The

Centre for Ecology & Hydrology

Centre for
Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

The **Centre for Ecology and Hydrology** is one of the Centres and Surveys of the Natural Environment Research Council (NERC). It was established in 1994 by the grouping together of four NERC Institutes, the Institute of Hydrology (IH), the Institute of Terrestrial Ecology (ITE), the Institute of Freshwater Ecology (IFE) and the Institute of Virology and Environmental Microbiology (IVEM). In 2000, the four component institutes were merged into a single research organisation.

The CEH mission

- To advance the sciences of ecology, environmental microbiology (including virology) and hydrology through high-quality and internationally recognised research leading to a better understanding and quantification of the physical, chemical and biological processes relating to land and freshwater and living organisms within these environments.

- To investigate, through monitoring and modelling, natural changes in the ecological, microbiological and hydrological environments, to assess both past and future changes and to predict man's impact on these environments.

- To secure, expand and provide ecologically and hydrologically relevant data to further scientific research and provide the basis for advice on environmental conservation and sustainable development to governments and industry.

- To promote the use of the Centre's research facilities and data, to provide research training of the highest quality and to enhance the United Kingdom's research base, industrial competitiveness and quality of life.

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 A2/A3 – May 2003

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

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

CONTRACT
No. CR0263

EXECUTIVE SUMMARY

This is the second report of the CS2000 Module 9 which presents 9A project activity and represents the deliverable against milestones A2 and A3.

The project has now reached the end of the Feasibility phase and the Operational phase is just beginning. The work completed has involved developing and testing methods of integrating FS and LCM2000, demonstrating the method for England and producing a prototype calibrated 1 km data set for England.

Correspondence analysis has determined calibration matrices between FS and LCM2000. The results have been examined in detail and at a range of spatial scales and for different spatial zonations.

The correspondence analysis has indicated the relative strengths and weaknesses of FS and LCM2000. The nature and spatial pattern of inter-class confusion between BHs has been examined and is now better understood.

Stratification using the National Land Classes (NLC) has been assessed as a method of off-setting the imprints in LCM2000 of the satellite image sampling framework. The NLCs can be merged as appropriate to maintain the statistical representativeness of FS statistics per zone.

The calibration procedure uses the matrices from the NLC stratified correspondence analysis to generate prototype calibrated 1 km data set. The calibration matrices have been used to scale the results of LCM2000 for a particular NLC to resemble those that would be derived by a comprehensive field survey.

The devised method incorporates the strengths of FS and LCM2000 to minimise the weaknesses of these two datasets. Thus coastal, urban and montane masks have been used to control the spatial application of the calibration procedure as appropriate.

The proposed method of calibration was discussed and accepted by the Module 9 Technical Advisory Group, in April 2003.

The first round of calibration iterations have generated a value-added product in which there has been an obvious shift away from the presence of image pair boundaries in the data set to the presence of NLC boundaries.

The use of NLC average calibrations has tended to 'distribute' BHs more evenly across an NLC zone. This reduces the spatial detail present in the original data whist calibrating the total BH estimates.

The prototype calibrated 1 km data set was launched at the user-oriented Module 9 Seminar in Bristol in May 2003 and will be made available on the Module 9 and CQC web sites in CIS format.

Further iterations are needed to address unresolved issues in the calibrated data set. This includes the effect of removing features that are anomalous within a zone or poorly sampled by the FS squares within that zone. This stage is now awaiting feed-back from the user community following the launch of the prototype calibrated 1 km data set.

Module 9 will now move into the operational phase in which the calibration procedure will be refined based on user feedback and completed, supplying calibrated 1 km data with information on i) aggregated land cover, ii) uncertainty and iii) pattern and structure information. This will feed directly into CQC for integration with character areas and typology and provide input to CS2000 Module 9B for the development of indicators.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

This is the second 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 and Land Cover Map 2000 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

The project has now reached the end of the Feasibility phase and the Operational phase is just beginning. The only change to the above list of activities is the delay in the production of estimates for the English Region until the calibrated 1 km data set has been finalised after refinement of the method. The production of a calibrated 1 km data set and regional estimates for Scotland and Wales awaits these countries joining the CS2000 Module 9 group.

This report will provide the background to the Module 9A work, a description of the main areas of activity and their results, report on the Technical Advisory Group meeting and Seminar and outline the way forwarded into the Operational Phase.

BACKGROUND

CS2000 and LCM2000

The Land Cover Map 2000 (LCM2000) was a module within the Countryside Survey 2000 (CS2000) project.

The main component of CS2000 was a Field Survey (FS) module that recorded a stratified sample of 569 1 km squares. The strata were the 40 National Land Classes (originally 32 ITE Land Classes); an environmental regionalisation based on physical geographical variables. The FS recorded areal features (e.g. fields), linear features (e.g. hedges) and point features (e.g. ponds) in great thematic and spatial detail, using 1:10 000 Ordnance Survey (OS) maps as the base; associated species (mostly plants) were also recorded. The main characteristics ('primary codes') denoted the type of feature (e.g. a wheat field, a hawthorn hedge, an individual tree); secondary codes recorded qualifying information (e.g. about species and cover, feature-size and management). It was possible to combine primary and secondary codes in an almost infinite variety of ways, to record some of the true complexity of the countryside. The information was necessarily simplified to generate the basic 'widespread Broad Habitat' (BH) classification of CS2000. BHs were based on selected combinations of primary and secondary codes, using objective rules. Further subdivisions were possible: for example, the CS1990 'baseline classes' give an objectively based, consistent, tried and tested classification.

The LCM2000 was based on the analysis of satellite image data with a spatial resolution of 25 m and provided a comprehensive map of widespread BHs. LCM2000 used image segmentation to identify relatively uniform areas within the images that were essentially distinct land parcels (e.g. fields, water bodies, urban areas and mosaics of semi-natural vegetation). The LCM2000 land parcels, or segments, were held in a vector format similar to the FS data. The segments were classified using the spectral character of the image data (i.e. reflectance, often from two different seasons). Enhancements were provided by knowledge-based corrections driven by ancillary data (e.g. elevation, soil sensitivity). LCM2000 used a hierarchical classification scheme consisting of 16 target classes, which were further subdivided to make 24 subclasses, with these in turn subdivided to give up to 72 class-variants. Most BHs were themselves target classes, though some were defined at the subclass level.

Calibration of Field Survey and Land Cover Map 2000

The 569 FS 1 km squares and the equivalent LCM2000 sections were inter-compared as part of the LCM2000 production programme to get a broad picture of LCM2000 map accuracy and to allow the generation of BH cover statistics at the national level, equivalent to those of the FS. The FS data were not 'ground truth': a quality assurance sample-survey recorded 88% agreement for re-survey of the original primary codes. In the absence of 'ground truth', the process of inter-comparison was one of 'calibration' where the FS and LCM2000 were quantitatively related. When differences occurred, the source or explanation was not always obvious without further analysis and interpretation.

These comparisons between FS and LCM2000 were raster-based at 2.5 m spatial resolution. The first step assessed the need for a shift in x- and / or y-directions of the LCM2000 data relative to the FS data. Further analysis then used, where appropriate, the shifted data set. The comparisons generated correspondence matrices, one for each FS 1 km square. Correspondences were recorded per-pixel (direct) and per-segment (LCM2000 segment labelled with FS data) and per-parcel (FS parcel labelled with LCM2000 data). To provide confidence limits for the measures of correspondence, a 'bootstrapping' procedure was adopted.

LCM2000 segments, compared with FS parcels, showed a basic correspondence of 63.4% in per-parcel comparisons at BH level (allowing for the FS generalisation of *Built up areas* and the LCM2000 omission of *Boundary and linear features* and *Rivers and streams*). As correspondence cannot realistically exceed the 88% repeatability of the FS, LCM2000 seems to be scoring at least 72% of its maximum potential. About 5% of the mis-match is explained by the 25 m grid underlying the image parcels, compared with the continuously variable structure of the FS. (If the field data are resampled onto the 25 m grid, the results show 95% correspondence with the original input data). The 0.5 ha Minimum Mappable Unit (MMU) of LCM2000 contrasts with the 0.04 ha MMU of the FS and explains many of the differences, especially for BHs which occur in less extensive stands (more than 4% of the area recorded by FS comprised parcels, not linear features, which were below the LCM2000 MMU). Time-differences explain other mis-matches: the FS was predominantly undertaken in 1998; LCM2000 used images mainly from 1998-2001. Evidently up to 15% of differences can be explained by the underlying structure of LCM2000 and, additionally, by its coarser MMU, and by date-differences. This suggest that LCM2000 may record Target classes with 87% success; to quote a figure of *c.* 85% accuracy at Target class level seems realistic.

