Policy of the Countryside Survey partners on access to survey square locations (http://www.countrysidesurvey.org.uk/square-access-policy)

CS square locations are considered confidential to preserve the representativeness of sampling sites and the goodwill of landowners. These are both essential elements to the future of the survey to ensure the scientific integrity of the sampling strategy, the protection of the environment, and help to ensure future permission from landowners to survey their land. Future surveys would be compromised if either of these elements were to be jeopardised and our capacity to reliably inform environmental policy would be diminished.

We believe our position on confidentiality to be in the public interest. Any requests for this information will be dealt with by CEH, on behalf of NERC and the CS partners, under the terms of the UK Freedom of Information Act 2000 (The Act) or the associated Environmental Information Regulations 2004 (The Regulations).

Spatial data from within survey squares (e.g. maps of habitats, linear features) could be used relatively easily to identify the location of a square. It is therefore considered to carry equivalence to the location data and will only be released under the same exceptional circumstances and terms as the six-digit grid references location data.

The numbers of individual survey squares and the figures containing images of those squares have therefore been redacted in the publicly available version of this report.

Ian McCulloch
CEH Archives and Records Coordinator
17th June 2011

CENTRE FOR ECOLOGY AND HYDROLOGY
(NATURAL ENVIRONMENT RESEARCH COUNCIL)

Project. C00324 (T02083j5)

Countryside Survey 2000 - Part I

Module 8: Airborne Scanner Applications

Classification of airborne CASI and LIDAR data of selected CS2000 sample squares

Fourth Interim Report

CSCL/Int4

R.A. Hill[1,] R.M. Fuller[1] & N. Veitch[2]

[1]Centre for Ecology and Hydrology, Monks Wood,
Abbots Ripton, Huntingdon, Cambridgeshire, PE28 2LS

[2] The Environment Agency, National Centre for Environmental Data and Surveillance,
Rivers House, Lower Bristol Road, Bath, Avon, BA2 9ES

Corresponding author:
R.A. Hill
Section for Earth Observation
Centre for Ecology and Hydrology
Monks Wood
Abbots Ripton
Huntingdon
Cambs PE28 2LS

