Hydrology
Monks Wood
Abbots Ripton
Huntingdon PE28 2LS

CONTRACT
No. CR0263

Executive summary

This is the third report of Countryside Survey Module 9 'Data Integration for Localised Results and Support for Indicators of Countryside Character and Quality' which addresses 9A project activities.

All activities from the Feasibility phase have been completed except the production of estimates for the English Region which will await finalisation of the calibrated 1 km data set in the operational phase.

The project is now approaching the end of the Operational phase and the final outputs are being prepared. A prototype calibrated 1 km dataset was presented at a number of meetings and was distributed to all the participants via email in a format compatible with Countryside Information System (CIS) version 7.

Through the process of meetings, presentations and email distribution the procedures developed and results produced by Module 9A have been communicated to a wide range of user groups. These procedures have in general been accepted and deficiencies identified for further considerations. This process has resulted in a clearer understanding of the issues associated with LCM2000, FS and their integration.

The slow feedback has delayed the work on the refinement of the methodology and production of estimates of uncertainty within the data set. Nevertheless, a certain amount of refocusing of the project has occurred in response to feedback we have received from the prototype calibrated 1 km data set.

Ancillary data have been identified as a possible means of correcting deficiencies in the calibrated 1 km data set and validating the final results. A range of data sets have been obtained and extensive comparisons undertaken with FS and LCM2000. These ancillary datasets have generated their own problems and decisions have been made on quality, definition and coverage along with suitability for a role within Module 9A.

Datasets useful for additional knowledge-based correction within the calibration process have been identified and suitable rule bases developed and tested. This has resulted in two further iterations of the calibrated 1 km data set to a point where it is in its 'final' form with respect to this project.

A final iteration may be carried out to improve the quality of the calibrated calcareous grasslands if new English Nature 'Priority Habitats inventory' data can be acquired prior to public release in April 2004.

A full validation will now be undertaken using a number of the ancillary datasets which are described. Uncertainty information will be derived for the calibrated 1 km data set by implementation of the full bootstrapping procedure.

A strategy for the measurement of landscape pattern has been developed and these values will now be derived for England at a 1 km level and the East Midland on a parcel by parcel basis.

Table of contents

Introduction

This is the third report of Countryside Survey Module 9 'Data Integration for Localised Results and Support for Indicators of Countryside Character and Quality' which addresses 9A project activities.

The remit of Module 9A is 'To determine how data from Countryside Survey 2000 Field Survey (FS) and Land Cover Map 2000 (LCM2000) can be integrated to produce consistent and robust estimates of stock and change at different scales.' The project was designed as follows:

Feasibility phase

1. Develop and test integration methods
2. Demonstrate methodology for England
3. Produce provisional estimates for English Regions, Wales and Scotland
4. Produce prototype calibrated 1 km data set for England

Operational phase

5. Refine methodology based on user feedback
6. Quantify and explain sources of error
7. Develop and evaluate measure of landscape pattern

All activities from the Feasibility phase have been completed except the production of estimates for the English Region which will await finalisation of the calibrated 1 km data set in the operational phase.

The project is now approaching the end of the Operational phase and the final outputs are being prepared. A certain amount of refocusing has been forced on the project due to a lack of feedback we have received from the prototype calibrated 1 km data set. This dataset was presented at a number of meetings and was distributed to all the participants via email in a format compatible with Countryside Information System (CIS) version 7.

The slow feedback has delayed the work on the refinement of the methodology and production of estimates of uncertainty within the data set. A point has now been reached where the work must progress to meet deadlines and most of these developments are reported here. Ancillary data sets have been explored to explain sources of errors and a strategy for the measurement for landscape pattern has been devised.

The production of a calibrated 1 km data set and regional estimates for Scotland and Wales was awaiting the point when these countries joined the CS2000 Module 9 group, but this now seems unlikely.

Assessment of additional datasets for calibration enhancement

The LCM2000 and the FS each have their strengths and weaknesses in generating land cover statistics, resulting from their methods of production. The process of integrating these two datasets to create 'best estimate' land cover statistics needs to draw together the individual strengths of each dataset, whilst reducing the weaknesses. The chosen spatial framework for this is the National Land Classes (NLC), with summary land cover data to be generated for each 1 km National Grid cell in England.

The starting point to achieve a 'best fit' integration of LCM2000 and FS for the supply of land cover statistics is their comparison with consistent nationwide datasets of known quality. There are few such relevant datasets that are readily available for England with a 1km or finer spatial resolution:

- National Soil Resources Institute (NSRI) National Soil Map for England and Wales (NATMAP1000),
- English Nature (EN) Grassland Inventory of England,
- Ordnance Survey (OS) 1 km Geographic Reference Data for GB,
- Forestry Commission (FC) Digital Woodland Map for England.

In addition to these 1 km spatial resolution datasets on land cover (or in the case of NATMAP1000, a key environmental variable that can directly influence land cover) CEH also has access to a 10 km spatial resolution GB dataset on indicator species composition.

These datasets can potentially contribute to three important elements of the integration process: (i) providing an independent comparison for LCM2000 and FS estimates of stock, thereby identifying the strengths and weaknesses of each survey; (ii) providing an input to the calibration procedure where the bootstrapped calibration matrices per NLC are not sufficient to generate an accurate output; (iii) providing a means of validating the integrated 1 km summary dataset, if not used in the calibration procedure.

The applicability of each of the above datasets as a means of identifying error in LCM2000 and FS, as a possible input to the calibration procedure and as a means of validating the calibrated output has been examined.

Soils

The NSRI NATMAP1000 is a 1 km raster data set with information on the dominant soil group per 1 km National Grid cell of England and Wales. The soil classification scheme is hierarchical, with 10 major soil groups which expand into 34 different soil groups (in total describing the composition and distribution of 300 soil associations). The current version of NATMAP1000 was launched in 2001 and more information is available from www.silsoe.cranfield.ac.uk/nsri.

The 34 different soil groups identified in NATMAP1000 were classified into calcareous, neutral, acid and peat soil types by Mark Hill at CEH Monks Wood. This resulted in a map of the dominant soil type per National Grid 1 km cell, against which

FS and LCM2000 estimates of Broad Habitats strongly influenced by soil base levels could be assessed. This includes the semi-natural grasslands, dwarf shrub heath and bog, which can be distinguished much more readily by field surveyors than by the spectral reflectance characteristics in satellite imagery. LCM2000 made use of soil sensitivity and peat drift maps to distinguish between these different classes during post image classification knowledge-based correction procedures. The deficiencies of the soils sensitivity and peat drift data for this purpose were recognised at the time of LCM2000 production, but nonetheless they were the only datasets available. This is borne out by the differences in spatial coverage of calcareous, neutral and acid conditions in England derived from soil type and soil sensitivity respectively:

	% by soil type	% by soil sensitivity
calcareous	11	74
neutral	59	15
acid	26	11

The spatial correspondence between these three classes is 15% for calcareous, 39% for neutral and 51% for acid (averaging just 23%). Note that according to the soil type classification, 3% of England is peat, in comparison with 6% based on the peat drift map. The spatial correspondence between those two data sets is 51%.

Comparison of soils data with FS and LCM2000 land cover estimates for England & Wales

NATMAP1000 offers a dataset for assessing the environmental conditions associated with distinguishing different semi-natural grasslands, or dwarf shrub heath from bog. As the FS estimates of these land cover types were not influenced by the vagaries of ancillary datasets, then in theory they should show a strong correlation with the soil type data. The LCM2000 statistics, by comparison, may be expected to show less of a correlation. Of course, the percentage cover of a particular land cover type in any particular NLC should not equate with the percent cover of its corresponding soil type. Thus for example, an NLC with 70% calcareous soils is unlikely to have 70% cover of calcareous grassland. This is because the dominant land cover in virtually all NLCs is either arable or improved grass. However, in an NLC with 70% calcareous soils it is reasonable to assume that approximately 70% of the semi-natural grassland present will be calcareous in nature. The comparisons shown below are calculated for the 24 NLCs of England and Wales.

Figure 1 plots calcareous, neutral and acid grassland, each as a percentage of the total semi-natural grassland, against the percent cover of the appropriate soil type per NLC for both FS and LCM2000. Perhaps the most notable distinction between the estimates of FS and LCM2000 is for neutral grassland. The FS estimates of cover per NLC form a tight scatter about the $y = x$ line, whilst the LCM2000 estimates bear no relationship at all to percentage cover of neutral soils. Thus, the proportion of semi-natural grassland identified as neutral in FS correlates strongly with the proportion of neutral soils per NLC. This is not the case with LCM2000 estimates. For acid grassland, the scatter of points about the $y = x$ line in the FS dataset is greater than for neutral grassland. Nonetheless, this is notably better than in the LCM2000 data, which has a general (but not exclusive) tendency to under-estimate acid grassland cover as a proportion of semi-natural grassland in relation to the proportion of acid soils.

Calcareous grassland land cover estimates do not correlate particularly well with percent calcareous soil cover per NLC in either FS or LCM2000. In the case of FS estimates, this probably reflects the comparative rarity of calcareous grassland in relation to the sampling density in each National Land Class. LCM2000 clearly over-estimates the likely proportion of semi-natural grassland that is calcareous in every NLC given the percent cover of calcareous soil. This reflects the fact that the soil sensitivity map used in assigning semi-natural grassland into a calcareous class was too generous.

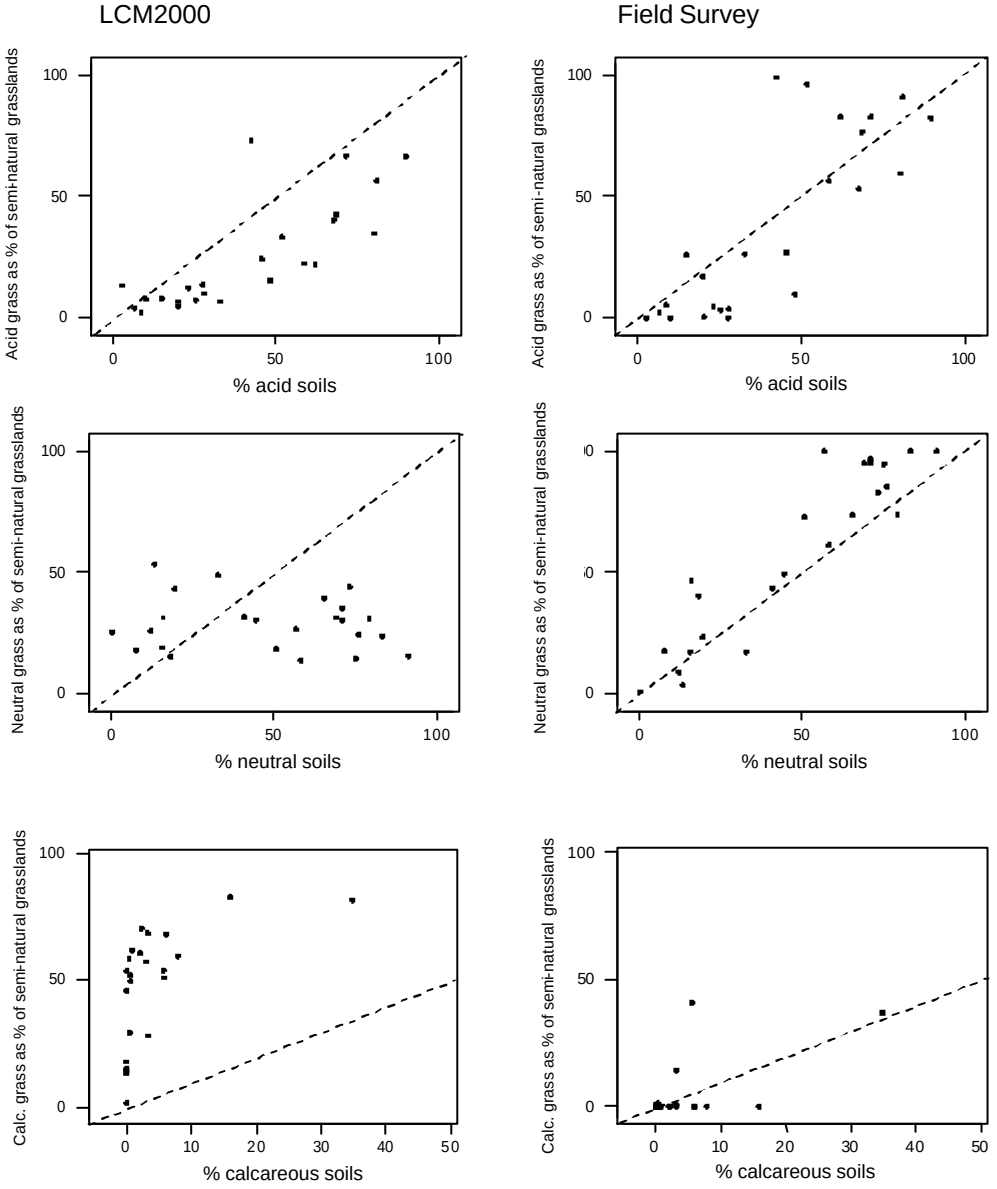


Figure 1. Scatter plots of calcareous, neutral and acid grassland as a percentage of the total semi-natural grassland against the percent cover of the appropriate soil type per NLC for LCM2000 and FS. (The $y = x$ line is also plotted for each graph).

The top two plots of Figure 2 show the cover of BH10 (dwarf shrub heath) as a percentage of BH 10 + BH 12 (bog) plotted against the percent cover of acid soils per NLC for FS and LCM2000. Neither the FS nor the LCM2000 data show a significant relationship with the proportion of acid soils per NLC. The LCM2000 statistics show

an apparent over-estimation in all but one NLC of England and Wales, whereas the FS statistics show both over- and under-estimation in relation to acid soil coverage. In the case of LCM2000, this reflects the already known fact that the knowledge-based correction procedure transferred too much land cover from the bog category into the most likely alternative categories (dwarf shrub heath and acid grassland). In the case of the FS statistics, the wide scatter of points about the $y = x$ line results from several causes. Firstly, for NLCs with a low percentage cover of acid soils, the remaining soil coverage is much more likely to be of a neutral type than peat. Thus, as occurs on the plot for FS, seven NLCs have less than 50 % acid soil and yet BH 10 comprises close to 100 % of the BH 10 + BH 12 total. Also confusing the picture is the fact that an NLC with high acid soil type coverage may be expected to have a high coverage of dwarf shrub heath, or acid grassland, or both. Also, it has to be borne in mind that in most NLCs both dwarf shrub heath and bog are rare land cover types; for example 16 of the 24 NLCs for England and Wales have a combined cover of BH 10 + BH 12 of less than 2 % of the total area. The FS may thus not have a high enough sampling density per NLC to map the percentage cover of dwarf shrub heath or bog adequately in many cases.

Figure 2 bottom row shows the percentage cover of bog plotted against the percent cover of peat soils per NLC for FS and LCM2000. Because the peat soils identified in the soil type map are located almost exclusively in semi-natural areas, it is reasonable to expect a direct relationship between percentage peat soils and percentage bog per NLC. However, both FS and LCM2000 estimates of bog coverage are low for the two NLCs with >30% peat soils.