For a given National Land Class (NLC) a single calibration matrix was produced by averaging the correspondence matrices for the FS squares that fall within it. A calibration could then made between the FS and LCM2000 by passing the LCM2000 proportions for the NLC through the calibration matrix. This process was repeated for each NLC and the results combined to given calibrated regional estimates of land cover. Such estimates were produced for Great Britain, England, Scotland, Wales and the six Environmental Zones. These results can be found in the Final Report of the Land Cover Map 2000 project (Fuller et al., 2002; http://www.cs2000.org.uk/Final_reports/M07_final_report.htm).

The process and procedures outlined above will be described in more detail in the following sections.

CALIBRATION METHODS FOR SMALL REGIONS

The calibration procedures developed during the LCM2000 production can be applied at a range of spatial scales / resolutions, although in this instance they will be tailored to provide results at a 1 km^2 spatial resolution.

This work is not simply a repeat of the LCM2000 production calibration as many issues that are masked at the national and Environmental Zone level become significant when working at a 1 km spatial resolution. The developments here therefore have three main components that consider i) the correspondence between the FS and LCM2000 results, ii) the stratification used to structure the calibration procedure spatially and iii) the calibration procedure itself. The components were dealt with in parallel at the beginning of the work, but were later merged as the calibration was refined.

The main components can be described thus:

i. A correspondence analysis was used during LCM2000 production to determine the calibration matrices between LCM2000 and the FS. The same approach was used during this project, but the results were examined in more detail and at a range of scales.
ii. The stratification using the NLC was assessed in the light of the fact that the calibration is between FS data and data derived from satellite images. The FS data were collected in support of the NLC, while the satellite image data collections were controlled by the satellites orbital parameters. This may have resulted in an alteration of the stratification as long as the calibration matrices were still statistically valid.
iii. Calibration was based on the simple approach used to generate the regional estimates of BH. This approach used the calibration matrices to scale the results of LCM2000 for a particular region to resemble those that would be derived by a comprehensive FS. The simple approach was enhanced in a number of ways to accommodate problems identified during LCM2000 production and from i. and ii. above.

Correspondence between LCM2000 and FS

Correspondence analysis was used to understand the similarities and differences between LCM2000 and FS for each of the 569 FS squares. This study used correspondence matrices generated by per-pixel comparisons; a direct overlay, with no regard for the structure of either dataset. FS parcels and LCM2000 segments were sampled onto a grid with a 2.5 m cell-size. To accommodate residual errors in the geo-registration of satellite images, the LCM2000 data were shifted to improve alignment. The correspondence analysis operated using shifted extracts (where appropriate) from LCM2000 Release 1. The overall mean shift distance was 53 m, with 48 % of squares shifted one pixel (25 m) or less in x- and or y- directions and 62 % shifted two pixels or less. Per-pixel scores of correspondence between the two data sets (160 000 samples at 2.5 m) were tabulated for each 1 km square. Table 1 shows an example of a simplified correspondence matrix. Values on the diagonal represent samples where the LCM2000 and FS agree, while those off the diagonal show confusion.

	FS									
LCM2000	Broadleaved, mixed and yew woodland	Coniferous woodland	Arable and horticulture	Improved grassland	Semi-natural / rough grass & bracken	Montain, heath & bog	Built up areas and gardens	Standing open water and canals	Coastal	Seas
Broadleaved, mixed and yew woodland	0.27	0.08	0.10	0.43	1.45	0.11	0.12		0.14	0.04
Coniferous woodland	0.11	3.36		0.19	0.96	0.01	0.05	0.02	0.54	0.18
Arable and horticulture	0.03		2.22	1.92	0.30		0.17		0.05	
Improved grassland	1.46	0.13	1.03	27.57	4.53	0.11	0.81		0.73	
Semi-natural / rough grass & bracken	1.64	0.05	0.20	4.70	4.92	0.36	1.35	0.01	1.15	0.09
Montain, heath & bog	0.34		0.01	0.36	3.39	1.89	0.50		1.06	0.06
Built up areas and gardens	0.21			0.44	0.66		3.71		1.21	0.03
Standing open water and canals					0.13	0.12				
Coastal	0.08		0.05	0.22	0.03		0.03		2.35	0.27
Seas	0.01			0.00	0.14	0.19	0.02		7.66	11.15
										57.43 Overall

Table 1. An example correspondence matrix for a single FS square, using aggregated classes for clarity on the page.

The overall correspondence statistics for the 569 squares are summarised in Figure 1. For mapping the 22 BHs, the range of correspondence between FS and LCM2000 for individual squares was 0% to 98%, with a mean of 53%. The modal percentile range was 70-80% correspondence. The CS2000 squares with lowest correspondence were frequently found in upland areas, where the ancillary data used in the knowledge-based corrections were insufficient to distinguish accurately between Dwarf shrub heath and Bog.

A surface representing the correspondence between LCM2000 and FS across GB is shown in Figure 2. Each 1 km cell has an interpolated correspondence value based on the actual correspondence between FS and LCM2000 for surrounding FS squares. This map was derived from an Inverse Distance Weighted spatial interpolation of the overall correspondence value for all 22 BHs for each FS square. The value in each interpolated cell will be determined by the correspondence of and distance to the nearest FS squares. The location of some of the FS squares can easily be seen in Figure 2 where they have an anomalously low (pale pink) or high (dark red) correspondence compared with surrounding values. This interpolated surface demonstrates a general gradient, with higher correspondence in the managed arable and pastoral landscapes of the south east and lower correspondence in the uplands of the north and west. Areas of low correspondence may reflect a high level of inter-class confusion between two or three particular BHs or a general high level of BH inter-class confusion.