Telephone: 01487 772 485
Fax: 01487 773 467
Email: rhill@ceh.ac.uk

June 2000

CONTENTS

1. EXECUTIVE SUMMARY 3

2. INTRODUCTION 5

3. IMAGE ACQUISITION 5
    3.1 Flight dates and specifications 5
    3.2 Data quality 5

4. METHODS DEVELOPMENT 6
    4.1 LIDAR data pre-processing 6
    4.2 CASI data pre-processing 8

5. CLASSIFICATION 10

6. KNOWLEDGE-BASED CORRECTION 11
    6.1 Phase-1 KBC procedure 11
    6.2 Phase-2 KBC procedure 14

7. VALIDATION 15

8. RESULTS OF VALIDATION EXERCISE 16

9. PROGRESS MEASURED AGAINST SCHEDULE 17

10. REFERENCES 18

Figures 1-9 20 *et sq.*

1. EXECUTIVE SUMMARY

- This project is evaluating the use of a Compact Airborne Spectrographic Imager (CASI) and LIght Direction And Ranging instrument (LIDAR) in Countryside Survey 2000, for environmental monitoring at an extent and scale of survey which is intermediate to the field and satellite surveys.
- Pairs of example survey squares are being studied in each of the Arable, Pastoral, Marginal and Upland Landscapes of GB - as defined in Countryside Survey 1990. Each pair has been divided into a *trial* and a *check* square, to allow the development, refinement, and validation of methods and their subsequent 'blind testing'.
- CASI and LIDAR data were flown during summer 1999. The quality of these data is as good as Environment Agency systems at the time would allow. The CASI images for the Arable, Pastoral and Marginal sites contain residual geometric distortions caused by aircraft movement and topographic variation, whilst the images of the Upland sites contain slight data shifts inserted by the EA geometric correction software. A slight z-displacement is present in the LIDAR data, particularly of the northern sites.
- A data processing flowline has been developed and followed through to completion for the four 1 x 1 km trial squares. The methods and data-sets derived should be directly applicable to the four check squares.
- LIDAR data pre-processing for each square involves the creation of: (i) a Digital Surface Model from point sample information; (ii) a Digital Elevation Model in which all features with 'above ground height' are removed to give landscape elevation; (iii) surface height data for tress, buildings, hedges, etc.; (iv) slope and aspect data.
- CASI image pre-processing involves: normalisation of the spectral characteristics (across swath, between flightlines, and between sites); geometric correction by co-registration with LIDAR data; and spectral segmentation to derive the parcels for use in classification.
- Classification of the 12-band CASI images is a per-parcel procedure based on CLEVER-Mapping in Laser-Scan IGIS software. The classifier is trained by assigning labels to parcels of known land-cover type, and performed by a Maximum Likelihood algorithm, using mean parcel statistics to select the most likely class in statistical terms. The parcel statistics are extracted from a shrunken area (eliminating 3 marginal pixels all round) to avoid edge pixels which may have a mixed signature.
- A two phase knowledge-based correction (KBC) procedure has been developed to address identified classification errors. Phase-1 KBC operates per-parcel, re-assigning the land-cover class of specified parcels using a combination of context, LIDAR height data, CS 1990 codes, and class probabilities. Phase-2 KBC operates by examining the constituent pixels of a parcel, or by aggregating all contiguous parcels of the same land-cover class.
- Validation of the classification output has been developed as a measure of correspondence with the digitised CS2000 data, with land-cover classes aggregated into Broad Habitats. Although both the field survey and classified airborne imagery have land-cover data in more detailed classes than the Broad Habitats, this is the only level at which automated validation can be performed readily.
- Subtle differences exist between the CS2000 Broad Habitat data and the classified CASI-LIDAR imagery, necessitating a degree of pre-validation editing to make the two data-sets more comparable. For example, linear features of the field survey such as hedgerows have been given a width on the maps, rather than their being treated as merely bounding lines.
- Correspondence has been calculated for the classified CASI-LIDAR imagery (after both Phase-1 and Phase-2 knowledge-based correction) against the field survey data (both with and without edits). Per-pixel (with the field survey data rasterised to a 1m grid) and per-parcel validation procedures have been investigated.

- For the Arable, Pastoral and Marginal squares, per-pixel correspondence between the Phase-2 corrected CASI-LIDAR classification and the CS2000 Broad Habitat data falls in the range 80-87% (raising by 2 percentage points for the edited CS2000 data). Correspondence is 72% for the Upland square where vegetation types occur as fine spatial mosaics, but the field survey used highly generalised 1990 parcel boundaries (with updates) to record the patterns.
- A per-parcel correspondence of 94% was calculated for the Arable square by labelling the field survey parcels with the dominant CASI-LIDAR class. This product would not be dissimilar to the result of using the field survey boundaries as vector outlines for CASI classification.
- Progress is currently on schedule: the data processing flowline is complete, and the methods developed, appropriate training statistics and correction rules should be directly applicable to the four check squares.

2. INTRODUCTION

This project aims to evaluate the use of airborne scanner applications in the context of Countryside Survey 2000. Data acquired by the Compact Airborne Spectrographic Imager (CASI) and LIght Detection and Ranging instrument (LIDAR) are at a spatial scale which is intermediate between field survey and satellite data. Such airborne data may enable a greater understanding of the links between ground-based sample survey and the satellite-derived census, potentially allowing elements of the field survey to be replaced; they may also allow extension of the 1 km study sites to record extra details of their wider contexts.

The focus of this work is on identifying the extent and spatial patterns of land cover, linear landscape features and widespread Broad Habitats. Pairs of example survey squares are being studied in each of the Arable, Pastoral, Marginal and Upland Landscapes of GB - as defined in Countryside Survey 1990 (Barr *et al*. 1993). Each pair has been divided into a *trial* and a *check* square, to allow the development, refinement, and validation of methods and their subsequent 'blind testing'. Analysis is currently focussing on 1 km trial squares using integrated CASI-LIDAR data, acquired in summer 1999. Additionally, for one site in the Arable Landscape, a comparison is being made with CASI imagery from summer 1998; and the 1 km square is being studied in the context of the surrounding 3 x 3 km area.