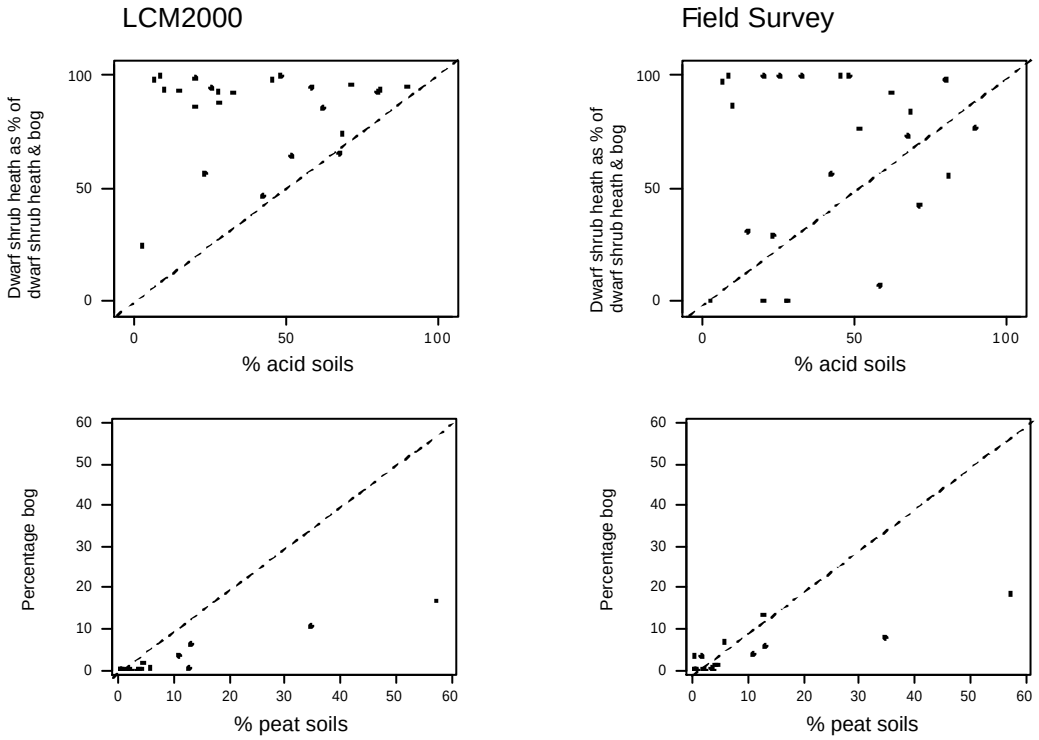


Figure 2. Top row: the cover of dwarf shrub heath as a percentage of BH 10 + BH 12 plotted against the percent cover of acid soils per NLC for LCM2000 and FS. Bottom row: the percentage cover of bog plotted against the percent cover of peat soils per NLC for FS and LCM2000. (The $y = x$ line is also plotted for each graph).

Grassland inventory

The English Nature (EN) Grassland Inventory of England is a vector dataset of semi-natural lowland grass communities (i.e. sites of enclosed grassland occurring at or below 300 m above sea level). The inventory is based on a range of sources, but the Phase 2 level surveys form the bulk of the data used, with sites mapped at the 1:50 000 level. Other data sources include surveys carried out by organisations such as Wildlife Trusts and Local Authorities. Both statutorily designated sites (e.g. National Nature Reserves and Sites of Special Scientific Interest) and undesignated sites are included within the inventory. Grassland sites were included according to the following criteria: high botanical diversity; post-1980 survey data; minimum size of 0.5 ha (100 m length for linear features); and information which is easily located and held as a readily accessible record. Only semi-natural communities were included in the inventory: neutral, calcareous and acid grassland, Calaminarian grassland (metallophyte vegetation), fen meadows and rush pastures, and selected swamp mire and mire communities. Maritime cliff grassland, salt marsh and sand dune grasslands were mostly excluded from the definition, as were improved or semi-improved grassland communities. These data were collected from sources spanning a 16 year timescale and consequently some sites in the inventory will have been lost to agricultural improvement or development.

The EN Grassland Inventory is not a complete measure of the extent and distribution of lowland grassland communities in England since the coverage of the Phase 2 grassland surveys were not comprehensive and some grassland types (e.g. calcareous grassland) have better coverage than others (e.g. acid grassland). Additionally, of the 8109 parcels making up the dataset, 481 have no information on grassland types present (i.e. are 'empty polygons' in which grassland type was not designated). This includes substantial areas of grassland at Salisbury Plain, Teesdale and the Brecklands. Furthermore, 25 % of the parcels list more than one grassland type, but the proportional composition is not stated. The EN Grassland Inventory could thus provide a dataset for comparing or validating the land cover composition of sample areas from LCM2000 or FS for which overlapping data exist. However, since the EN Grassland Inventory is not a complete census, has unlabelled parcels and the sampling strategy was not systematic, then this dataset is not particularly useful for comparing the land cover statistics per National Land Class or as input to the calibration procedure.

It should be noted that the Grassland Inventory will soon be superseded by Biodiversity Action Plan (BAP) Priority Habitat inventories. In total, 23 BAP Priority Habitat inventories will be available for England, which includes seven grassland habitats. These data are currently being quality assessed by EN before release in April 2004 via the website natureonthemap.org.uk. The dataset will still contain no information as to the location or proportional coverage of each Priority Habitat within the boundaries of designated areas. Nonetheless, compared with the Grassland Inventory, the dataset will be more of an exhaustive survey and will contain habitat labels for all parcels. Thus, if this dataset can be obtained from EN prior to public release in April, then the parcel information could be simplified to recorded presence or absence per 1 km National Grid square and used as a spatial mask in the calibration procedure for semi-natural grasslands.

Ordnance Survey 1km summary data

Ordnance Survey (OS) 1 km summary Geographic Reference Data for GB is available through the Countryside Information System (CIS) and lists the percent cover for each 1 km National Grid cell of GB for a range of features. Of relevance to the widespread Broad Habitat reporting framework of CS2000 are the classes: built up towns, built up villages, canals, inland water, rivers, foreshore, sand, and woodland.

The woodland data can provide a comparison with BH 1 and BH 2 combined into a generic woodland class for LCM2000 and FS. In addition to woodland, the OS Geographic Reference Data can easily provide a 1 km summary comparison for BH 13 by combining canals and inland water, BH 17 by combining built up villages and towns, and a combined 'coastal' habitat (BHs 18-21) by combining foreshore and sand. The OS Geographic Reference Data is dated as 1998.

Comparison of 1km OS98 data with FS and LCM2000 land cover estimates for England & Wales

In contrast to the comparisons made using the soils data which compared proportional coverage per NLC, with the OS98 dataset it is possible to make direct comparisons of land cover estimates. However, it has to be noted that the OS98 and LCM2000 data on land cover are both a complete census, whereas the FS land cover estimates are based on the proportional composition of samples within a NLC multiplied by the area of each NLC. However, even this does not supply land cover statistics for the same spatial coverage as LCM2000 and OS98 since:

- the NLCs do not extent as far off shore as the LCM2000 and OS98 datasets,
- the FS statistics are not intended to be projected into National Grid cells with greater than 75% urban coverage.

Thus, the FS statistics are not extrapolated to cover as large a spatial area as LCM2000 and OS98. This has obvious implications for the FS stock estimates for coastal and built up land covers. This is borne out in Table 1, which shows land cover statistics derived from all three datasets for England.

Table 1. Total coverage of land cover types (in km^2) in England as reported by FS and LCM2000, compared with OS98 data.

	***FS**	**LCM2000**	**OS98**
Woodland	12 950	13 910	6 500
Inland water	890	590	500
Coastal	1 560	1 900	2 000
Built up	10 420	13 800	13 700

* FS has a restricted spatial coverage compared with LCM2000 and OS98

FS clearly under-estimates the total area of coastal and built up land cover in England compared with OS data (by 440 km^2 and 3280 km^2 respectively). LCM2000, by contrast estimates these land covers to within 100 km^2 of the OS estimates. However, we can calculate from the urban mask used in the CIS that 5 038 1 km National Grid cells which have an urban coverage of greater than 75 % are absent from the FS estimate of built up. As these 1 km cells must have an urban composition of between

75 % and 100 %, we can calculate the total urban estimate of England for FS to be in the range of 14 200 km^2 to 15 458 km^2. This is an over-estimate compared with the OS98 data (of 500 km^2 to 1758 km^2). This probably relates to the Field Survey method of placing a boundary around any built area and assigning a land cover of 'built up' irrespective of the actual variations in land cover within that built up boundary.

For inland water, the LCM2000 estimate of the total coverage for England is once again within 100 km^2 of the OS estimate. The FS estimate is too high by 390 km^2 compared with the OS data. The implication here is that the sampling density of FS squares in England is not sufficient to represent adequately the spatial coverage of inland water bodies.

For woodland, both LCM2000 and FS statistics over-estimate the total coverage for England compared with the OS data (by 7 410 km^2 and 6 450 km^2 respectively). The LCM2000 estimate of woodland for England is for a larger spatial coverage than the FS estimate, since the LCM2000 statistic also includes woodland mapped in an urban context. This would account for some, but presumably not all, of the 960 km^2 greater estimate of woodland by LCM2000 than by FS.

A more directly comparable analysis is to extract land cover statistics from the OS98, FS and LCM2000 datasets in the CIS using the NLC boundaries and urban mask. This reduces the spatial coverage of all three datasets to the lowest common denominator (i.e. the lower spatial coverage for the projected FS data, which has urban and coastal data gaps compared with the complete census coverage of the OS98 and LCM2000 datasets). For ease of data extraction in the CIS, this was carried out for the total area of NLCs 1-24, which cover England and Wales.

Table 2. Total coverage of land cover types (in km^2) in England and Wales for National Land Classes 1-24 with an urban mask applied as reported in CIS by FS and LCM2000, compared with OS98 data.

	FS	**LCM2000**	**OS98**
[#] Woodland	15 508	16 730	8 437
[#] Inland water	1 054	669	548
[#*] Coastal	1 856	1 614	1 276
[#] Built up	11 801	10 890	10 560

[#] Excludes National Grid cells with > 75% urban coverage

* Coastal includes estuary and inshore data only

The land cover statistics for England and Wales in Table 2 are directly comparable for FS, LCM2000 and OS98 since they are estimates for the same spatial area. For inland water, coastal and built up land cover types, the stock estimates for LCM2000 provide a closer match with OS98 than the FS estimates. All three land cover types are over-estimated by both LCM2000 and FS. For LCM2000, the total coverage of inland water is apparently over-estimated by 121 km^2, coastal by 338 km^2, and woodland by 330 km^2. By comparison, FS over-estimates these classes by 506 km^2, 580 km^2, and 1 241 km^2 respectively compared with OS98 statistics. For woodland, FS provides a slightly closer match with OS98 than the LCM2000 statistics, but both heavily over-estimate the total (by 7 071 km^2 and 8 293 km^2 respectively).

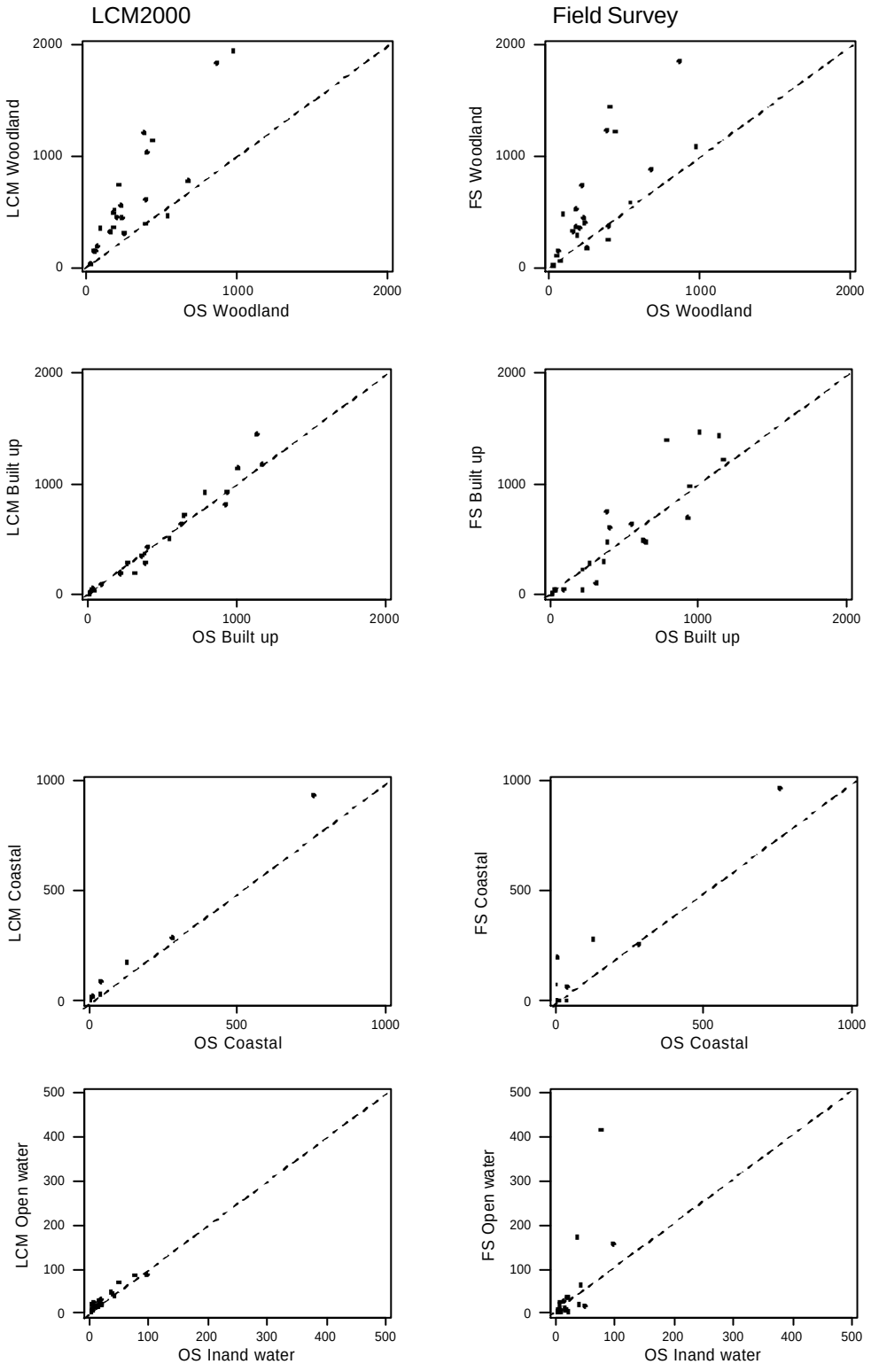


Figure 3. Scatter plots of the LCM2000 and FS estimate of woodland, built up, coastal and inland water land cover against the OS98 total per National Land Class. (The $y = x$ line is also plotted for each graph).