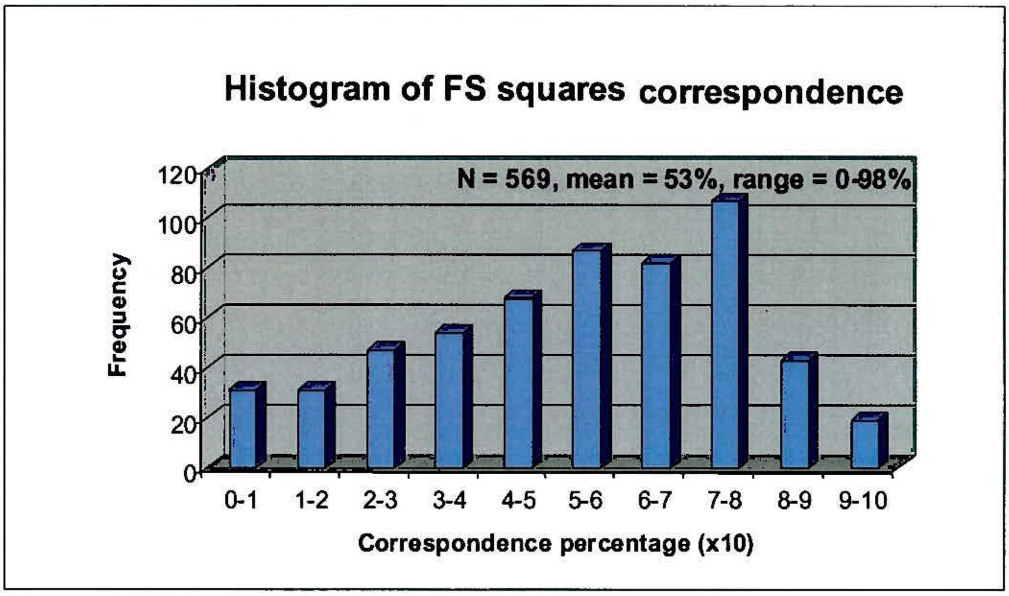

Figure 1. Correspondence between FS and LCM for mapping 22 Broad Habitats

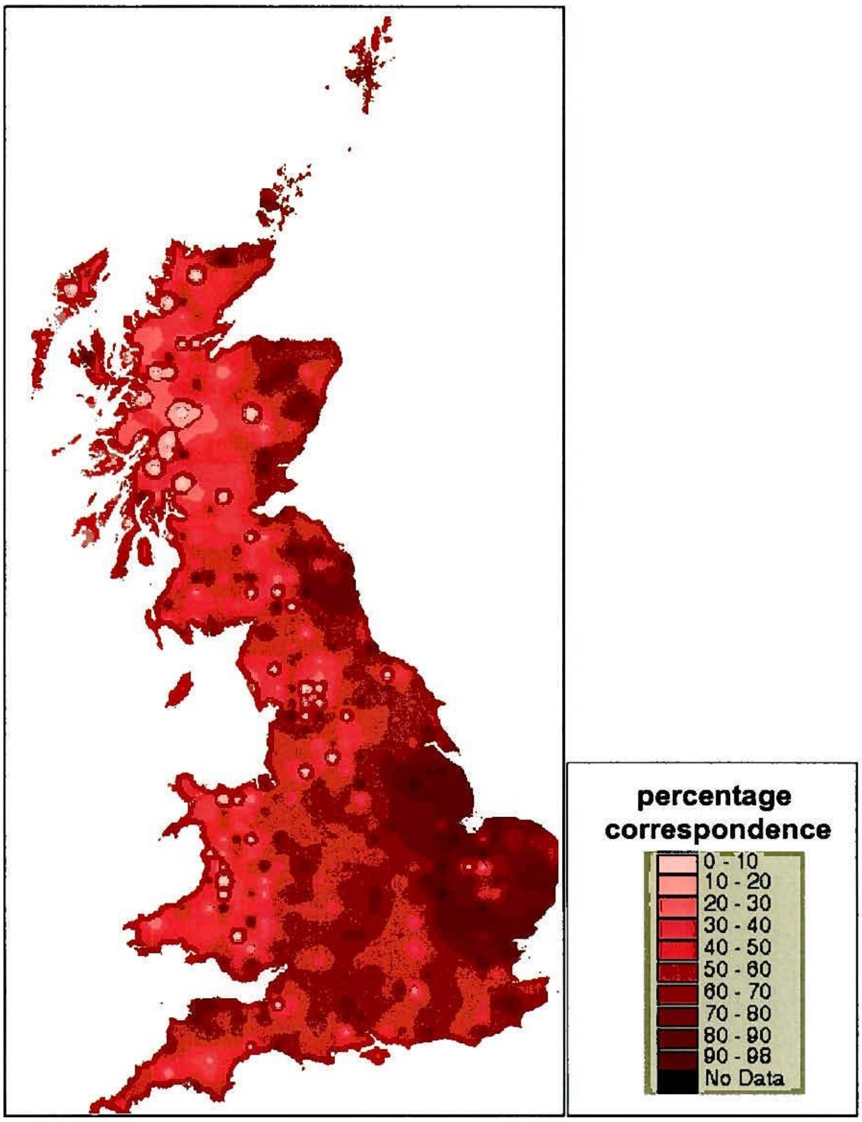

Figure 2. Interpolated correspondence between FS and LCM for mapping 22 Broad Habitats.

The differences in land-cover statistics for 22 BHs across GB as sampled by the FS and censused in LCM2000 are discussed in the Module 9 Report A1 (Watkins *et al.*, 2003). In summary, the following discrepancies were revealed:

Broad Habitats where LCM2000 estimates are lower than the 95% confidence limits of FS	**Broad Habitats where LCM2000 estimates are higher than the 95% confidence limits of FS**
Improved grassland	Neutral grassland
Bracken	Calcareous grassland
Fen, marsh & swamp	Dwarf shrub heath
Bog	Montane
Supralittoral rock	Inland rock
Supralittoral sediment	Built up & gardens*
Littoral rock	

* Note that the FS does not sample Built up & gardens within core urban areas

In general, the principal relative differences between FS and LCM2000 in terms of Broad Habitat mapping are:

- LCM2000 apparently over estimates Calcareous and Neutral grassland, at the expense of Improved grassland due to the problems inherent in mapping a continuum of grassland intensity and the use of a soil sensitivity mask with deficient class boundaries;
- LCM2000 apparently over estimates Dwarf shrub heath, at the expense of Bog due to problems with the peat mask used in knowledge-based correction;
- Some land-cover classes (e.g. Bracken, Fen, marsh & swamp, Supra-littoral and Littoral classes) are too rare or of too limited an extent to be recorded consistently in LCM2000 due to limitations of the training data and the MMU of 0.5 ha;
- Montane habitats were identified in LCM2000 using a decision rule (elevation > 600m) that was too generalised and based on coarse spatial resolution vegetation records;
- Inland rock was over estimated in lowland Britain due to spectral similarity with un-vegetated arable fields, and this could not be corrected by knowledge-based procedures as both inland rock (e.g. quarries) and un-vegetated fields can occur in a lowland context;
- FS does not sample within core urban areas and so extrapolated national statistics based on FS will inherently under estimate the spatial coverage of this land class. Within squares where FS does map urban, it may be over estimated as no distinctions are made for urban green space etc.

Broad Habitat inter-class confusion is more complex than the above GB areal estimates of land-cover would suggest. The patterns of BH inter-class confusion are also discussed in the Module 9 Report A1 (Watkins *et al.*, 2003). Across GB it is possible to find examples of inter-class confusion between virtually all BHs. However, the most frequently occurring examples of BH inter-class confusion are between Improved grassland and Arable & horticulture, Neutral grassland, and Calcareous grassland, between Neutral grassland and Calcareous grasslands, and between Bog, Acid grassland and Dwarf shrub heath.

The reasons for BH inter-class confusion can be summarised as:

1. Distinctions between FS and LCM2000:

- Different surveying dates, only half of the LCM2000 image pairs were from the FS 'target period' and therefore land-use rotation between crops and ley grass can create apparent non-correspondence;
- Different boundary positions and MMUs can result in non-correspondence and apparent patch effects when FS and LCM2000 are compared at a 2.5 m pixel scale;
- The FS and LCM2000 have different approaches to mapping within urban and woodland boundaries;
- Varying state of tides between FS and the time of satellite image acquisition can result in different extents of coastal BHs being mapped;
- When comparing FS and LCM2000 there is an issue of the representativeness of 569 FS squares, especially if sub-divided into a spatial stratification such as the NLC.

2. Misclassification in LCM2000 due to:

- Compromised image dates, early summer or late winter imagery can reduce spectral distinctions between certain land cover types that are strongest in mid-summer and mid-winter imagery;
- Spectral similarity occurs between land cover types (e.g. bare and un-vegetated land, different grassland types);
- Differing illumination levels due to aspect can cause increased spectral ranges of land cover types, increasing the chances of spectral confusion between land cover types;
- Varying detail and quality of ancillary data used in knowledge-based corrections resulted in some localised misclassification;
- Difficulty of identifying the rarer land cover types means these can often be under-represented in LCM2000.

Stratification issues

Local and regional scale patterns occur in the BH inter-class confusion described above resulting from the various boundaries in the combined data sets used to create LCM2000. These include:

- boundaries between satellite image pairs;
- boundaries within satellite image pairs resulting from summer or winter only data or local in-filling of cloud holes with LCMGB 1990 data;
- boundaries of ancillary data masks, such as soil sensitivity, peat depth, coasts;
- boundaries in the application of knowledge-based correction rules, e.g. thresholds in elevation, slope, vegetation indices, etc.

The results of correspondence analysis have been examined in detail at a range of spatial scales and for different spatial zonations. There are a number of stratifications that could be identified as the spatial framework for the calibration procedure. The nature of BH inter-class confusion and of apparent BH over or under estimation varies with different stratifications. As the stratification becomes increasingly multi-zonal so the number of FS squares per zone declines, rendering the calibration matrices less statistically representative. For example, 36 pairs of satellite images were used to cover GB in LCM2000. Within these 36 image pairs, the number of FS squares varies between a minimum of 1 and a maximum of 42. However, this stratification could be subdivided according to whether classification was based on summer-winter composite, summer-only or winter-only data, or whether the image pairs were within or outside the 'target period', etc.