This report is the fourth of a series of Interim Reports on the CASI-LIDAR Module of CS 2000 (see Fuller *et al.* 1998, Hill *et al*. 1998, 1999) and covers work done over a 12-month period from June 1999.

3. IMAGE ACQUISITION

3.1 Flight dates and specifications

All eight sites were re-flown by the Environment Agency (EA) with both the CASI and LIDAR instruments in the summer of 1999. Flying dates for the CASI were 25th June for the southern England sites (Arable and Pastoral), and 26th July for the northern England sites (Marginal and Upland). The LIDAR was flown on 8th and 17th June for the southern and northern England sites respectively.

The specification for the airborne data retrieval was the same as for 1998. Thus, at least a 3 x 3 km area was recorded for each Countryside Survey square, from which the central 1 km square can be extracted. The LIDAR data were first pulse only, capturing height information for the tops of vegetation canopies. The CASI data were recorded in twelve wavebands, which focused particularly on the red and near infrared spectral boundary, with a pixel size of 3 m.

3.2 Data quality

The 1999 CASI and LIDAR data were assessed at CEH. A slight edge-of-flightline *z*-displacement occurs in the LIDAR data, particularly of the northern sites. For the CASI data, the atmospheric quality is excellent and the geometric quality is as good as the EA systems will allow.

Pre-processing of the 1999 CASI data at the EA involved roll-correction only for flightlines covering the Arable, Pastoral, and Marginal squares. This was because of a problem in their Itres 'geocor' software, which generated erroneous data shifts if applied to the CASI imagery for geometric correction. For the Arable, Pastoral, and Marginal sites, the CASI data contain residual geometric distortions where aircraft roll has been either under- or over-compensated.

In addition, geometric distortions also result from underlying topography, which has not been accounted for in the pre-processing. The two Upland squares, however, were given the higher order geometric conversion, as conventional geometric correction of imagery can be a near impossible process in upland areas, where fewer prominent landmarks (e.g. field boundaries, crossroads) are found.

4. METHODS DEVELOPMENT

This project has several novel aspects that pose a unique set of challenges:

- The integration of multi-spectral CASI imagery with LIDAR elevation data to derive information on landscape features and structure.
- The use of remotely sensed data products gathered as part of an operational airborne remote sensing programme, rather than under research specifications.
- The use of automated image segmentation procedures, e.g. CLEVER-Mapping, for analysis of airborne images.

Methods development (in terms of both image acquisition and processing) has thus constituted a major part of this project for both the EA and CEH. A processing flowline has been developed at CEH that involves: elevation and height data generation from LIDAR point sample data; CASI normalisation, geometric correction and segmentation; classification; knowledge-based correction; and validation. Processing has focussed entirely on the four trial squares (Countryside Survey Squares [redacted], [redacted], [redacted], and [redacted]) which were selected by virtue of image data quality and land-cover diversity.

The methods, training statistics and correction rules developed for the trial squares should be directly applicable to the check squares (Countryside Survey Squares [redacted], [redacted], [redacted], and [redacted]). However, the initial phases of image pre-processing were highly interactive due to residual data quality issues. Although influencing the results of this project, the considerable developments in the EA image acquisition system should eliminate data quality problems from any repeat exercise.

4.1 LIDAR data pre-processing

4.1.1 Creating a Digital Surface Model

The LIDAR data were supplied by the EA as ascii files of *x*, *y*, *z* point information. As data retrieval is achieved in the LIDAR instrument by sampling across the aircraft flight path (i.e. across swath), the point sample information forms a zig-zag pattern with distribution varying, but typically falling *c.* 3-4 m apart. The LIDAR swath width is approximately 750 m, and the flightlines are flown to overlap. A 1 km square will contain around 165000-175000 sample points.