Figure 3 plots the LCM2000 and FS estimate of each of the above land cover types against the OS98 total per NLC. Ideally, land cover estimates per NLC in each plot should fall along the $y = x$ line. It is clear that for inland water, coastal and built up land cover types the LCM2000 estimates are much closer to the $y = x$ line and with less scatter than the FS estimates. For woodland, both the LCM2000 and FS estimates per NLC show considerable scatter and a tendency to over-estimate the total compared with the OS98. For both LCM2000 and FS datasets, the tendency to over-estimate the woodland land cover total becomes worse with increasing aerial coverage; more so for LCM2000 than for FS. The OS98 data thus provide a useful indication of the relative strengths and weaknesses of the FS and LCM2000 estimates of stock, but only for generic land classes and are therefore not useful for the calibration procedure which operates at greater thematic detail.

Forestry Commission Digital Woodland Map for England

The Forestry Commission Digital Woodland Map for England is a vector dataset based on interpretation of 1:25 000 aerial photography (flown in 1991-2000) and plotted against OS 1:25 000 mapping. Woodland parcels consist of areas of tree cover with a crown density of at least 20 %, with a minimum width of 50 m and a minimum size of 2 ha. Woodland classes are: coniferous, broadleaved, mixed, shrub, coppice and young trees. In addition, parcels also identify ground prepared for planting, felled woodland, and young trees. Data were updated by Woodland Surveys for the National Inventory of Woodland and Trees to include FC new planting and New Woodland Grant Schemes, as at 31st March 2000.

The FC dataset, by contrast to the EN dataset, is a complete survey and all 136 286 polygons have a land cover label. This can readily supply overall woodland land cover statistics for England. In addition, the vector data can be compressed into proportional composition of the different 'woodland' classes and 'other' non-woodland cover per 1 km National Grid cell and combined as appropriate to match woodland widespread Broad Habitats identified in LCM2000 and FS.

Comparison of FC data with FS and LCM2000 woodland estimates for England

In the FC dataset, the classes broadleaved, mixed, coppice, and shrub were combined into a broadleaved & mixed woodland Broad Habitat, whilst the classes conifer and young trees were combined into a coniferous woodland Broad Habitat. The estimates of spatial coverage for these two Broad Habitats in England derived from the FC, FS and LCM2000 are shown in Table 3.

Table 3. Total coverage of woodland types (in km^2) in England as reported by FS and LCM2000, compared with FC 2000 data.

	***FS98**	**LCM2000**	**FC 2000**
Broadleaved & mixed woodland	9 970	10 930	6 510
Coniferous woodland	2 980	2 980	3 593

* FS has a restricted spatial coverage compared with LCM2000 and FC 2000

Both FS and LCM2000 apparently under-estimate the total coverage of coniferous woodland in England (both by 613 km^2), but over-estimate the total coverage of broadleaved & mixed woodland (by 3 460 km^2 and 4 420 km^2 respectively). Figure 4 plots the LCM2000 and FS estimate of broadleaved & mixed woodland and coniferous woodland against the FC total for the 21 NLCs that cover England. For coniferous woodland the estimates of LCM2000 lie much closer along the $y = x$ line than for the FS estimates. For broadleaved & mixed woodland, both LCM2000 and FS show a tendency to over-estimate the total coverage in almost every NLC. Note that the FC estimate of the total woodland cover in England (10 103 km^2) is greater than that of the OS98 dataset (8 437 km^2), but this still represents an apparent over-estimate of 2 846 km^2 and 3 806 km^2 for LCM2000 and FS respectively. This may relate to the Minimum Mappable Unit of the FC data (2 ha) compared with LCM2000 (0.5 ha) and FS (0.04 ha). The apparent over-estimate of broadleaved & mixed woodland by both FS and LCM2000 thus probably represents the total of < 2 ha sized patches across England. The parcel information of the FC dataset could thus be simplified to recorded presence or absence per 1 km National Grid grid square and used as a spatial mask in the calibration procedure for woodlands. However, the stock statistics cannot be used for validation of the calibrated product since they do not contain enough spatial detail.

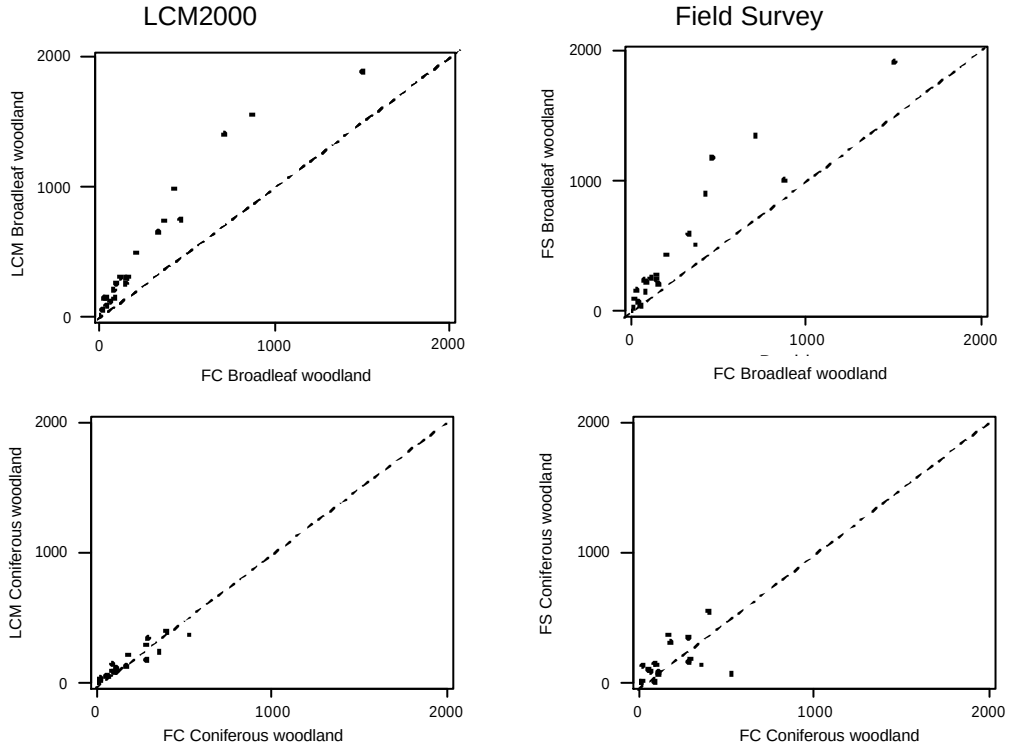


Figure 4. Scatter plots of the LCM2000 and FS estimate of broadleaved & mixed woodland and coniferous woodland against the Forestry Commission total for the 21 National Land Classes of England.

Indicator Species

The Biological Records Centre (BRC) at CEH Monks Wood has a GB dataset on species composition in a 10 x 10 km grid. Preston et al (2003) identified all species associated with BAP Broad Habitats and identified the percentage of acid, calcareous and neutral species of the GB total in each grid square. Therefore, the score of each grid cell could be up to 300 %. This dataset has been used in the Critical Loads project (see http://critloads.ceh.ac.uk) to refine the LCM2000 data on semi-natural grassland identification. They used a cut-off of 50% to determine a 10 km square of calcareous grassland (i.e. a grid cell where at least 50 % of the calcareous species pool for GB is present). They used a cut-off of 40 % for acid and neutral grasslands. The nature of the dataset and these rules means that grid cells can be deemed to have more than one type of semi-natural grassland (in some cases all three). No information is provided in the dataset as to where within a 10 x 10 km grid cell the areas dominated by acid, neutral or calcareous indicator species occur, or indeed of how much of a 10 x 10 km cell could be deemed as being acid, neutral or calcareous grassland.

A breakdown of the acid, neutral and calcareous species composition by soil type (here including peat soils as acid) gave inconclusive results (Table 4). Note that the highest percentage is always for neutral species indicators, irrespective of soil type. This occurs because neutral species indicators are very prevalent and because of the effects of comparing a 10 km species dataset with 1 km soils data. These indicator species data are thus not suited to provide comparisons with FS or LCM2000 stock estimates or spatial masks for use in the calibration procedure.

Table 4. The proportion of the GB stock of indicator species present per 10 x 10 km square on acid, neutral and calcareous soils.

	Indicator species	Average (%)	Minimum (%)	Maximum (%)
Acid soils	acid species	53.44	19.0	80.0
	neutral species	62.48	29.9	83.1
	calcareous species	43.20	13.8	73.5
Neutral soils	acid species	43.21	6.9	78.0
	neutral species	66.13	35.7	84.5
	calcareous species	46.77	13.8	73.1
Calcareous soils	acid species	39.97	15.4	77.6
	neutral species	67.25	46.7	77.4
	calcareous species	60.13	31.9	72.5

Recommendations for the calibration process from the comparisons with nationwide datasets

Comparisons of land cover statistics for the generic 'coastal' and 'built up' land cover types derived from FS and LCM2000 with OS estimates clearly demonstrate that LCM2000 offers by far the better set of statistics. This has been shown to be the case for several reasons. Firstly, LCM2000 can map further off-shore than FS and maps the percentage cover of all National Grid 1 km cells, regardless of the percentage urban content. This obviously has the greatest impact on the coastal and built up land

cover types respectively. LCM2000 thus provides data on the land cover of more National Grid cells than FS. Secondly, even for the areas of overlap in both FS and LCM2000, it has been shown that LCM2000 provides better estimates of total coverage for coastal and built up land cover types. FS apparently over-estimates the cover of both coastal and built up land classes. In the case of coastal land classes, it is perhaps not surprising that the FS statistics over-estimate the total land cover, when the estimates are based on the area of each NLC rather than the length of coastline. There is not likely to be a relationship between the length of coastline and the area of hinterland for each NLC. The over-estimation of urban total land coverage in FS probably relates to the way in which urban areas are mapped by the FS. The whole area within an urban boundary is classed as built up, irrespective of whether there are open spaces, woods, lakes etc. These features are all identified in LCM2000 if above the minimum mappable unit of 0.5 ha. Thus, it is concluded that it would be detrimental to the quality of the LCM2000 statistics to use FS data to calibrate the total coverage per NLC of 'built up' or 'coastal' land classes. However, using the calibration procedure to re-distribute the 'coastal' land class between littoral and supra-littoral Broad Habitats is worth attempting, since these are not well distinguished in LCM2000.

The comparison of land cover statistics for inland water derived from FS and LCM2000 with OS98 estimates also demonstrated that LCM2000 offers by far the better set of statistics. This almost certainly relates to the fact that water bodies are amongst the most readily mapped 'land' cover types as water has very identifiable spectral reflectance characteristics compared with vegetation, bare soil or artificial surfaces. The poorer statistics for FS in this particular case reflects the distinction between extrapolated sample data and well identified census data. Thus, once more it must be concluded that it would be detrimental to the quality of the LCM2000 statistics to use FS data to calibrate inland water.

The difficulties in mapping deciduous woodland in single date or two-date imagery from too early or late in the seasons (as occurred for considerable areas of LCM2000 production) resulted in poor land cover statistics for this land cover class compared with FS statistics. Calibration of LCM2000 broadleaved & mixed woodland and coniferous woodland using FS statistics is thus recommended. This process can be improved by using the FC Digital Woodland Map for England to provide a spatial mask to maintain the within-NLC spatial distributions.

The inadequacies of the soil sensitivity and drift maps used in the post-classification knowledge-based correction procedures in LCM2000 production are highly apparent in the comparisons of FS and LCM2000 cover statistics with soil type. FS estimates per NLC are much closer to what would be predicted based on soil type, especially in the case of neutral and acid grasslands. It is thus recommended that calibration of the LCM2000 estimates for the semi-natural grasslands, dwarf shrub heath and bog make use of FS statistics. However, it is also recommended that the soil type data be used directly in the calibration process, because questions remain for many NLCs as to whether the FS coverage was adequate to characterise the proportional cover of the semi-natural land cover types where these occur only in small, rare or isolated patches. For example, of the 21 NLCs that occur in England, only two have any appreciable amount of calcareous grassland according to FS estimates, and even they have less than 3% of their total area. For neutral and acid grassland, and for dwarf

shrub heath and bog only five of the 21 English NLCs have more than 6% coverage. As an example, Table 5 indicates the number of FS squares dominated by semi-natural vegetation, the proportion of those squares identified with the same dominant vegetation cover in LCM2000 and how that proportion would change if LCM2000 land classes were re-assigned using the soil type data.

Table 5. The affects of calibrating LCM2000 estimates of semi-natural land classes using the soil type data.

Dominant Broad Habitat in FS	No. of FS squares	% correct in LCM2000	% that would be correct using soil data to calibrate LCM2000
Neutral grassland	3	0	33
Calcareous grassland	0	-	-
Acid grassland	17	29	76
Dwarf shrub heath	21	48	76
Bog	4	25	50

Note, this is for the dominant Broad Habitat per FS square; calibration is for the percent composition per FS square.

The data on soil types, derived from the re-classification of soil groups in NATMAP1000, should be used in the calibration process for the semi-natural grasslands, dwarf shrub heath and bog. This will provide a mask to maintain spatial distribution patterns of these Broad Habitats within NLCs, preventing the smearing effect of calibrated stock estimates that would otherwise result. Acid grassland and dwarf shrub heath both occur on acid soils and comparisons with FS data show that mis-classification occurs between these two land cover types in LCM2000. Whilst the soil type data can be used to correct for mis-classification in LCM2000 between these two classes and calcareous or neutral grassland and bog, they cannot be used to correct mis-classification between acid grassland and dwarf shrub heath. Taking a more detailed look at the soil data, it is not possible to identify a particular soil class that can separate acid grassland from dwarf shrub heath, as both most commonly (but not exclusively) occur on stagnohumic gley soils.

For the more rare Broad Habitats of bracken, fen marsh & swamp, inland rock and montane, no consistent nationwide datasets are currently available with which to compare the land cover estimates of FS and LCM2000. In the case of these relatively rare, often isolated or fragmented, and, in the case of bracken, temporally variable land classes, both sample-based field survey and satellite-based land cover mapping struggle to generate meaningful nationwide statistics. Under these circumstances it must be asked whether either survey technique should be expected to provide robust land cover statistics for these Broad Habitats at a range of spatial scales, and whether the calibration of these land cover statistics by integration is a worthwhile aim. Finally, the two most prevalent land cover types in England, arable & horticulture and improved grassland are known to show mis-classification in LCM2000 when compared with FS. This results from rotation farming which is picked up when LCM2000 satellite images were not acquired in the target period coinciding with the timing of FS. Because these two Broad Habitats are so abundant in all NLCs in England, the stock estimates of FS are statistically representative to be used for calibrating the LCM2000 estimates.

Enhancement of calibration process to derive the final product

Feedback on trial 1km summary dataset

The calibration method for integrating LCM2000 and FS data to produce consistent and robust estimates of stock at different scales was designed as an iterative process. By May 2003 the calibration process had reached a fourth iteration, and this was described and explained in the last Module 9a Interim Report (Smith and Hill, 2003). This report also contained sample extracts of the calibrated 1 km summary product, which has been released as a trial data set in CIS format. A copy of the Interim Report was sent to all members of the Technical Advisory Group (TAG) and the attendees of the Module 9 Seminar held in Bristol on 19th May. (Details of the attendees of both meetings were in the last Interim Report). The trial dataset was demonstrated at the May Seminar (Bristol) and at the CIS Data Forum (London) on November 24th, and was presented briefly at the CQC Advisory Group meeting on January 7th 2004 (London). The trial 1km summary calibrated dataset has been sent to all members of the TAG, attendees of the May Seminar and to selected attendees of the CIS Data Forum. In addition, the trial dataset has been presented at CEH Monks Wood and discussed with several field ecologists. Feedback has helped to finalise the calibration method and steer the final set of iterations. The feedback is thus summarised below.