Figure 3 shows some examples of the spatial variation in correspondence between FS and LCM2000 for sample BHs when stratified by image pair. For each BH shown, the correspondence between FS and LCM2000 has been calculated using all of the confusion matrices from CS squares located within the boundaries of each satellite image pair. For each BH shown, the correspondence is displayed per image pair and thus many of the image boundaries can easily be seen. In calculating and displaying the correspondence results in this way, the spatial variation between image pairs in the strength and direction of non-correspondence for individual BHs becomes apparent. The dark colours represent where LCM2000 has a higher proportion of a particular BH within an image pair compared with FS, whilst pastel colours represent where LCM2000 has a lower proportion of a BH. For the sample BHs shown in Figure 3, note that there is no consistent trend at the GB level in whether LCM2000 apparently over or under estimates percentage coverage compared with FS, or by what extent. However, regional patterns are more detectable, such as the south east to north west gradient of increasing apparent over estimation of Dwarf shrub heath in Scotland.

The spatial patterning of correspondence between FS and LCM2000 reported above is an artificial one resulting from the distribution of images dictated by the satellite orbiting parameters. Stratification using the NLC has been assessed as a method of off-setting the imprints in LCM2000 of the satellite sampling framework as it is more relevant to the biogeographical pattern of GB. The number of field survey squares per NLC varies between 6 and 30, and there is potential to combine the data for similar NLCs. The results of correspondence analysis based on the NLC stratification show a different spatial pattern and different values for apparent over or under estimation per zone compared with the satellite scene stratification. Nonetheless, the overall nature of BH inter-class confusion remains largely consistent between the satellite scene and NLC based stratifications. Thus, for both stratifications, compared with FS estimates LCM2000 frequently under estimates the percent coverage of Fen, marsh & swamp, Bog and Built up & gardens, and over estimates the percent coverage of Arable & horticulture, Neutral, Calcareous and Acid grasslands, and Dwarf shrub heath. However, in the satellite scene stratification LCM2000 also frequently under-estimates the percent coverage of Broadleaf, mixed & yew woodland, Improved grassland and Bracken compared with FS.

The calibration procedure will thus use the matrices from stratified correspondence analysis based on the NLC because they remove the imprints in LCM2000 of the satellite sampling framework. The NLCs are based on real landscape characteristics (such as topography and geology), and they can be merged as appropriate to maintain the statistical representativeness of FS statistics per zone.

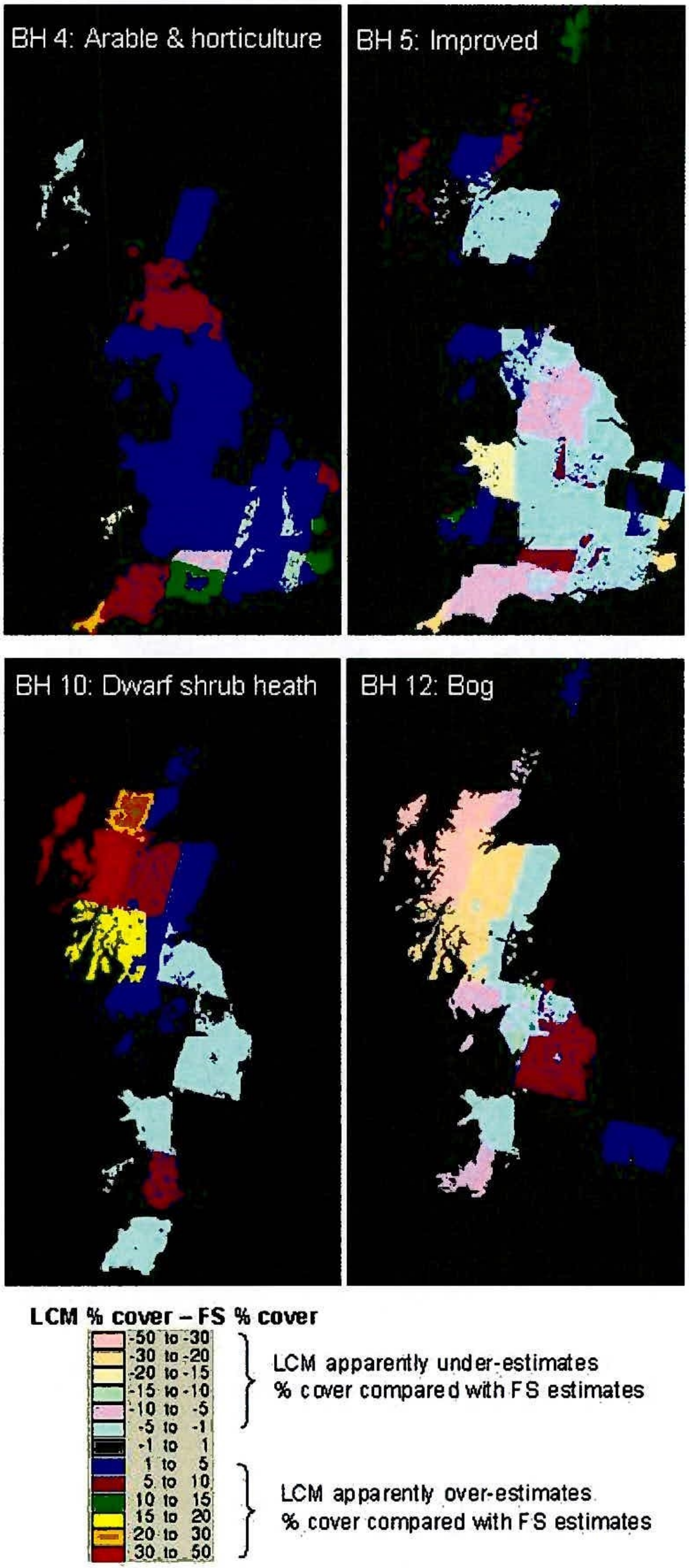

Figure 3. Correspondence between FS and LCM2000 per image pair.

Calibration method

The calibration method uses the correspondence matrices for a given NLC to scale the results of LCM2000 for each 1 km square within that NLC to resemble the results that would be derived by a compromise between the comprehensive LCM2000 and sampled based FS. The emphasis in the development of the calibration method has been on incorporating the strengths of both FS and LCM2000 and to minimise their weaknesses.

The calibration has to be more than the simple application of correspondence matrices (Figure 4). The results from the correspondence analysis identified areas where additional information was required to correct weaknesses and guided the formulation of a set of knowledge-based corrections. For instance, coastal and urban masks were required to control the spatial application of the calibration method as appropriate.

The initial step of the method is the production of a set of calibration matrices, one for each NLC. Each of the NLC calibration matrices is in fact the average correspondence matrix, derived from the set of correspondence matrices for each of the FS squares within the NLC. Thus,

$$A_k = \frac{1}{S}\sum_{l=S}^{1} M_l \qquad \textbf{Equation 1}$$

where A_k is the calibration matrix for NLC k, M are the individual correspondence matrices for the FS squares within NLC k and S is the number of FS squares within NLC k. Each element of the calibration matrix, A_{ij}, denotes the value for row i column j of the calibration matrix, i.e. the proportion of LCM2000 type i classified as FS type j.