The first stage of the LIDAR pre-processing was to interpolate a continuous surface from the point sample information. This was achieved by the creation of an irregular triangular mesh (Triangulated Irregular Network) from the sample points. This was then transformed into a lattice with a rectangular array of mesh points with a chosen constant sampling interval in the *x*- and *y*- direction of 1m. Because the LIDAR data were first response only, this 1 m spatial resolution interpolated grid is a Digital Surface Model (DSM), as trees, buildings, etc, are present in the data with height expressed in metres above mean sea level (OS Datum) (Figure 1).

4.1.2 Creating a Digital Elevation Model

The second, and more complicated phase of LIDAR pre-processing, was the generation of a Digital Elevation Model (DEM) in which all features with ‘above-ground height’ had been removed to give genuine landscape elevation. To achieve this, all features with ‘above-ground height’ (e.g. trees, hedges, buildings) had to be removed from the DSM to allow re-interpolation of surface elevation across the gaps generated. Various methods of feature removal were investigated including the use of surface variance filters, and the mean filtering and statistical approach recommended by Jaafar *et al.* (1999). These approaches identified either variance in surface height or mean surface height, over a specified area using a spatial filter. The size of the spatial filter had to be determined statistically for each image, depending on the nature of landscape and surface feature variance. Once the appropriate filter size had been decided, a threshold was identified to distinguish between pixels representing the ‘true ground’ and those which represent unwanted features such as buildings. Using height variance filters, the threshold was applied directly to the resultant image whilst using mean filtering, the threshold was applied to the product which results from subtracting the filtered image from the DSM. In general, the variance filtering approach identified the edges of features such as hedges or buildings, whilst the mean filtering approach masked the centre of features. An additional stage was to ‘grow’ a mask outwards to capture a greater proportion of the unwanted surface features, or to use the two methods together to identify both the centre and edges of features. However, these approaches were too simplistic, with the complete removal of surface features such as hedges and buildings achieved typically at the expense of removing considerable areas of ‘true ground’. This influenced the potential accuracy of the surface interpolation across the masked areas, especially large blocks of woodland that contain areas of ground sampling in glades and rides.

Interpolation of heavily masked LIDAR data did, however, give a rough indication of the ground surface. This was used to put surface elevation information back into the original masked image, where the difference between the original DSM and interpolated surface were within a specified limit (e.g. + or - 0.5 metres). This enabled the creation of a mask which removed virtually all unwanted surface features (such as hedges and buildings) but considerably fewer true ground samples.

The actual method of interpolating across the masked off data gaps was selected from an operational standpoint. Possible procedures included: triangulation, splining, kriging and inverse distance weighted methods of interpolation. Of these, surface triangulation was the preferred choice since it represented a continuation of the method used to create the original DSM from the LIDAR point sample data. In addition, the other interpolation methods proved highly intensive on computer and analyst time to identify the optimum input parameters, which varied spatially depending on the nature of the landscape.

4.1.3 Creating a surface height model and other products

Once the Digital Surface and Elevation Models were complete, it was a simple task to create height data for surface features by subtracting the two data-sets (Figure 1). In addition, slope and aspect data were generated directly from the DEM.

Height data were generated only for Arable, Pastoral and Marginal squares, since no features with significant above ground height occur in the Upland sites. The accuracy of the surface height data was examined in Square [redacted] by comparing tree height estimates, derived using the 1999 LIDAR data and measurements taken in the field in 1998. Correspondence in height estimates was found to vary between 5 cm and 80 cm.

It should be noted that because of the slight *z*-displacement in the LIDAR data, a degree of manual editing was necessary to ‘clean’ the surface height imagery. However, both this process and the above method of height data generation would be unnecessary in any repeat exercise, since a replacement LIDAR system at the EA gives first and last pulse data, allowing surface height to be generated automatically.