The calibration method of using bootstrap re-sampling of the confusion matrices for each 1 km FS square to obtain 'calibrated' estimates of Broad Habitat coverages was approved for deriving national estimates. In addition, the subsequent corrections by knowledge-based use of masks for known areas or geographical limits of coastal, urban and montane cover types was also regarded as sensible to avoid obvious contextual errors. However, it was felt that averaging the 1 km square correspondence matrices within an NLC had the effect of smoothing (or reducing/eliminating) some of the potentially real spatial variation in Broad Habitat types as mapped in LCM2000. This led to a discussion about the nature of the spatial stratification chosen for the calibration procedure, and whether LCM2000 itself could provide a multivariate, multi-level stratification that would be more detailed than the NLCs. Previous analysis, reported in the first Module 9a Interim Report (Watkins *et al.*, 2003) highlighted an inconsistency in correspondence between LCM2000 and FS stock estimates for different satellite image pairs, and also between single and two-date image classification and between target and non-target date images. However, it was concluded that a spatial stratification based on the NLCs offered the benefit of (i) optimal use of the spatial distribution of FS squares, as the original ITE Land Classes provided the stratification for the systematic random location of FS squares, and (ii) replacing the artificial image boundaries present in LCM2000 with boundaries that are based on environmental variables.

The prototype calibrated 1 km dataset contained no estimates of uncertainty and this was requested by many users. However, it was recognised that the bootstrapping procedure could give some estimate of uncertainty for the calibrated product at national or regional scales, similar to the confidence limits or standard errors from the field survey. Concern was raised on the issue of how to estimate, report and indicate the uncertainty in estimates of cover for Broad Habitat types for much smaller regions and even individual squares. Providing estimates of uncertainty for small-area calibrations is much harder than for large areas, because not only do the 'average'

misclassification rates have to be considered, but also the tendency for misclassifications to be correlated between adjacent parcels. This is equivalent to the two types of confidence interval in regression, one for the fitted line, and the other for points about the fitted line. However, as no obvious solutions were apparent from the discussions it was felt that uncertainty measures should be derived where possible and health warnings should be clear in the CIS output as to the spatial scale at which the calibrated statistics should be considered valid.

The final set of comments relating to the prototype calibrated 1 km dataset were Broad Habitat specific errors highlighted in certain geographic locations. Examples of these included the 'disappearance' of Salisbury Plain in the calibrated product for calcareous grassland and the transfer of Thetford Forest from a high percentage coniferous woodland to broadleaf & mixed woodland. Numerous examples were identified, in particular for neutral and calcareous grassland, coniferous and broadleaf & mixed woodlands and bog. As the strength of the calibrated product is at the level of land cover estimates per NLC rather than the individual 1 km squares, a suggestion was made as to the appropriateness of releasing the calibrated product as a CIS data layer, as this package automatically displays data as 1 km resolution maps. The suggestion was to have a tabular display of Broad Habitat stock estimates per NLC (with associated uncertainty measures) rather than a cartographic output with different values per 1 km square. However, as far as possible the individual geographical 'errors' in the calibrated product have been addressed by incorporating ancillary datasets into subsequent iterations of the calibration procedure. Hence, the National Soil Map for England and Wales (classified into calcareous, neutral and acid soil types) and the Forestry Commission Digital Woodland Map for England (rendered into proportional coniferous and broadleaf woodland per 1 km^2) were incorporated into the fifth and sixth iterations respectively.

Enhancements to the calibration process for the 'final' version

The soil type data were used in the calibration procedure to alter the average correspondence matrix for each NLC if one of the three soil types (acid, neutral or calcareous) was found to be dominant. The alteration merged the sections of the calibration matrix that calculate semi-natural grasslands, so that only semi-natural grasslands related to soil type were produced. The total area of semi-natural grasslands remains the same. Thus, for example, in an NLC with dominant acid soil the calibration coefficients for neutral and calcareous grasslands were added to the calibration coefficient for acid grassland.

The ancillary woodland data were used to prevent the creation of large amounts of woodland by the calibration procedure which altered the balance of broadleaf, mixed & yew (BH 1) and coniferous (BH 2) woodland. The Digital Woodland Map for England was used to identify 1 km squares where only one woodland class is present (and with > 5 % coverage) and where one woodland class is dominant (i.e. greater than twice the percentage cover of the other, but both with > 5 % cover). The calibration rules applied were as follows:

- broadleaf, mixed & yew woodland only – do not create any new coniferous woodland;
- broadleaf, mixed & yew woodland dominates – do not create new coniferous woodland if that will make the 1 km square coniferous dominated;

- coniferous woodland only – do not create any new broadleaf, mixed & yew;
- coniferous woodland dominates – do not create new broadleaf, mixed & yew woodland if that will make the 1 km square broadleaf dominated.

These woodland rules were achieved by setting a single cell of the calibration matrix to zero and then re-normalising that row. The above sets of rules were applied to the calibration matrix immediately prior to the calibration matrix being used.

'Final' calibration output

The 'final' calibrated 1 km dataset is shown in Figure 5, as displayed per Broad Habitat in CIS version 7. In general, an obvious distinction between LCM2000 (Release 1) and the calibrated product (Iteration 6) for the terrestrial Broad Habitats is a shift away from areas with 0 % coverage to a smoothing effect whereby the proportional cover is smeared across NLCs to give a low level background percentage cover. As a result, the boundaries of NLCs are often highly apparent in the calibrated dataset, generally replacing any satellite image boundaries that were present in LCM2000 (Release 1). There are some specific points that can be made for individual Broad Habitats:

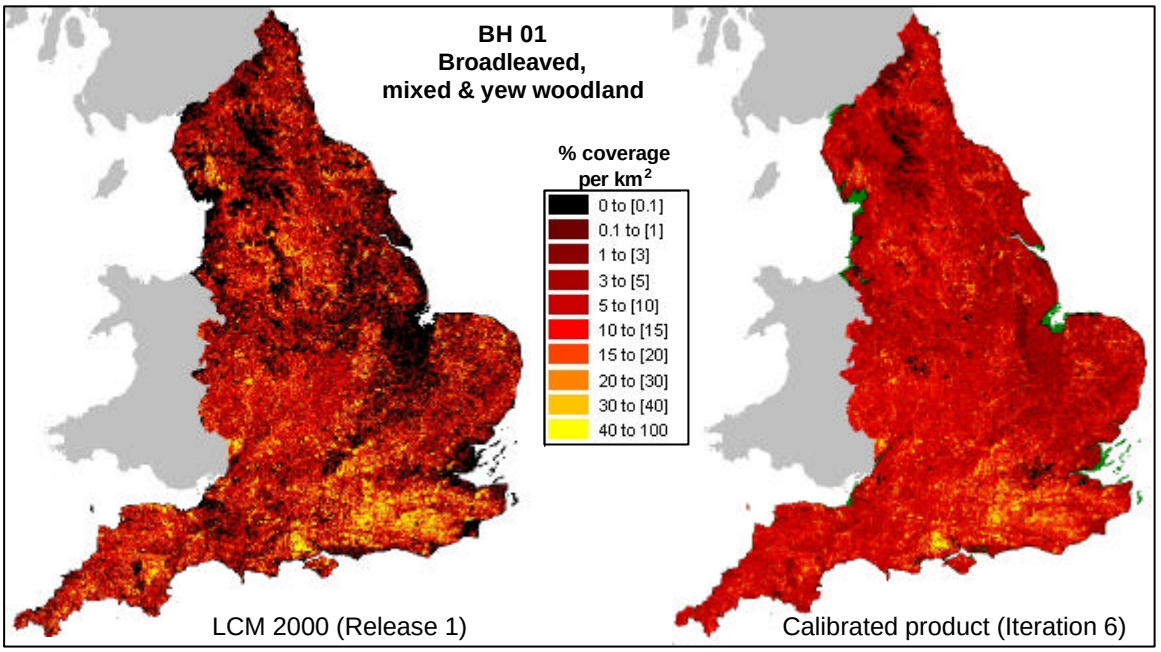
Broadleaf, mixed and yew woodland: a general decrease in the higher proportional coverage areas;
Coniferous woodland: a general decrease in the higher proportional coverage areas in the south, particularly Thetford Forest;
Arable and horticulture: a reduction in proportional coverage around the Pennines, and a loss of landscape features such as Salisbury Plain, Thetford Forest and Bodmin Moor (which should be devoid of this class);
Improved grassland: a general increase in proportional cover in the east of England, also an increase in cover on Dartmoor;
Neutral grassland: a general reduction of proportional cover around the Pennines where LCM2000 had high coverage, but a general increase in 'background levels' in much of England;
Calcareous grassland: a loss of all areas of high proportional coverage and restriction to areas of calcareous soils;
Acid grassland: an increase in proportional cover around the Pennines (especially south Pennines), an increase in background levels in East Anglia, and the high concentrations in the South West peninsula associated with the various moors has been smoothed across the whole area;
Bracken: a decrease in proportional cover in those areas in LCM2000 of high percentage coverage, but an increase in 'background levels' in most parts of England;
Dwarf shrub heath: an increase in cover in the surrounding areas of LCM2000 with high proportional cover, and a background increase in East Anglia and the SW peninsula (one image boundary remains);
Fen marsh and swamp: an overall increase in 'background levels' and an obvious shift away from satellite image boundaries;
Bog: a general spread in distribution around the high proportional coverage areas in LCM2000, and an increase in East Anglia;
Standing open water and canals + Rivers & streams: a shift from linear features to a general 'background level';

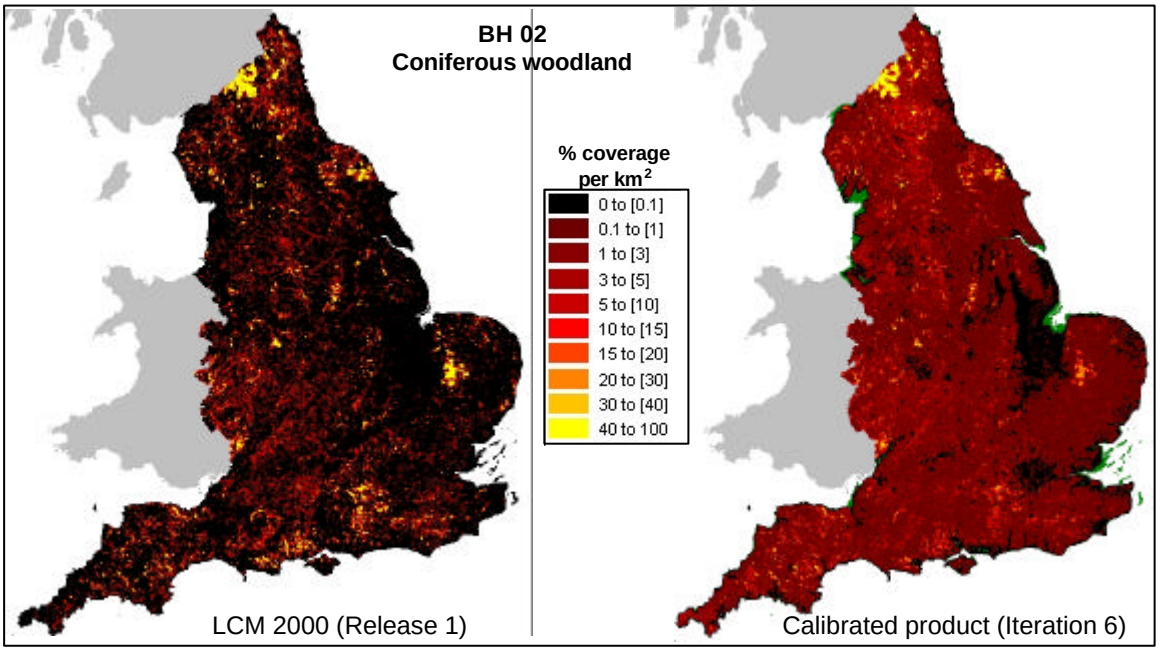
Montane: an appearance at very low percentage coverage in the high ground of the Lake District;
Inland rock: an overall increase in 'background levels' and an obvious shift away from satellite image boundaries;
Built up and gardens: an overall increase in 'background levels';
Supralittoral rock: increase in proportional coverage, especially in the south;
Supralittoral sediment: increase in proportional coverage, especially in the south;
Littoral rock: a complete reduction around the south west and north east;
Littoral sediment: a reduction in the spatial coverage in all the major English bays.

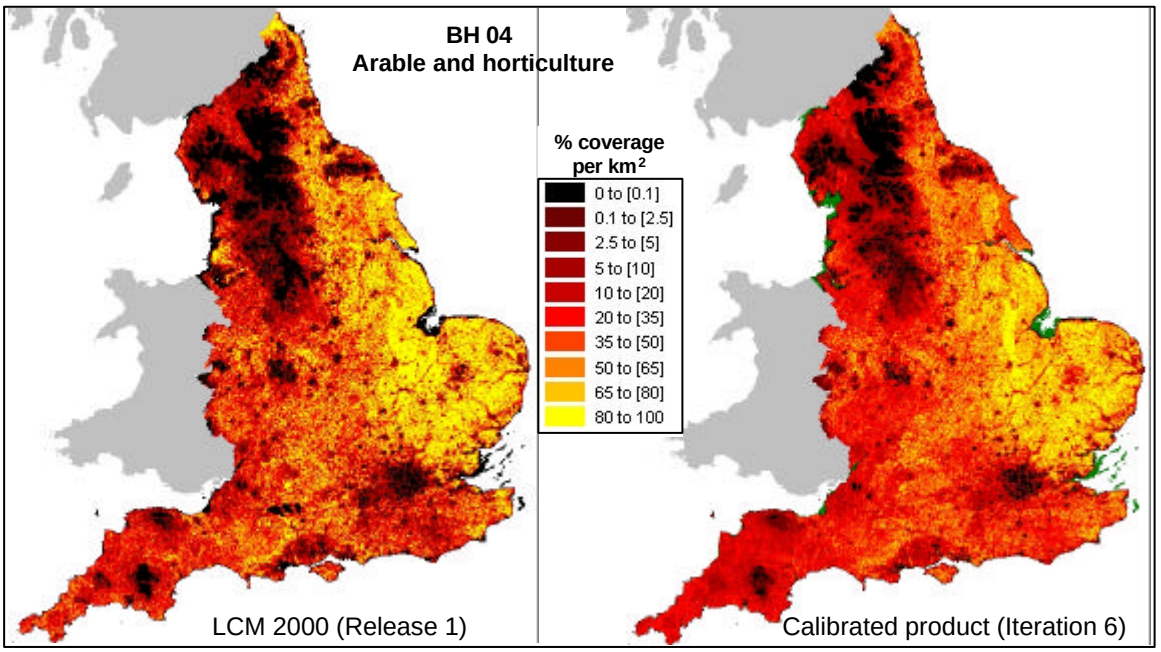
The validation of the final calibrated product has yet to be undertaken. However, a visual QA and discussions with field ecologists at CEH Monks Wood has led to the general conclusion that the calibrated product is in almost all cases better than the original 1 km summary LCM2000 dataset. It has to be accepted that the calibration procedure (i) often results in a very low (i.e. 1-2%) cover of Broad Habitats where they should not be expected to occur; and (ii) causes a general smearing of land cover across the landscape (constrained by the spatial boundaries of NLCs). This smearing effect can sometimes reduce the definition of obvious landscape features. This occurs, for example with Dartmoor in bog, acid grassland and improved grassland. However, it has to be borne in mind that Dartmoor does have a varied land cover and so its boundaries should not be expected to be sharp when individual Broad Habitats are displayed as percentage cover per 1 km square. The two obvious cases of where the calibrated product is worse that the original LCM2000 1 km summary product is for standing open water & canals/rivers & streams and littoral sediment; the former loses its geographical distribution as linear features, the later is reduced in area by the restricted off-shore limits of the NLCs.