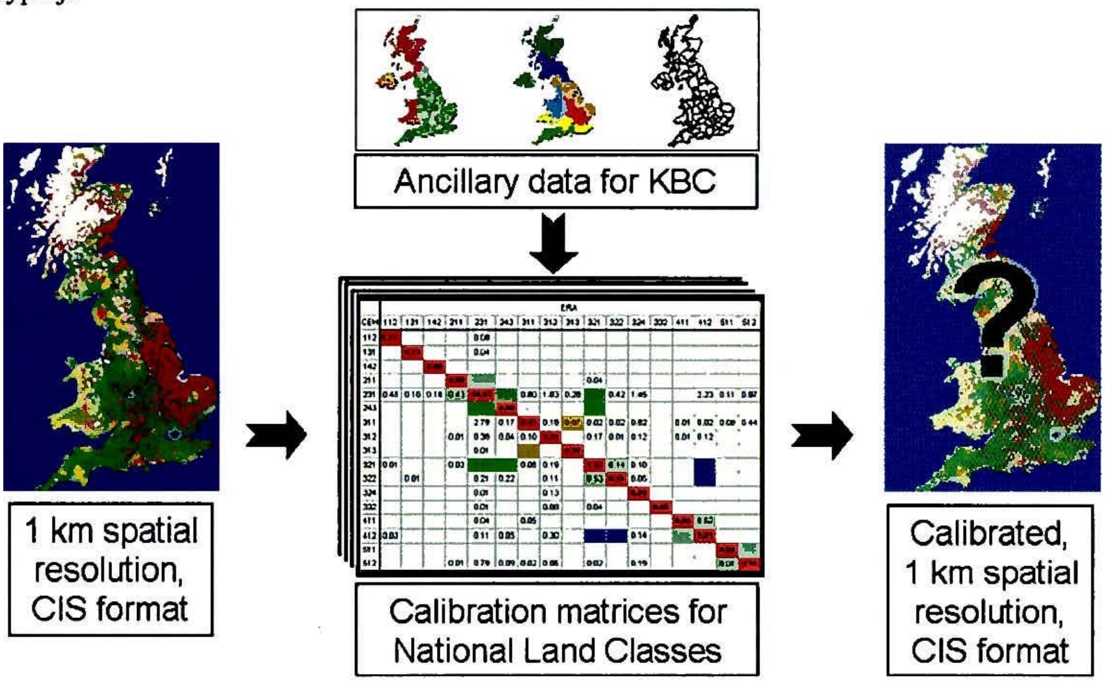

Figure 4. Calibration method to produce the calibrated 1 km data set from calibration matrices and ancillary data.

The application of the calibration matrices to the original 1 km data set uses the same procedure as that used for the regional estimates within the LCM2000 production. For a given 1 km square the LCM2000 data can be calibrated to FS equivalent values by multiplying the LCM2000 land cover proportions for that 1 km square by the calibration matrix for the NLC within which the 1 km square lies, i.e.

$$FS_m = A_k \times LCM_m \qquad \textbf{Equation 2}$$

where FS_m and LCM_m are vectors of the proportions of each BH within the mth 1 km square. LCM_m is the vector from the original LCM2000 data and FS_m is the resulting vector which forms the new calibrated data set.

As an example, consider this hypothetical calibration matrix:

	Field survey BH		
LCM2000 BH	***broadleaved***	***conifer***	***urban***
broadleaved	0.75	0.10	0.15
conifer	0.10	0.85	0.05
urban	0.05	0.05	0.90

Results have been normalised so values sum to '1' across the rows. It shows the proportions of each of the land cover types that was mapped as the same and different land cover type in the other survey. For instance, of the area mapped by LCM2000 as broadleaved, 75 % was also mapped as broadleaved by the FS, but 10 % was mapped by the FS as conifer and 15 % as urban. The calibration matrix allows the LCM2000 proportions to be altered by performing the following analysis:

		Field survey BH		
LCM2000 BH	***Values***	***broadleaved***	***conifer***	***urban***
broadleaved	*1000*	750	100	150
conifer	*500*	50	425	25
urban	*200*	10	10	180
Calibrated output values		**810**	**535**	**355**

The LCM2000 proportions of each BH (*Values*) are multiplied by the fractional amounts in the calibration matrix to give output proportions which, if summed (bold), show how the same 1 km square might have been recorded by a comprehensive FS.

The above procedure is valid only where the calibration matrix is fully representative of the landscape within the 1 km square which is being calibrated. The NLCs range in extent from just over 800 km^2 to in excess of 15000 km^2. With a maximum of 30 FS squares per NLC, not all of the spatially variability in the landscape within a NLC will be present in it's calibration matrix. Also, some of the NLCs straddle a number of landscape types which can not be mixed in a calibration exercise such as this. For instance NLC 8 is defined as 'Coastal, often estuarine, mainly pastures, otherwise built-up'. NLC 8 is found along the coast of The Wash, but also along the river courses that cross The Fens. Therefore a simple application of the calibration matrix for NLC 8 would produce coastal habitats along the river valleys of the Nene, Ouse and Welland.

To control the operation of the calibration matrix a number of knowledge-based corrections were developed. These can be divided into three groups based on their method of operation and the impact on the results.

The simplest knowledge-based correction works by disabling the calibration process completely where it worsens the results. For instance, the FS is not designed to map dense urban areas and the selection of FS squares specifically avoided areas with greater than 25 % urban. In this case, dense urban areas are therefore mapped most optimally by the LCM2000 data without calibration.

The second type of knowledge-based correction relates to 1 km squares where a particular class is known not to be present via some additional contextual information. In this case the output column of the calibration matrix for the class that is not present should be set to zero to prevent any of it being produced by the calibration. As the input rows of the calibration matrix will no longer sum to 1 it will be necessary to renormalise the calibration matrix to produce the correct total of the output proportions. This type of correction can be used with the example described above to prevent coastal habitats appearing in river valleys far from the sea.

The third type of knowledge-based correction relates to a known mis-classification within a particular spatial context which the calibration matrix at the NLC level cannot fully correct. In this case the column of values for the mis-classified class can be combined with the column of values for the correct class. This correction could be applied to grassland types where ancillary data, such as soils information, can identify which of the grassland types is correct.

The process for developing the calibration method was one of iteration (Figure 5). The 1 km summary data set from LCM2000 Release 1 was the starting point. The calibration method is applied, the calibrated results are produced and these results are validated. The results of the validation are then used to refine the calibration method and the process is repeated.

To date there have been four iterations to produce the current version of the calibrated data set:

1. Apply the calibration matrices without knowledge-based corrections,
2. As above, but with the calibration disabled in dense urban areas,
3. As above, but coastal habitats excluded outside the coastal zone, and
4. As above, but montane habitats excluded outside of a montane mask and urban excluded within it.

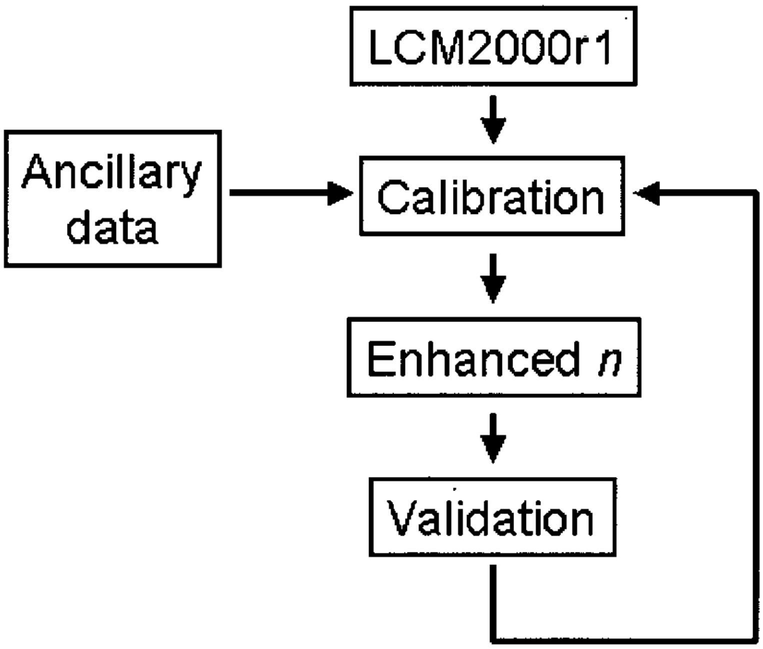

Figure 5. The iterative process for developing the calibration method.

Results of the first calibration iterations

The initial calibration procedures have caused an obvious shift away from the presence of image pair boundaries in the data set to the presence of NLC boundaries (Figure 6, bottom). The use of NLC average calibrations has tended to 'distribute' BHs more evenly across NLC zones. This reduces the spatial detail present in the original data whist calibrating the total BH estimates. For all the terrestrial BHs (except Calcareous grassland), the effects of the first calibration iterations have been to generate an increased spatial distribution ranging from slight to significant. Calcareous grassland (Figure 6, middle) has a significant reduction in spatial distribution.