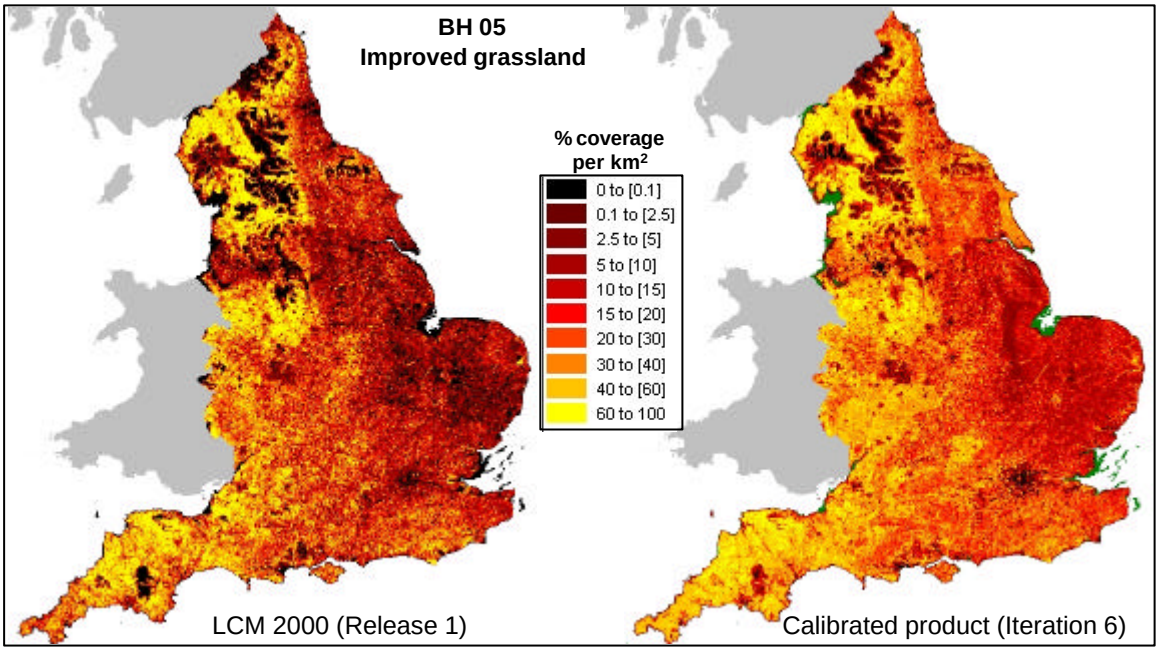
The one Broad Habitat for which issues still remain in the 'final' calibrated product is calcareous grassland. LCM2000 identified too great a spatial distribution of this Broad Habitat, and in places with too high a percentage coverage. The calibrated product has a better spatial distribution of calcareous grassland for lowland England (approximately to the south east of a line from Exeter to Durham), but with low proportional coverage in places. However, for upland England (i.e. to the north west of a line from Exeter to Durham) the calibrated product misses virtually all of the calcareous grassland as these occur on non-calcareous soils. This reflects several problems with how the calibration procedure operates:

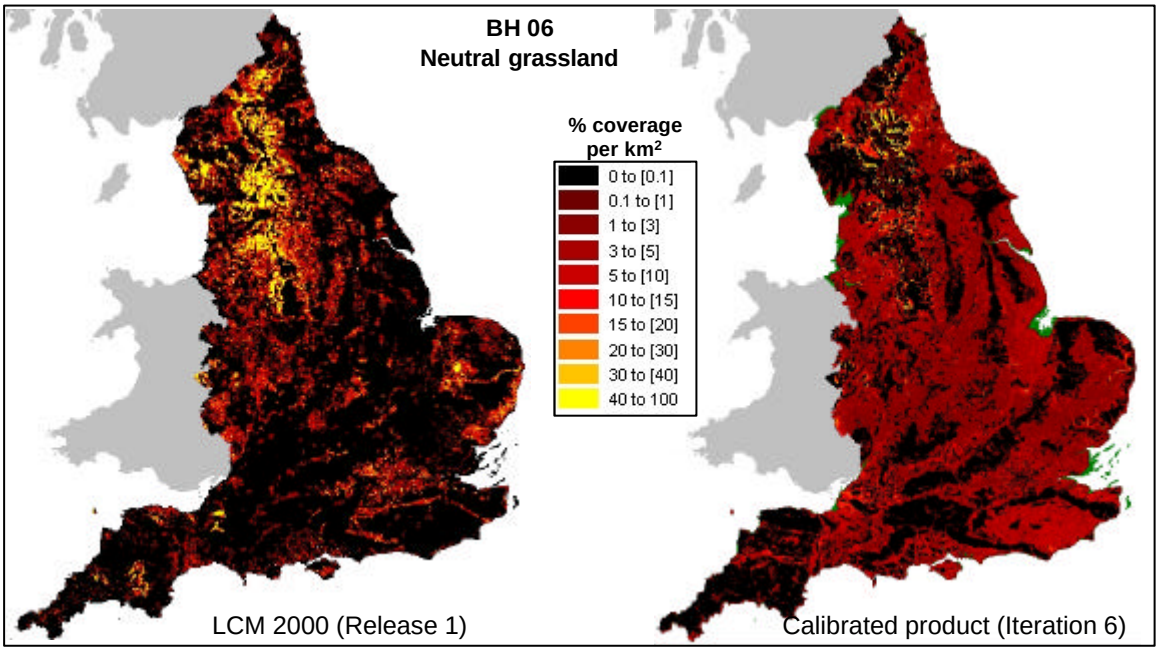
1. Calcareous grassland is a reasonably rare land cover in England (compared with other grassland types). In many areas the spatial distribution of FS squares was not adequate to detect calcareous grassland in a representative way. An example of this is Salisbury Plain, which is missed completely by the stratified random sample of the FS squares. Thus, in FS the total predicted coverage of calcareous grassland for the NLC which contains Salisbury Plain is too low. Calibrating the statistics of LCM2000 in accordance with the FS estimate results in a lowering of the estimated coverage of calcareous grassland for this area and the subsequent disappearance of Salisbury Plain as a landscape feature.
2. In the north of England calcareous grassland occurs on non-calcareous soils. Thus constraining the calibration procedure to allow calcareous grassland only on calcareous soils is too restrictive. However, without this constraint the calibrated product is significantly worse (Figure 6).

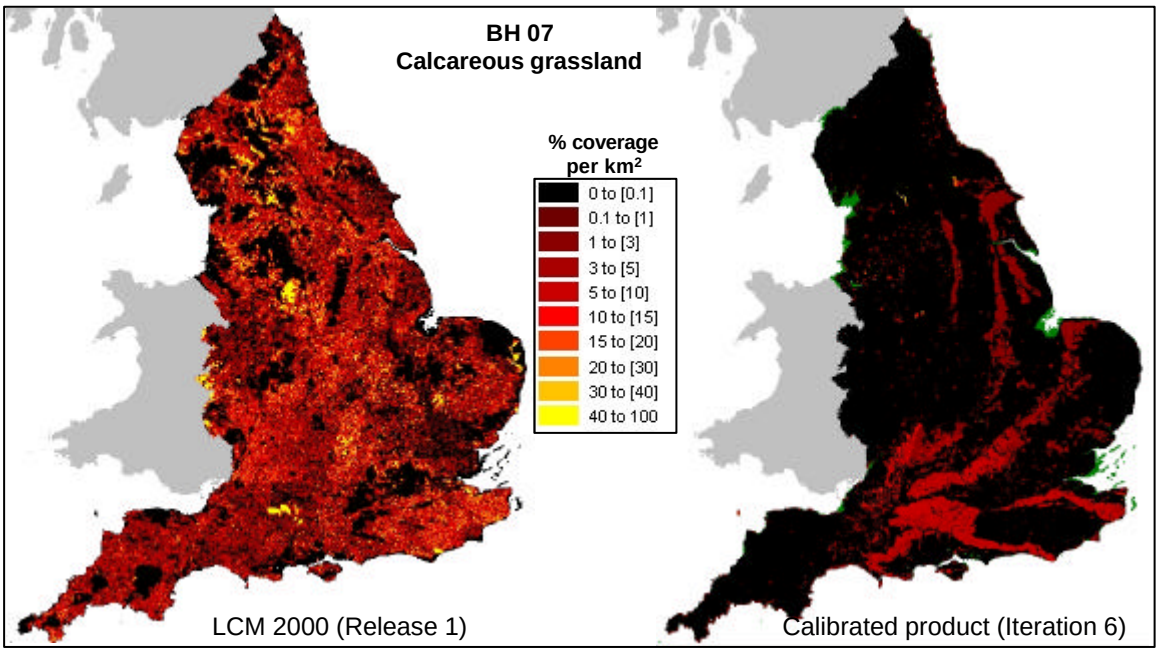

BH 01
Broadleaved,
mixed & yew woodland
% coverage
per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

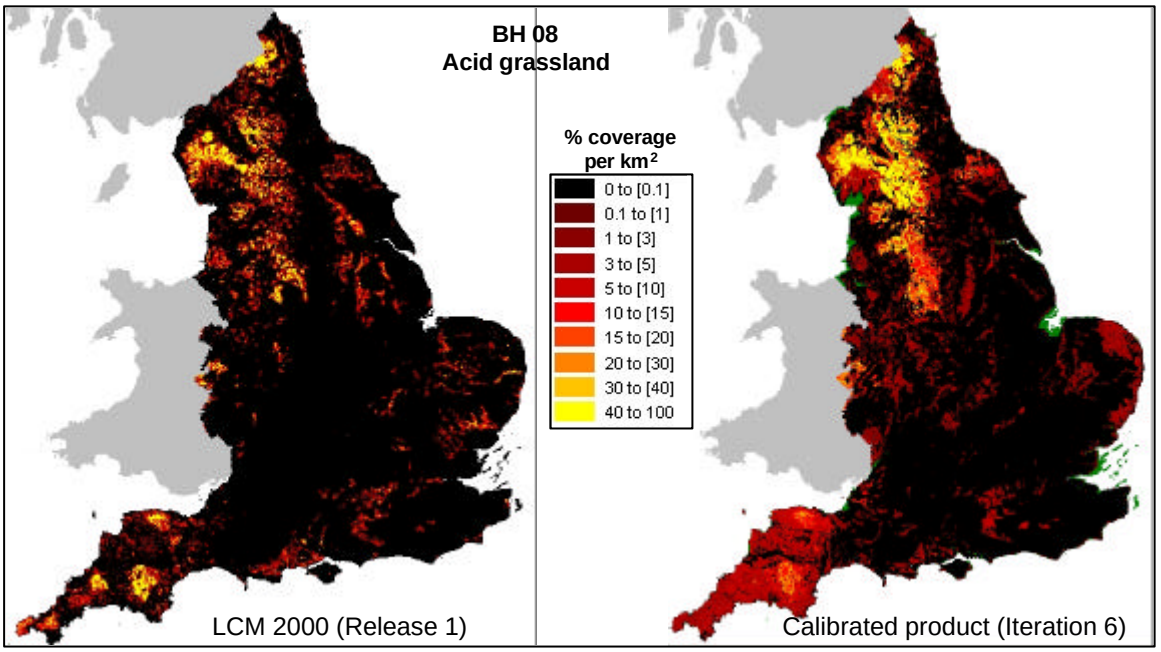

BH 02
Coniferous woodland
% coverage
per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

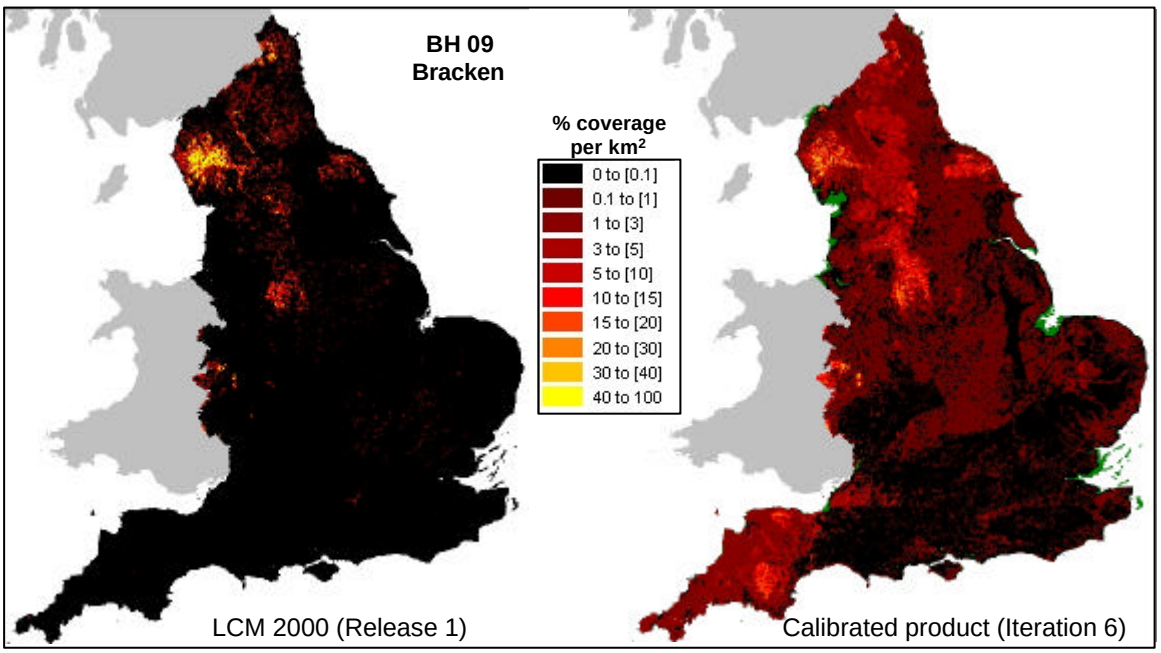

BH 04
Arable and horticulture
% coverage
per km²
0 to [0.1]
0.1 to [2.5]
2.5 to [5]
5 to [10]
10 to [20]
20 to [35]
35 to [50]
50 to [65]
65 to [80]
80 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

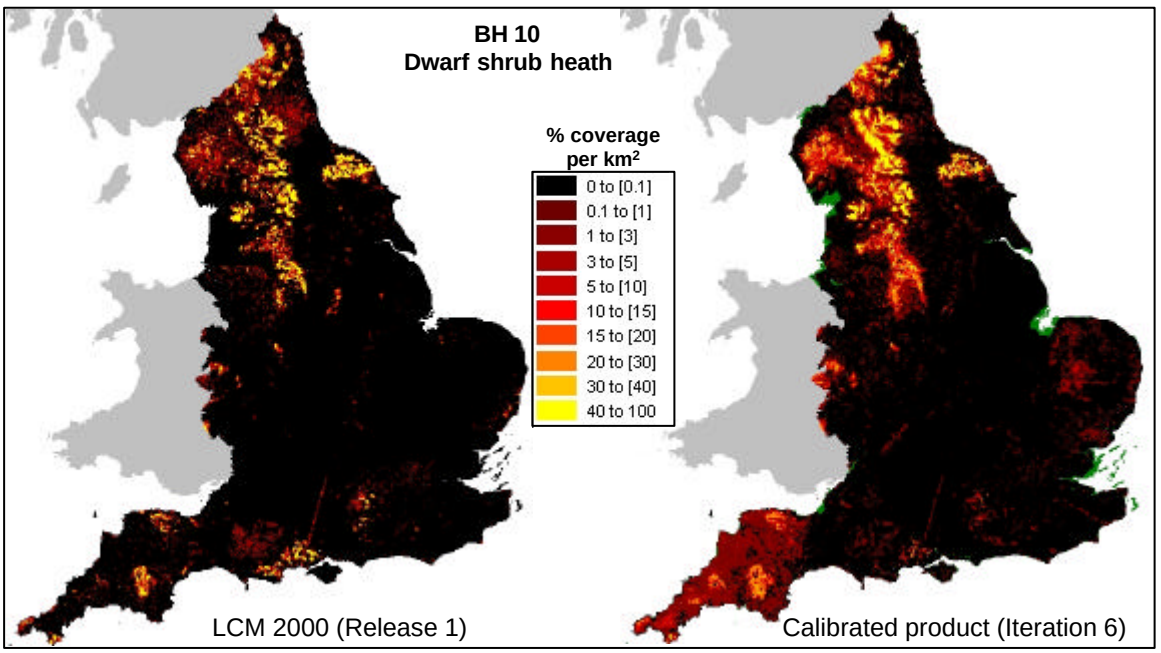

BH 05
Improved grassland
% coverage per km²
0 to [0.1]
0.1 to [2.5]
2.5 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to [60]
60 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