Changes in the total coverage of BHs is more varied, with some increasing and others decreasing (Table 2). The range in the size of the change matches in many respects the summary results of the correspondence analysis. Fen, marsh & swamp shows the greatest change followed by a number of BHs known to have a limited extent. Those BHs with the smallest amount of change are those which both surveys have considerable success in mapping. Table 2 does also show a slight anomaly, as the total areas before and after calibration are marginally different. This will be explored in the refinement stage and is likely to be related to the use of NLCs which have a different extent to the LCM2000.

Broad Habitat	Area (km^2) Original	Calibrated	% change
Broadleaf woodland	10930	10610	-3
Coniferous woodland	2980	2750	-8
Agric. & horticulture	48390	46070	-5
Improved grassland	32020	38910	22
Neutral grassland	5000	4240	-15
Calcarous grassland	7850	440	-94
Acid grassland	2790	3930	41
Bracken	700	1400	100
Dwarf shrub heath	2650	2870	8
Fen, marsh & swamp	180	1230	583
Bog	1060	1060	0
Standing open water	590	520	-12
Montane habitats	0	0	+
Inland rock	1100	170	-85
Built up & gardens	13810	15500	12
Supralittoral rock	0	160	
Supralittoral sediment	110	380	245
Litoral rock	20	0	-100
Litorral sediment	3140	810	-74
Oceanic seas	1770	1160	-34
Total	135090	132210	

Table 2. Total coverage of England for all 22 BHs in LCM2000 Release 1 (left column) and the calibrated product (right column). Note Montane Habitats do increase but not sufficiently to be represented in the table.

The calibration procedure not only has the effect of 'smoothing' the spatial distribution of Broad Habitats across the NLC zones, but also of removing features that are anomalous within a zone or poorly sampled by the FS squares within that zone. This reflects the fact that the FS was designed to be representative for generating land cover statistics at the national or regional level, rather than for precise spatial mapping of land cover within NLC zones. An example of this is Thetford Forest, which is currently converted from a high percent coverage of Coniferous woodland in LCM2000 to a high percentage coverage of Broadleaf, mixed and yew woodland in the calibrated product (Figure 6, top). This occurs because the FS squares within the NLC zone containing Thetford Forest fail to provide an adequate sample of the Coniferous woodland present at Thetford. Other examples of local features that are not well represented in FS data at the level of NLC zones include Salisbury Plain, Bodmin Moor and the limestone scenery around Buxton.

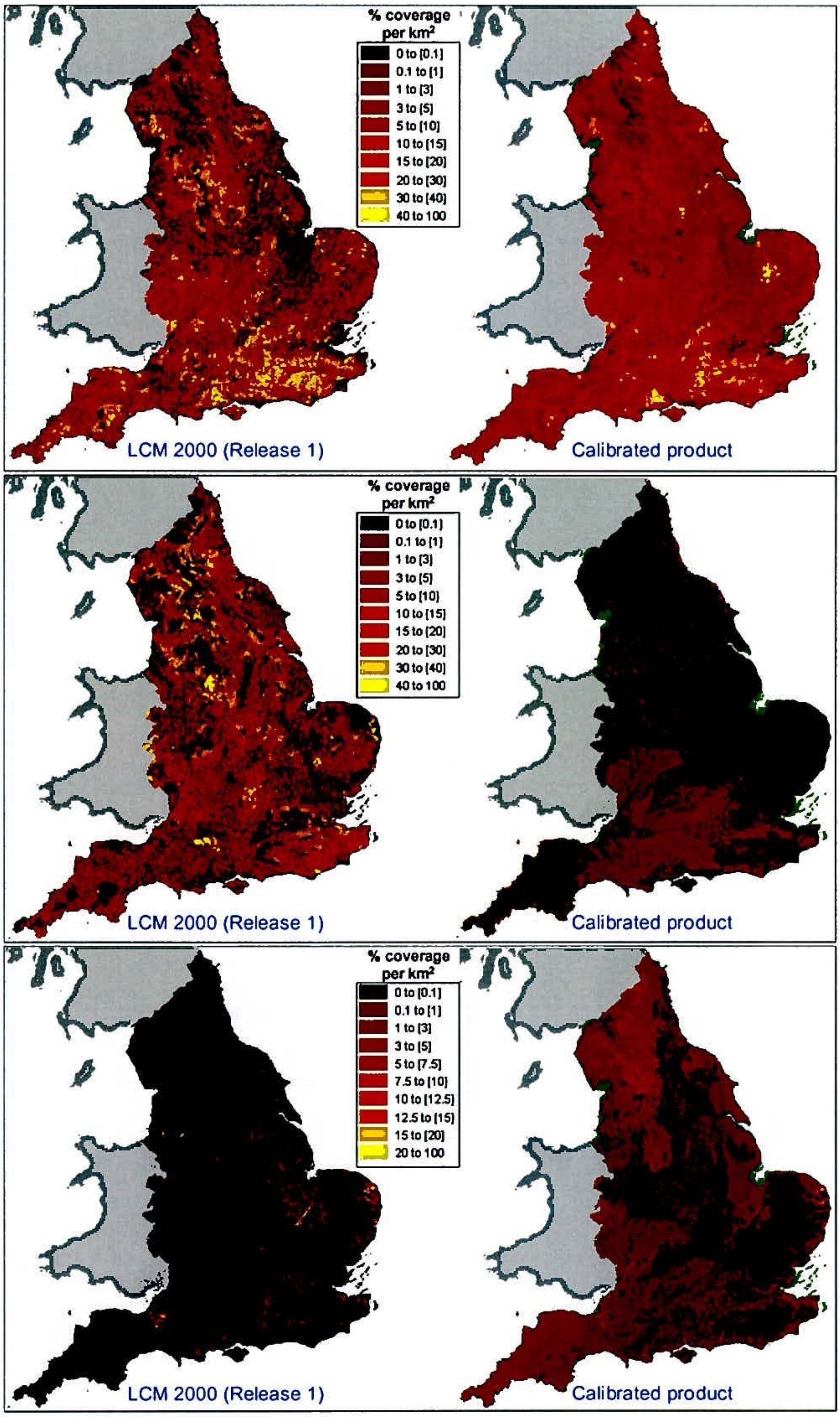

Figure 6. Examples of LCM2000 and the calibrated output for Broadleaf, mixed and yew woodland (top), Calcareous grassland (middle) and Fen, marsh & swamp (bottom).

At this point it would be useful to visualise the operation of the current set of knowledge-based corrections. The group of nine images in Figure 7 show original 1 km summary LCM2000 Release 1 data for a single BH in the left column, calibrated data using the calibration matrices only in the middle column and calibrated with knowledge-based corrections in the right column.