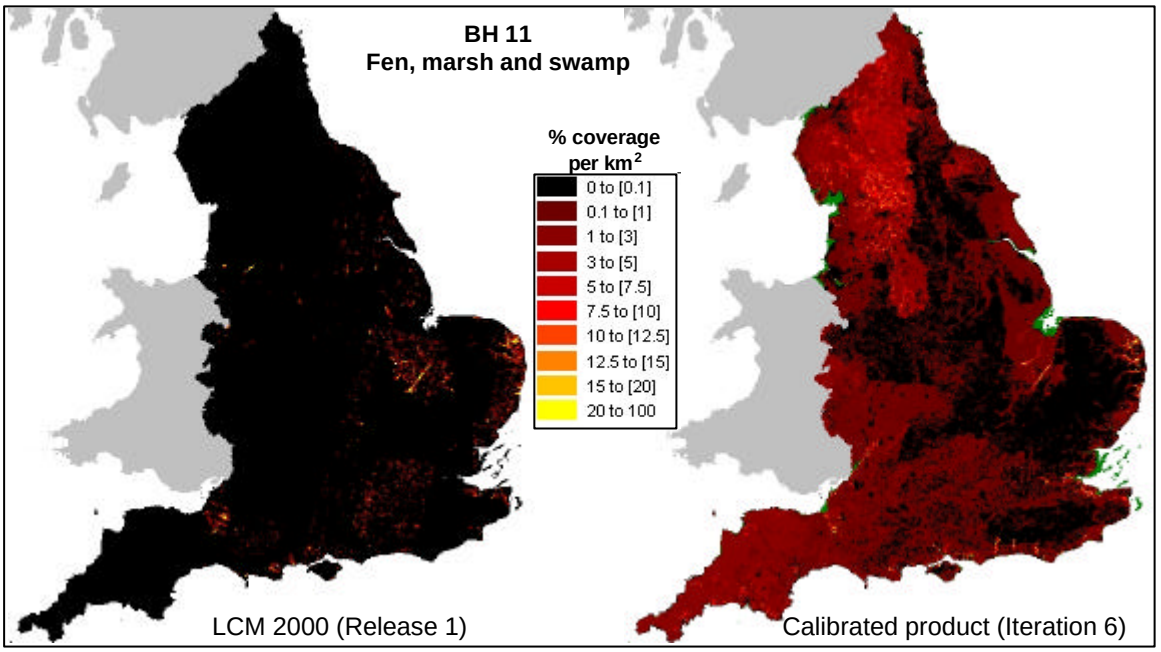

BH 06
Neutral grassland
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

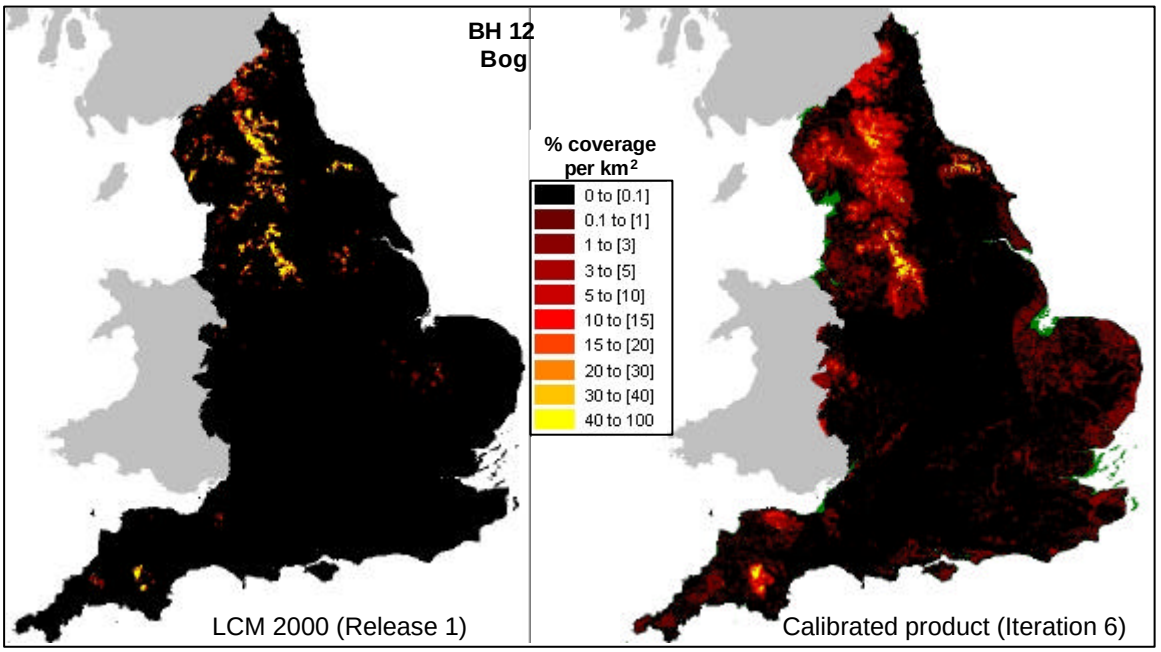

BH 07
Calcareous grassland
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

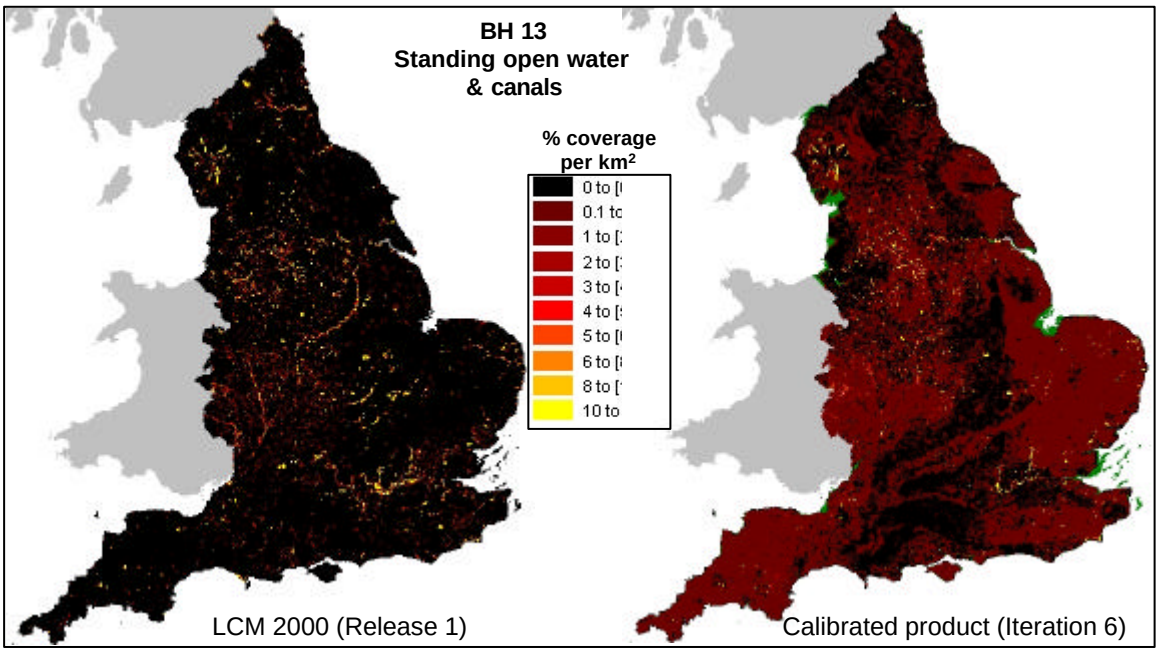

BH 08
Acid grassland
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

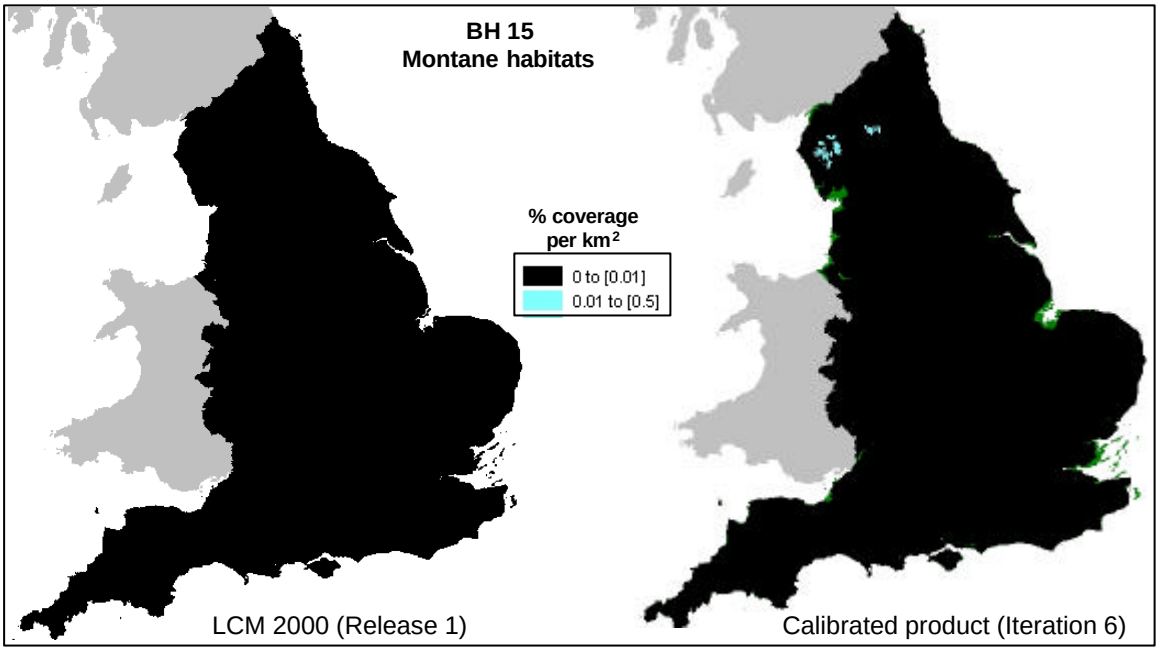

BH 09
Bracken
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

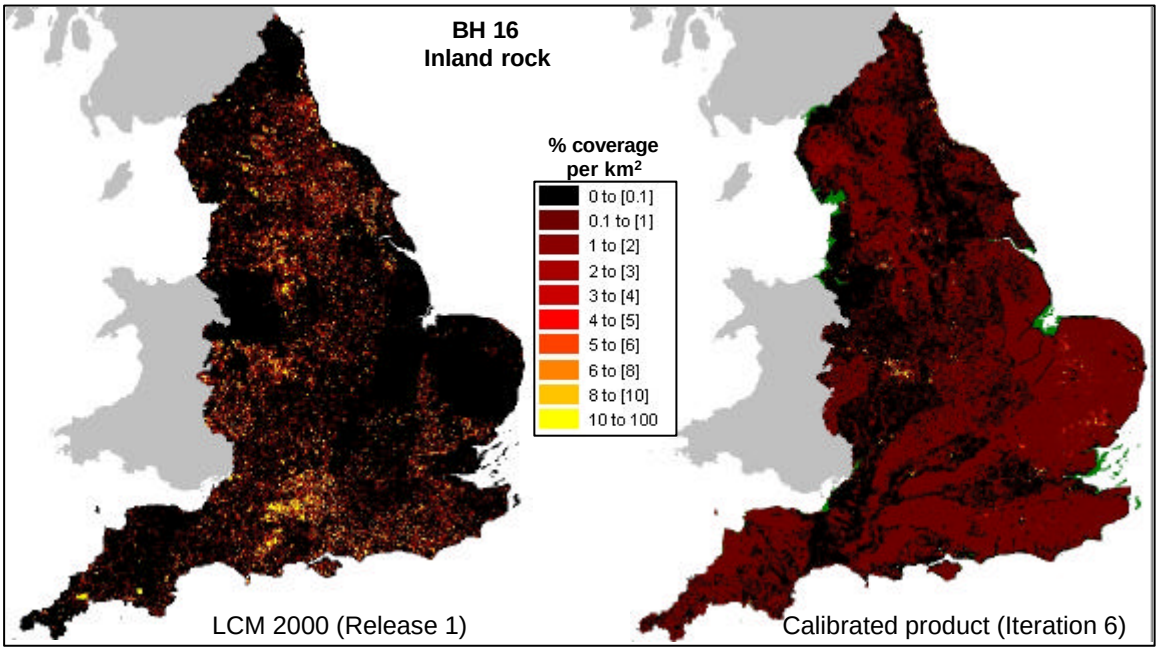

BH 10
Dwarf shrub heath
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

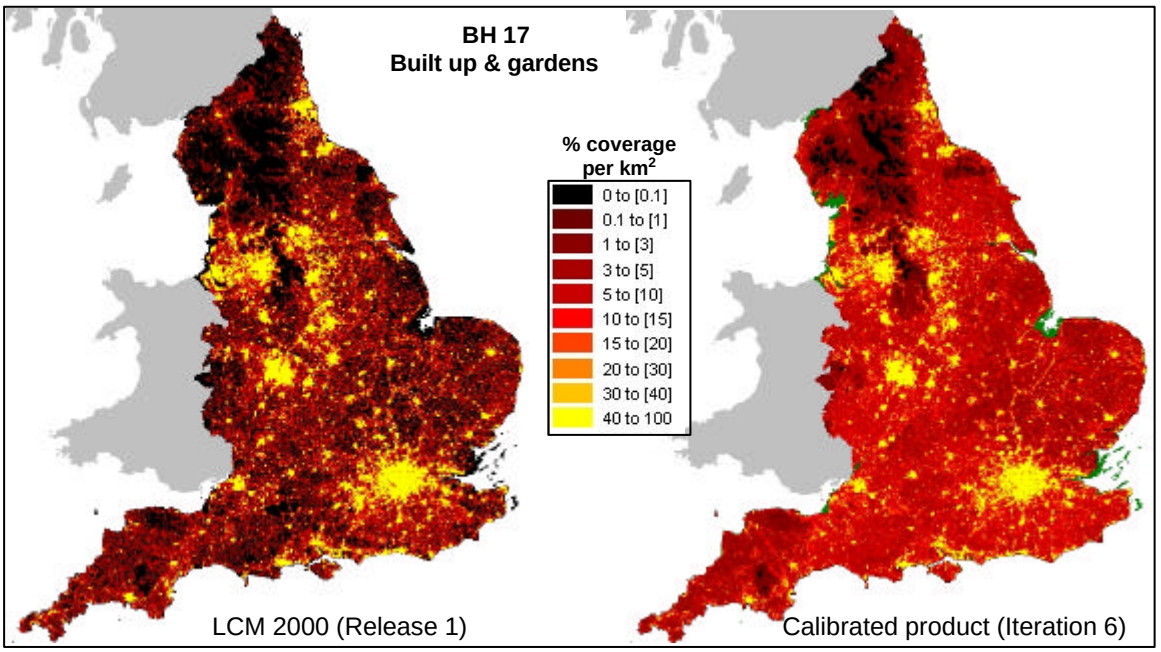

BH 11
Fen, marsh and swamp
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [7.5]
7.5 to [10]
10 to [12.5]
12.5 to [15]
15 to [20]
20 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

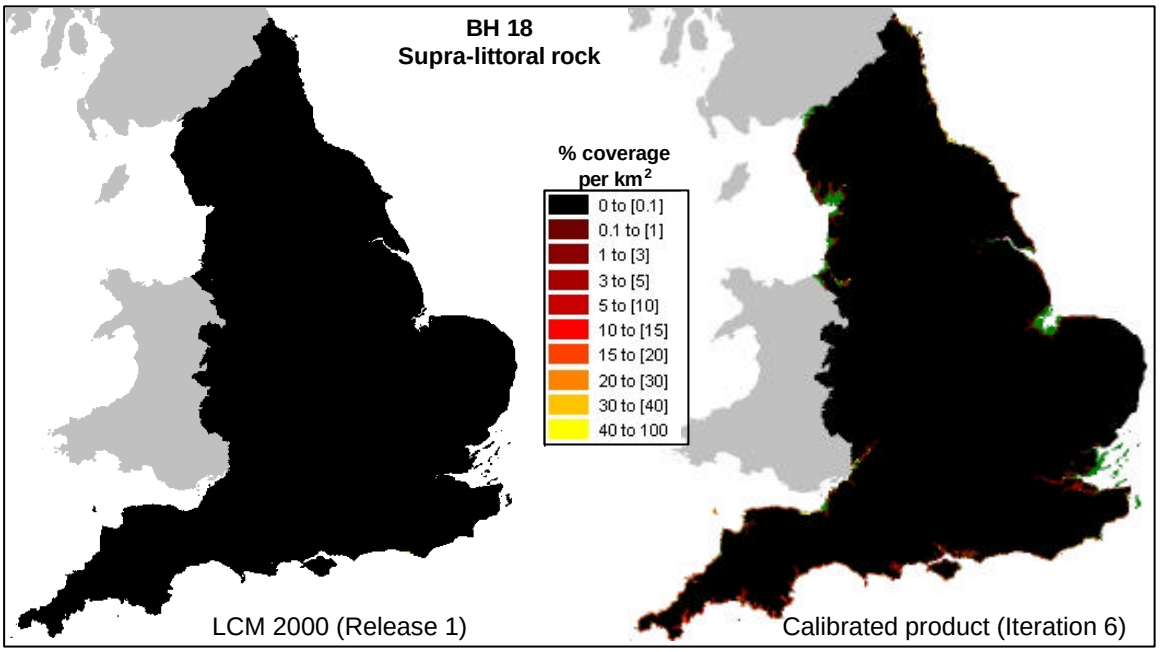

BH 12
Bog
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

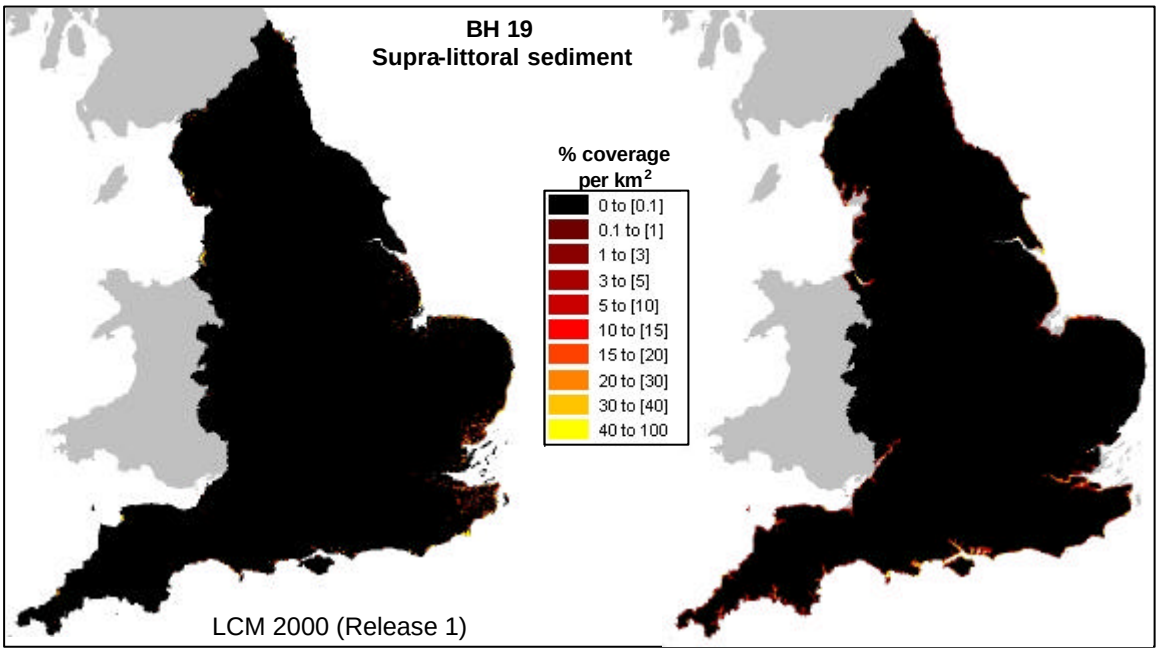

BH 13
Standing open water & canals
% coverage per km²
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

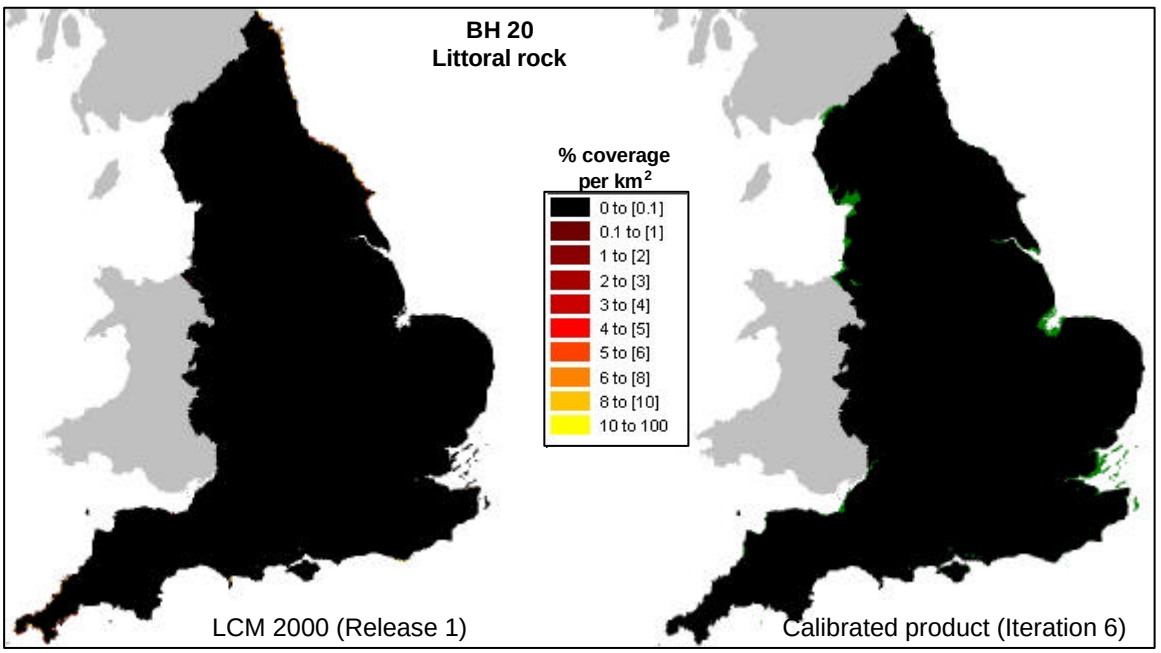

BH 15
Montane habitats
% coverage per km²
0 to [0.01]
0.01 to [0.5]
LCM 2000 (Release 1)
Calibrated product (Iteration 6)


BH 16
Inland rock
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [2]
2 to [3]
3 to [4]
4 to [5]
5 to [6]
6 to [8]
8 to [10]
10 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)


BH 17
Built up & gardens
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)


BH 18
Supra-littoral rock
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)


BH 19
Supra-littoral sediment
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [3]
3 to [5]
5 to [10]
10 to [15]
15 to [20]
20 to [30]
30 to [40]
40 to 100
LCM 2000 (Release 1)


BH 20
Littoral rock
% coverage per km²
0 to [0.1]
0.1 to [1]
1 to [2]
2 to [3]
3 to [4]
4 to [5]
5 to [6]
6 to [8]
8 to [10]
10 to 100
LCM 2000 (Release 1)
Calibrated product (Iteration 6)

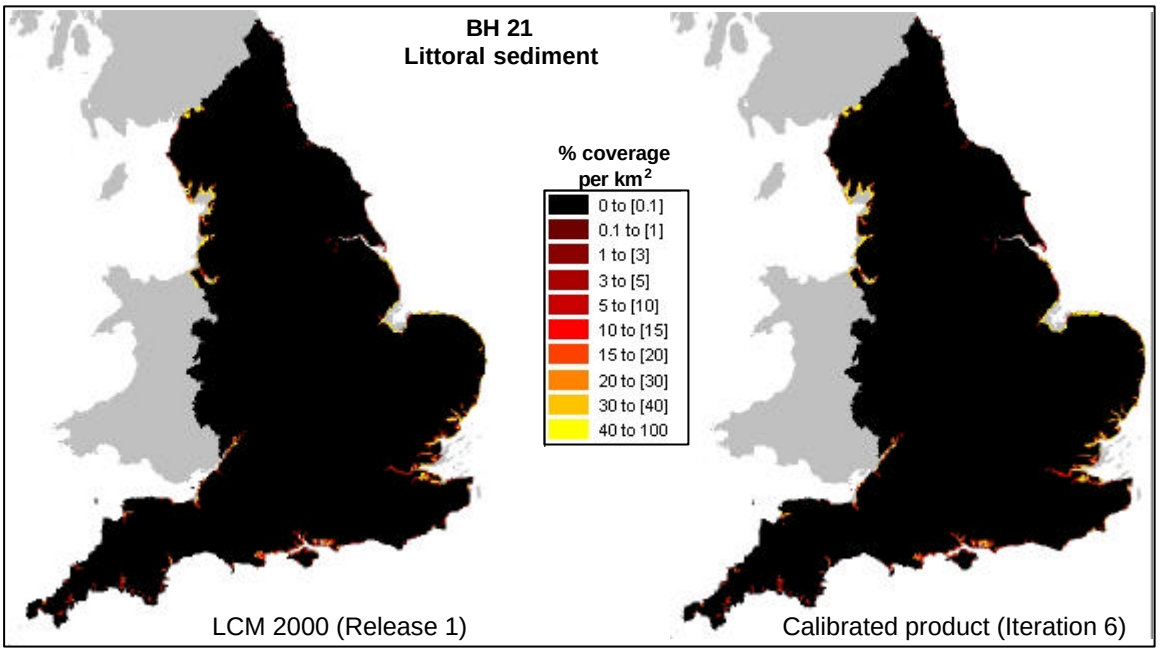


Figure 5. The 1km summary land cover data shown for individual Broad Habitats; (left) derived from LCM2000 Release 1 (right) the calibrated product, Iteration 6.

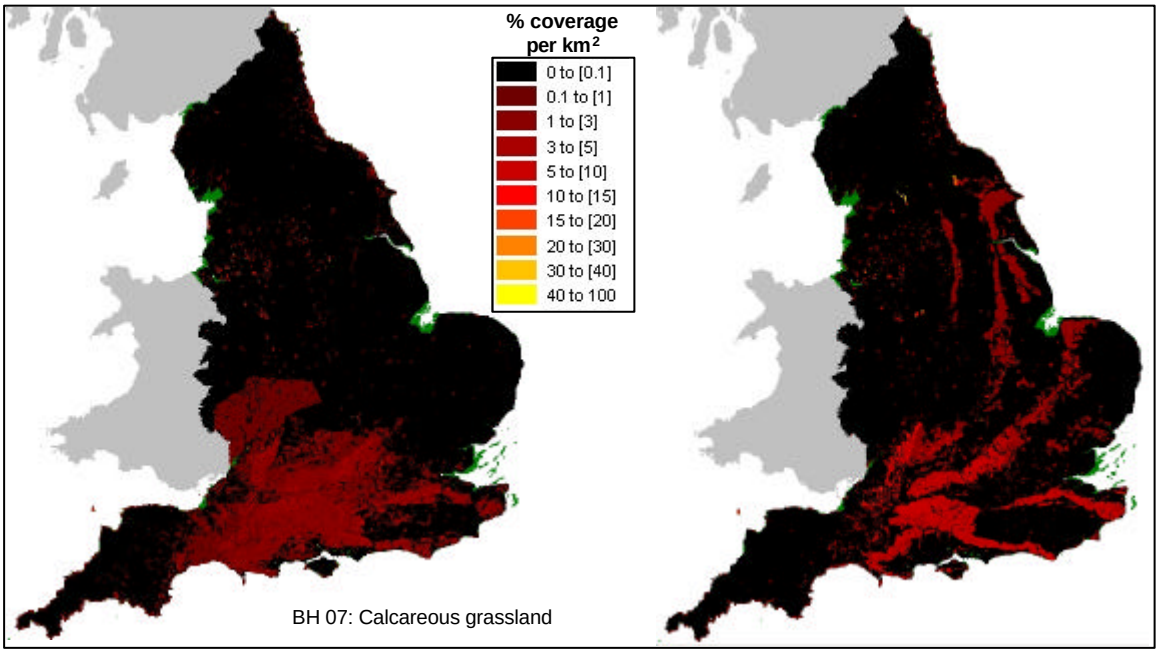


Figure 6. The calibrated 1km summary data showing calcareous grassland; (left) Iteration 4, (right) Iteration 6.