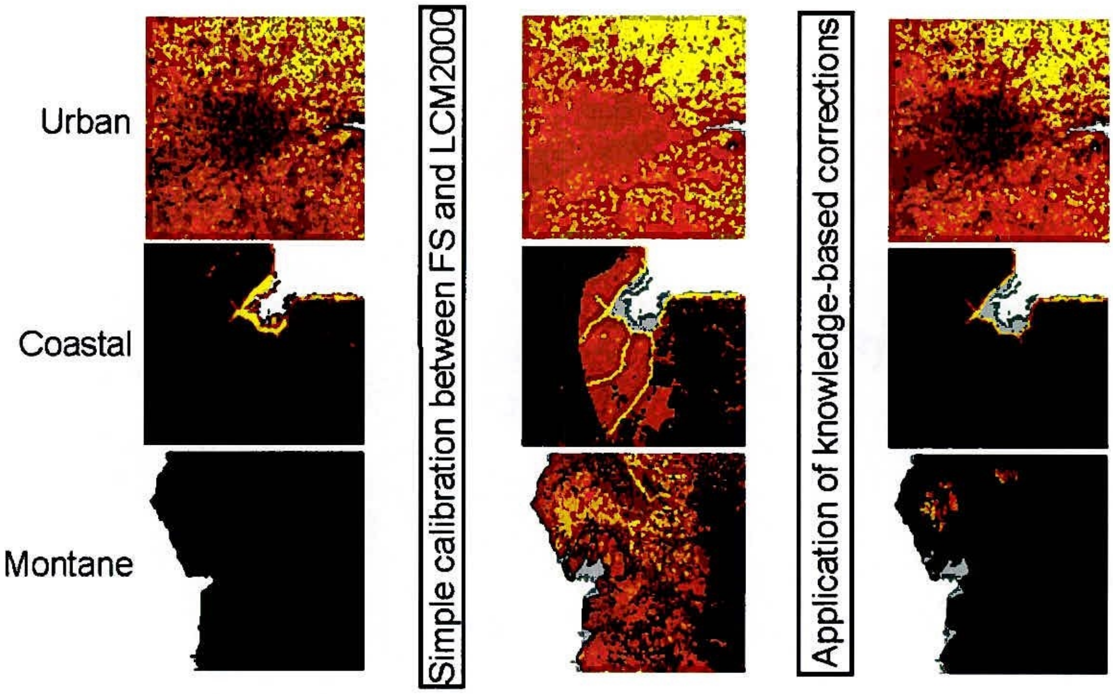

Figure 7. Visualisations of the operation of the knowledge-based correction rules for an Urban, Coastal or Montane context.

To demonstrate the urban knowledge-based correction, the Arable and horticulture data is displayed (top row) to emphasis the process. In the original LCM2000 data London is clearly defined with either zero or extremely low values for Arable & horticulture. The application of the calibration matrices incorrectly adds Arable & horticulture to the dense urban areas, which is removed when the calibration process is disabled in these areas. A similar effect occurs with coastal classes (middle row) being added to inland areas, but the application of a coastal mask restricts these classes to their true context. Also, the use of NLCs reduces the intertidal areas of The Wash since these have not been allocated to a NLC. In the final example above, LCM2000 did not map any Montane habitats in England (bottom row), although from species records these habitats are known to be present. The application of the calibration matrices suggests extensive areas of Montane due to its the mis-classification with extensive classes such as Dwarf shrub heath. A mask for Montane was generated from species data and a digital elevation model to limit the occurrence of this class to an appropriate context. The same mask was also used to exclude Built up & gardens from these areas.

REPORTS ON PROJECT MEETINGS

As part of the feasibility phase of module 9, two meetings were held to involve the wider research and user communities in the project.

The first meeting was of a Technical Advisory Group (TAG) (Annex 1) mid way through the Feasibility phase, where the rational and methods to be employed in Module 9 were put before a group of experts. The group's experience ranged from spatial statistics, through landscape ecology and data integration to remote sensing. The TAG were able to provide useful inputs to the methodological developments and approved the approach adopted by CEH.

The second meeting was a Seminar (Annex 2) at the end of the Feasibility phase where the prototype calibrated 1 km data set for England and the work so far on indicators was presented to the user community. The results were well received and support was given for evaluating the prototype calibrated 1 km data set within user organisations.

DISCUSSION

Module 9A has now reached a point where the user community needs to provide feedback to CEH to guide the final set of refinements, before the definitive calibrated 1 km data set is produced. Based on our analysis of the results, there are a number of issues that should be addressed in the refinement stage or noted as qualifiers against the data set.

- There is a loss of local spatial detail that the LCM2000 originally provided through the application of average calibration matrices within NLC zones.
- The smearing of BHs within zones is often at very low proportions, therefore a threshold could be applied and could be dependent on the overall extent of the BH.
- The calibration procedure has increased the general 'urbanness' of England since FS over estimates this Broad Habitat in FS squares.
- Neutral and Calcareous grasslands no longer match the soil sensitivity map, even at the broad scale. A new soils / grassland mask should be developed to deal with this issue.
- The development of more complex knowledge-based corrections will require the use of calibration coefficient merging rather than exclusions. These operations will be difficult to define in a consistent manner nationally and their use should be thoroughly tested.
- The NLC boundaries are now perhaps too obvious in the calibrated output.
- The use of other data sources will be explored; e.g. soils and species information for semi-natural grasslands, CORINE data for inland bare (arable / rock), OS rivers and lakes for water features.

Once the refinement process has been completed Module 9A will move on generate other integrated products.

It will firstly, examine the sources of error within the calibrated data set and attempt to provide measures of uncertainty. To calculate confidence limits for the calibrated LCM2000 estimates a 'bootstrapping' procedure was developed. The bootstrap samples are obtained by randomly sampling the correspondence matrices within the NLC, with replacement (i.e. such that the same square might have been drawn more than once or not at all), to obtain a number of samples equal to the total number of samples in the NLC. From the bootstrap samples a new calibration matrix is calculated and the calibration formula applied. This process is repeated 1000 times and mean BH proportions are calculated. The 1000 values are then ranked from the lowest to the highest levels and the 25th and 975th in rank taken to represent the '95-percentile range' (i.e. that encompassing 95% of all estimates). A bias-correction method is applied to correct the distribution for any asymmetry.

The LCM2000 1 km summary data set as used in this project and found in CIS is based on a simple aggregation of land cover statistics. Other metrics can be calculated for each 1 km cell, including the pattern, fragmentation, and connectivity of landscape features. The LCM2000 full resolution vector data provides a measure of landscape spatial structure derived from its land parcel structure. This does not exactly match the structure as recorded on the surface by FS because the LCM2000 land parcels are based on the spectral characteristics of the satellite image pairs. For instance, fields containing a crop with variable growth patterns may be subdivided as this could produce a significant spectral change. Using the FS 1 km squares it will be possible to compare the pattern recorded by FS and LCM2000. There are numerous possible indices of spatial pattern that can be measured, such as shape, area, connectivity, patch distribution or number, distance to the nearest feature or land cover of interest, or the number, type, spatial coverage and connectivity of features within a given radius. A selection of these will be produced from the segment structure within the LCM2000 at 1 km resolution. Two versions of the measures will be generated, one simply using LCM2000 data, the other derived from an integration of LCM2000 and FS with quantifiable measures of confidence.

CONCLUSIONS

The feasibility phase of Module 9A has been completed. This has involved developing and testing methods of integrating FS and LCM2000, demonstrating the method for England and producing a prototype calibrated 1 km data set for England.

Correspondence analysis has determined calibration matrices between FS and LCM2000. The results have been examined in detail and at a range of spatial scales and for different spatial zonations. The correspondence analysis has indicated the relative strengths and weaknesses of FS and LCM2000. The nature and spatial pattern of inter-class confusion between BHs has been examined and is now better understood. The reasons for non-correspondence between FS and LCM2000 can be categorised under i) the distinctions between the two data sets and their method of production and ii) genuine mis-classification.

Stratification using the NLCs has been assessed as a method of off-setting the imprints in LCM2000 of the satellite image sampling framework. The spatial pattern of satellite image pairs results from the satellite orbiting parameters, whilst the NLCs are based on real landscape characteristics (such as topography and geology). The NLCs have the extra advantage that they can be merged as appropriate to maintain the statistical representativeness of FS statistics per zone.

The calibration procedure uses the matrices from stratified correspondence analysis to generate a prototype calibrated 1 km data set. The approach has used the calibration matrices to scale the results of LCM2000 for a particular region to resemble those that would be derived by a comprehensive field survey. The emphasis has been on incorporating the strengths of FS and LCM2000 to minimise the weaknesses of these two datasets. Thus coastal, urban and montane masks have been used to control the spatial application of the calibration procedure as appropriate. The proposed method of calibration was discussed and accepted by the Module 9 Technical Advisory Group, in April 2003.

The first round of calibration iterations have generated a value-added product in which there has been an obvious shift away from the presence of image pair boundaries in the data set to the presence of NLC boundaries. The use of NLC average calibrations has tended to 'distribute' BHs more evenly across an NLC zone. This reduces the spatial detail present in the original data whist calibrating the total BH estimates. This prototype calibrated 1 km data set was launched at the user-oriented Module 9 Seminar in Bristol in May 2003 and will be made available on the Module 9 and CQC web sites.