A suitable dataset is not currently available to allow an accurate calibration of LCM2000 for calcareous grassland percentage cover per km square. Unfortunately, this includes the FS data, the NSRI soil data and the EN Grassland Inventory. A calibration is required, however, since the LCM2000 estimates far too much of this particular Broad Habitat. It may be possible to acquire a pre-release copy of the EN BAP Priority Habitat inventories for lowland and upland calcareous grassland. This data could then be used to create a 1 km spatial resolution dataset of presence/absence and used as a mask for the calibration process. However, although this has been discussed with EN, early completion and release of this dataset cannot be guaranteed. The options are thus: (i) to release the 1 km calibrated dataset for calcareous grassland from Iteration 6 with a strong health warning indicating its weaknesses; (ii) to wait for the release of Iteration 7 as a new 'final' 1 km calibrated datset once the EN data have become available; or (iii) to release a 10 km calibrated product for just this one Broad Habitat and use the BRC indicator species data to constrain the calibration process. Figure 7 highlights in red the 10 km squares of England which have at least 50% of the GB stock of calcareous indicator species. This much more closely matches the known distribution of calcareous grasslands; the stock estimates from LCM2000 and FS could provide the input to derive a calibrated dataset of percent coverage.

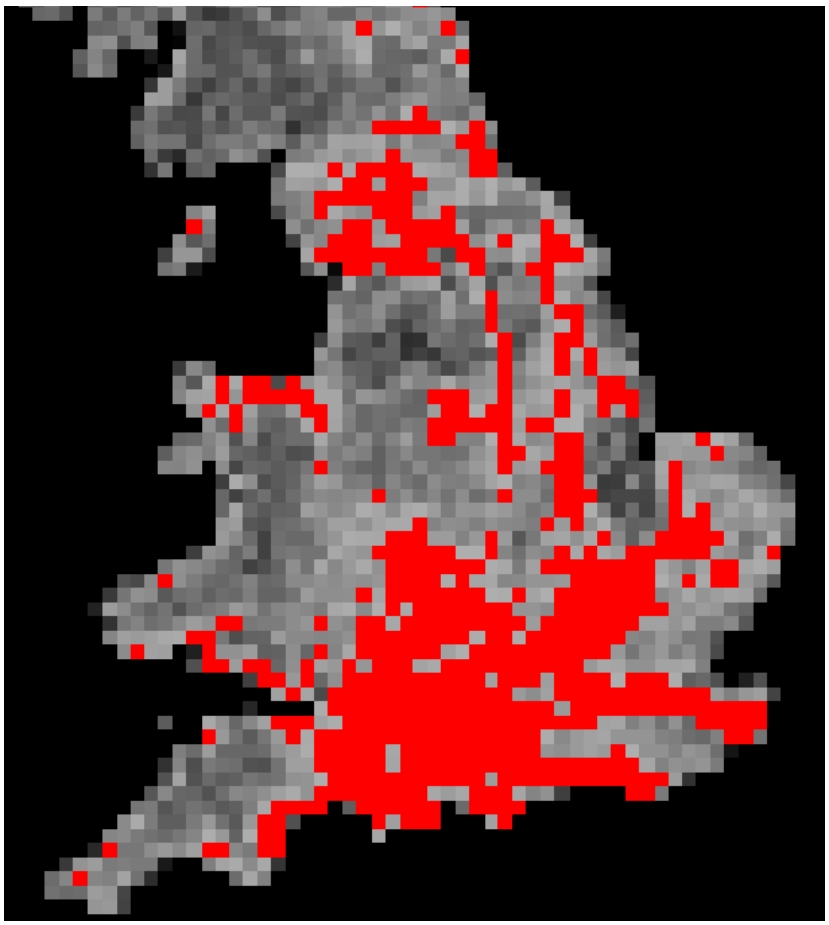


Figure 7. Percentage of the GB stock of calcareous indicator species per 10 km^2 (black 0%, shades of grey 1-50%, red > 50%).

Validation analysis

The validation exercise has yet to be carried out on the final calibrated product. In terms of statistical analyses, the first task will be to compare the FS and calibrated product land cover estimates per NLC. These datasets should be correlated and the relative differences between FS, LCM2000 and the calibrated 1 km will be very revealing.

The soils data cannot be used for validation, since they form part of the enhanced calibration procedure. The FC woodland data were also used in the enhanced calibration procedure, but only as a spatial mask to identify where to 'switch on' or 'off' the calibration. Hence, statistical validation of the per-NLC estimates of broadleaf, mixed & yew woodland and coniferous woodland could be performed using the FC woodland data. However, as reported above, in mapping only woodland parcels larger than 2 ha, this dataset tends to under-estimate the total woodland coverage (especially for broadleaf, mixed & yew woodland). Therefore, the FC woodland data are not suitable for validation purposes.

The EN Grassland Inventory data were not complete or detailed enough to provide a validation dataset. The forthcoming Priority Habitat inventories from EN would be more useful, but not statistically representative as parcels for designated areas with more than one Priority Habitat present will appear more than once in the dataset.

Whilst the BRC indicator species dataset has a spatial resolution too coarse for the calibration procedure it can be used for a basic QA, but not to provide statistical measures of the calibrated product validity. The OS 1 km summary Geographic Reference Data for GB can be used to provide validation for the generic classes of woodland, inland water, coastal and built up.

The June Agricultural Census data for 1998 can be used to provide a statistical comparison of arable & horticulture and improved grasslands. The June Agricultural Census is an annual census of agricultural activity conducted by a postal questionnaire which collects information from farm holdings including land use, crops, livestock, and horticulture. In 9 years out of every 10, the Census is in fact conducted as a sample survey. A stratified random sampling approach is adopted in which holdings are divided into groups (strata) on the basis of their economic size, with higher sampling rates being used in the larger strata. National and local figures are then estimated based on the data received. The Census data from 1998 have been made available from the CQC project and have been processed into Countryside Character Areas. This dataset thus offers the ability to examine the validity of the calibrated product for the two most prevalent English Broad Habitats and for smaller geographic units than the calibration process was applied to. This will help to address one of the issues raised in the feedback from Iteration 4 of the calibrated product.

Uncertainty

The 1 km results of Iteration 6 represent the calibration procedure in its 'final' form. It is now being extended to provide uncertainty information by bootstrapping the sample squares used to generate the calibration matrices for each NLC. The software developments are in hand, but the results are not yet available in time for this report.

Landscape pattern metrics per 1 km square derived from LCM2000

LCM2000 data provides a measure of landscape spatial structure identified in land parcels which were based on spatial boundaries in spectral reflectance characteristics of the Earth's surface recorded in satellite imagery. As a result, the landscape boundaries identified in LCM2000 often do not match those identified in the field. Thus, a single large field containing a crop with a variable growth pattern may be subdivided in LCM2000, whilst a landscape composed of several small fields of the same land cover type and condition may be amalgamated as a single feature in LCM2000. In enclosed landscapes, the parcel boundaries of LCM2000 are generally (but not exclusively) less reliable than the FS boundaries. LCM2000 frequently over-segments the landscape, dividing individual fields into several parcels. FS boundaries for woodland or urban areas, however, fail to separate within-parcel land cover heterogeneity that LCM2000 identifies, such woodland clearings or urban open space. In unenclosed (semi-natural) environments, LCM2000 boundaries are more reliable than FS boundaries, having been derived objectively from satellite spectral reflectance characteristics.

Firstly, it must be considered that landscape structure refers to the individual features or building blocks which make up a landscape, whilst landscape pattern refers to the spatial configuration of those building blocks. Statistical measures of landscape structure and pattern derived from LCM2000 and FS will thus be different, as the two datasets identify different building blocks. However, at a more aggregate level (i.e. combining neighbouring parcels of the same land cover type) LCM2000 and FS can show more similar structure and pattern in landscape patch dynamics. In other words, the overall picture may be similar, but the breakdown of patch dynamics into individual parcels may be very different. Therefore, measures of landscape structure and pattern derived from LCM2000 and FS will vary depending on whether the focus is on the parcel or patch level. Which of these takes the focus depends largely on whether the interest is on land cover areas or the boundaries of landscape features.

A second issue for consideration when looking at the parcel-based structure of LCM2000 and making comparisons with the FS data is that the LCM2000 data offer complete spatial coverage, whilst the FS squares are 1 x 1 km extracts. Thus, it is possible to extract LCM2000 data selecting all parcels that intersect an area (giving a set of parcels that extend beyond the 1 km survey square boundaries). The FS data however are trimmed down to a 1 x 1 km area only. As a result many of the parcels in the FS data are truncated at the FS square boundary, thereby placing artificial boundaries on the landscape. In a recent study carried out at CEH Monks Wood (Swetnam, unpublished data) it was found that for 23 FS squares spanning the English and Welsh National Land Classes, on average 40 % (15-73 %) of parcels were entirely contained within the 1 km square. The two suggested approaches for dealing with this were either to use only the fully contained parcels or to digitise the complete outlines of the truncated parcels from the OS MasterMap database. The first option would in many cases significantly reduce the amount of remaining data for analysis and the second option is not practical within the time frame of this project.

For the work package concerned with developing and evaluating measures of landscape structure and pattern there must therefore be two separate areas of investigation.

1. It is important to identify the extent to which LCM200 over- or under-segments the landscape, and how this differs spatially in relation to different landscape types. It will be necessary therefore to investigate patch number, patch size and boundary information between LCM2000 and sample FS squares. A suggested study area is the East Midlands, which has proved a data rich test area for the CQC project. The focus will be on comparing patch and boundary information recorded by LCM2000 and FS and developing a method for calibrating the two sets of statistics and thereby extrapolating to non-FS 1 km squares. Simple metrics such as number of patches per km, average patch size, average patch area & perimeter, average patch area-perimeter ratio, and the largest patch index will be influenced by both how a 'patch' is determined in a parcel-based dataset and whether or not the truncated edge parcels of FS are included in the analysis. More complex metrics such as Landscape Shape Index, Mean Patch Fractal Dimension and Interspersion and Juxtaposition Index will also be influenced by how a 'patch' is defined, but have been shown to be largely unaffected by the issue of FS squares having truncated edge parcels (Swetnam, unpublished data).

2. Landscape structure and pattern are important attributes that may describe countryside character. Therefore it is important to derive meaningful landscape pattern metrics at a 1 km scale for all of England that can be used in the CQC project. These metrics must be independent of the issues raised above concerning truncated edge parcels in FS data and definitions of patches and boundaries in LCM2000. Metrics such as the Simpson's Index of Diversity, the Evenness Index, and the Bray-Curtis Similarity Index can be computed using the proportional composition of Broad Habitats per 1 km square. This would not be affected by patch dynamics or edge effects. Furthermore, this would have the combined benefit of being easy to derive from the LCM2000 data or the calibrated 1km summary dataset and would be conceptually linked as extra data layers in the final CIS release. (Information on these spatial metrics is enclosed in Appendix 1 below)

Conclusions

Through the process of meetings, presentations and email distribution the procedures developed and results produced by Module 9A have been communicated to a wide range of user groups. These procedures have in general been accepted and deficiencies identified for further considerations. This process has resulted in a clearer understanding of the issues associated with LCM2000, FS and their integration.

Ancillary data have been identified as a possible means of correcting deficiencies in the calibrated 1 km data set and validating the final results. A range of data sets have been obtained and extensive comparisons undertaken with FS and LCM2000. These ancillary datasets have generated their own problems and decisions have been made on quality, definition and coverage along with suitability for a role within Module 9A.

Datasets useful for additional knowledge-based correction within the calibration process have been identified and suitable rule bases developed and tested. This has resulted in two further iterations of the calibrated 1 km product to a point where it is

in its ‘final’ form with respect to this project. (However, one final iteration may be carried out to improve the quality of the calibrated calcareous grasslands if the EN Priority Habitats data can be acquired prior to public release).

A full validation will now be undertaken using a number of the ancillary datasets which have been described.

Uncertainty information will be derived for the calibrated 1 km data set by implementation of the full bootstrapping procedure.

A strategy for the measurement of landscape pattern has been developed and these values will now be derived for England at a 1 km level and the East Midland on a parcel by parcel basis.

References

Preston, C.D., Telfer, M.G., Arnold, H.R., Carey, P.D., Cooper, J.M., Dines, T.D., Hill, M.O., Pearman, P.D., Roy, D.B., & Smart, S.M. (2003). The Changing Flora of the UK. (DEFRA, London,).

Smith, G.M. and Hill, R.A. (2003), CS2000 Module 9: Data integration for localised results and support for indicators of countryside character and quality, Report A2/A3 - June 2003. Unpublished CEH report to DEFRA.

Watkins, J.W., Hill, R.A., Howard, D.C., Smith, G.M. and Cox, R. (2003), CS2000 Module 9: Data integration for localised results and support for indicators of countryside character and quality, Report A1 - Jan 2003. Unpublished CEH report to DEFRA.

Appendix 1

Simpson's Index of Diversity

$$D = \frac{1}{\sum p_i^2}$$

D = Simpson's Index
p_i = % cover of land cover types

Evenness

Evenness = D / s

$\boldsymbol{D}$ = Simpson's Index
$\boldsymbol{s}$ = species (land cover) richness

Bray-Curtis Similarity Index

$$I_{BC} = 1 - \frac{\sum |x_i - y_i|}{\sum (x_i + y_i)}$$

$\boldsymbol{I_{BC}}$ = Bray-Curtis measure of similarity; ranges from 0 to 1
$\boldsymbol{x_i}$ = % cover of a land cover type in one square
$\boldsymbol{y_i}$ = % cover of the same land cover type in an adjoining square

DIVERSITY
Simpson Diversity Index (D = 1 → s)

- *Gives more weight to common species (land cover types)*
- *D = s when all species (land cover types) are equally represented*

EVENNESS
Species Evenness – ("Simpson" 1/s → 1)

SIMILARITY
Bray-Curtis Similarity Index (IBC: 0 → 1)
0 = no species shared, 1 = all species shared
Based on species abundance similarity
Uses density or % cover, NOT relative density or relative % cover