Further iterations are needed to address unresolved issues in the calibrated data set. This includes the general effect of removing features that are anomalous within a zone or poorly sampled by the FS squares within that zone. This stage is now awaiting feed-back from the user community following the launch of the pilot data set.

Module 9 will now move into the operational phase in which the calibration procedure will be refined based on user feedback and completed, supplying a calibrated 1 km data set with information on i) aggregated land cover, ii) uncertainty and iii) pattern and structure information. This will feed directly into CQC for integration with character areas and typology and provide input to CS2000 Module 9B for the development of indicators.

REFERENCES

Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, R.A., Thomson, A.G., Cox, R., Brown, N.J., & Gerard, F.F. (2002) *Countryside Survey 2000 Module 7: Land Cover Map 2000. Final Report, CSLCM/Final.* 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.

ANNEX 1

CS2000 Module 9
Data Integration For Localised Results And Support For Indicators Of Countryside Character And Quality
Technical Advisory Group
First Meeting 4th April, 2003 Reading University
Minutes/Notes

Attendees

Peter Atkinson	University of Southampton
Lex Comber	University of Leicester
Peter Fisher	University of Leicester
Geoff Griffiths	University of Reading
Ross Hill	CEH Monks Wood
David Howard	CEH Merlewood
Margaret Oliver	University of Reading
Paul Robinson	University of Leicester
Peter Rothery	CEH Monks Wood
Geoff Smith	CEH Monks Wood
Andrew Stott	Defra
John Watkins	CEH Merlewood
Barry Wyatt	CEH Monks Wood

Background

AS outlined Defra's objectives and requirements from the project; a dataset is needed that will allow changes in landscape features to be measured by habitat at a national scale. By integrating the two CS2000 datasets (field survey sample and census satellite map) it was hoped that more sense could be made of the results and they could be subdivided with confidence to a regional scale. JWW then described the need for the data to contribute to the production of an indicator of landscape change and the relationship between Module 9 and the Countryside Quality Counts (CQC) project.

Integration of LCM2000 and CS2000 FS data

GS and RH made presentations of the approach and progress towards a unified dataset. The presentation included a summary of the work done in the LCM2000 generation project, the initial comparison of LCM2000 and CS2000FS and the proposed calibration procedure. The TAG accepted the approach and appreciated the work carried out so far but made several suggestions and recommendations:

1.1 Examine the geographic distribution of differences in correspondence between the datasets and relate them to the image pairs. Make a comparison between image pairs that are contemporary to CS2000FS and those that are not. Investigate the use of the image pair identifier in the calibration process. The correspondence values generated should be kriged to give a correspondence surface.

1.2 Investigate a calibration procedure adjusted by the confidence level of each data component [within each land class and/or image pair]. The correspondences could be interpolated to the 1 km level and used as a scale.

1.3 Separate random noise from directional bias in the correspondence. Start by reworking the correspondence tables with a threshold higher than 3%.

1.4 Identify ways of determining the confidence in different levels of zonation and propose a limit below which the data should not be subdivided. The use of the meta-data in LCM2000 (especially the class probability information) should be investigated; measures of confidence in BH classification could drive the calibration process.

Development of indicators for countryside character and quality

DH described how provisional indicators derived from landscape characteristics were being generated and applied. Countryside Character Areas were being used as units and amalgamated into regions using stock and change statistics. An initial investigation of kriging CS2000 FS data to generate surfaces that can be subdivided to produce CCA statistics was presented. LCM2000 was being used as a mask to define the limits of extent and the work was intending to look at co-kriging using LCM2000. The TAG felt the analysis was interesting, not valid in all cases and worthy of taking on in different directions. A number of specific suggestions and recommendations were made:

2.1 Geostatistical analysis could assess the effectiveness of the CS2000 FS sample and guide the selection of additional sites.

2.2 The relationship between the kriging standard error and the predicted values should be investigated. It was probably due to the transformation/back transformation that had been applied but other problems need to be ruled out.

2.3 The variogram can be used to constrain the calibration/classification analysis.

2.4 The estimates and distributions generated should be compared to the LCM2000 statistics and the performance of the geostatistical approach be compared with that of the ITE Land Class. The ITE LC and sample strategy were designed to provide national estimates. Now that local estimates are being derived from the same information then a new approach will be necessary.

2.5 Other styles of kriging should be investigated (co-kriging, regression kriging, etc).

2.6 Surface generation describing change should be carefully examined. The two options are to generate the best surface for each year and measure the difference or to identify the differences between the sample dates and generate a single surface from one data set. The best approach will be dependent on the covariance of the two data sets.

Discussion

Module 9 is concerned with indicator development, but will not be involved in the assessment of quality derived from any indicator. That task will be performed by the CQC. However, information pertaining to the quality of the information and derived indicators will need to be identified for the CQC. Information is also needed to identify the most appropriate data set to use to generate statistics for different size regions. At present LCM2000 is used for local, CS2000 FS for national and both for sizes in between. The generation of alternative estimates is only a disadvantage when it is unclear which statistic to use in which situation, although discrepancies between the data sets always need explaining.

There may be different levels of confidence in maps for different Broad Habitats generated using the same approach. It is possible that some Broad Habitats (e.g. Fen, Marsh and Swamp) cannot be mapped with any confidence using either or both data sets. Other ancillary datasets may need to be considered. Alternatively, the BH classes could be aggregated in a similar manner to the LCM2000 production final reporting. It is important that the Broad Habitats are reviewed in the light of the performance of the two approaches.

It was agreed that there was value in the TAG meeting again to review progress on the suggestions and recommendations. Members of the TAG were willing to be involved in discussions outside the meeting to offer specific advice.

The Centre for Ecology and Hydrology has 600 staff, and well-equipped laboratories and field facilities at nine sites throughout the United Kingdom. The Centre's administrative headquarters is at Monks Wood in Cambridgeshire.

This report is an official document prepared under contract between the customer and the Natural Environment Research Council. It should not be quoted without the permission of both the Centre for Ecology and Hydrology and the customer.

CEH Sites

CEH Directorate
Monks Wood, Abbots Ripton,
Huntingdon,
Cambridgeshire PE28 2LS
Telephone +44 (0) 1487 772400
Main Fax +44 (0) 1487 773590

CEH Oxford
Mansfield Road,
Oxford,
Oxfordshire OX1 3SR
Telephone +44 (0) 1865 281630
Main Fax +44 (0) 1865 281696

CEH Merlewood
Windermere Road,
Grange-over-Sands,
Cumbria LA11 6JU
Telephone +44 (0) 15395 32264
Main Fax +44 (0) 15395 34705

CEH Bangor
University of Wales, Bangor,
Deiniol Road,
Bangor, Gwynedd LL57 2UP
Telephone +44 (0) 1248 370045
Main Fax +44 (0) 1248 355365

CEH Wallingford
Maclean Building, Crowmarsh Gifford,
Wallingford,
Oxfordshire OX10 8BB
Telephone +44 (0) 1491 838800
Main Fax +44 (0) 1491 692424

CEH Dorset
Winfrith Technology Centre,
Winfrith Newburgh, Dorchester,
Dorset DT2 8ZD
Telephone +44 (0) 1305 213500
Main Fax +44 (0) 1305 213600

CEH Edinburgh
Bush Estate,
Penicuik,
Midlothian EH26 0QB
Telephone +44 (0) 131 4454343
Main Fax +44 (0) 131 4453943

CEH Windermere
The Ferry House, Far Sawrey,
Ambleside,
Cumbria LA22 0LP
Telephone +44 (0) 15394 42468
Main Fax +44 (0) 15394 46914

CEH Monks Wood
Abbots Ripton,
Huntingdon,
Cambridgeshire PE28 2LS
Telephone +44 (0) 1487 772400
Main Fax +44 (0) 1487 773467

CEH Banchory
Hill of Brathens,
Banchory,
Aberdeenshire AB31 4BY
Telephone +44 (0) 1330 826300
Main Fax +44 (0) 1330 823303

Further information about CEH is available on the World Wide Web at www.ceh.ac.uk

NATURAL
ENVIRONMENT
RESEARCH COUNCIL