4.2 CASI image pre-processing

4.2.1 Image normalisation

In airborne optical imagery, the spectral signal recorded for surface features will be ‘distorted’ by atmospheric effects of scattering and absorption. The degree of atmospheric noise in an image will vary with atmospheric conditions, and with both view and sun angle. Atmospheric attenuation needs to be accounted for to achieve comparability in spectral reflectance of features across image flightlines, or of areas sampled at different times or dates. Achieving this by detailed atmospheric modelling is far beyond the scope and time-frame of this project and, to-date, no generalised atmospheric correction model exists for airborne imagery. For operational purposes (in the absence of atmospheric correction models), it would be necessary to visit each site at the time of airborne data acquisition to record calibration reflectance data for target surface features. Given the spatial coverage of these trial data-sets, it is virtually impossible to find surface features within or between sites that should have identical surface reflectance spectra, since building materials, crop maturity, grassland nutrient status, and semi-natural land-cover mosaics, will all vary spatially.

It was possible, however, to perform some basic normalisation procedures. For example, a procedure for correcting view angle differences across the swath has been devised based on mean nadir values. In this procedure normalised pixel values were calculated as follows:

$$x'_{ij} = x_{ij} - (\bar{x}_j - \bar{x}_{nadir})$$

where x_{ij} was the original pixel value at row *i* and column *j*, $\bar{x}_j$ was the average of column *j* after smoothing using a moving average (100 pixels) and $\bar{x}_{nadir}$ was the average of the nadir column after smoothing (Figure 2). The radiance values of adjacent flight lines could be made comparable by normalising each to the mean scene values of the central flight line for each site, and the same approach could be used to normalise between sites of the same Landscape type. However, since an inherent assumption in this procedure was that the type and proportions of land cover were similar between flightlines and different sites, there was a limit to the degree to which normalisation could be performed. This restricted the assumed transferability of identified spectral characteristics. Thus, the 1999 CASI data of the Arable and Pastoral sites (which have a mixture of grassland and agriculture land-covers) were normalised to enable their combined training and classification. For the Marginal and Upland sites, however, the check squares will be normalised to the trial squares as two separate data-sets.

4.2.2 Geometric correction

Correction of the CASI imagery was necessary to remove geometric distortions remaining in the data following pre-processing by the EA (Figure 3). This was achieved by registering the required sections of each CASI flightline to the matching LIDAR data by identifying ground control points (GCPs) and performing ‘rubber sheeting’ to warp the image around those identified points. This was an extremely labour intensive process, requiring anything up-to 200 GCPs to correct a 1 km square. Furthermore, because only the specified control points were guaranteed to link the LIDAR and registered CASI imagery, it was virtually impossible to get a perfect correspondence. Registration was performed using a nearest neighbour algorithm, resampling the CASI imagery to match the 1 m spatial resolution of the LIDAR data. This method had the advantage of maintaining the original spectral value of pixels whilst achieving a more detailed spatial matching by sub-dividing each 3 m CASI pixel. It must be remembered, however, that the minimum mappable unit will not be reduced in size by this apparent increase in image spatial resolution.

For many of the sites (Squares [redacted], [redacted], [redacted], [redacted] and [redacted]), the central 1 km square did not fall entirely within one flightline but was split across two adjacent runs. In these circumstances, the registered flightline sections had to be mosaicked to generate a single data-set (Figure 3).

4.2.3 Image segmentation

The image segmentation procedure was based on the same software package being used in LCM2000 (Fuller *et al.* 1999). Written originally in the Microsoft Windows environment by the Cambridge University Geography Department, Laser-Scan has now implemented a fully operational version of the segmentation software, in a Unix environment for use with IGIS.

Important methodological issues for image segmentation include:

- Band selection for edge-detection and segmentation,
- Setting thresholds to identify edges and generate segments,
- Post-segmentation boundary rejection and generalisation.

It was only possible to use three bands for the edge-detection / segmentation process and so the optimum choice of wavebands was investigated using the four trail 1 km squares of 1999 CASI data. Principle Components Analysis of the 12-band CASI images, demonstrated these data to be two-dimensional (with at least 96% of variance contained in PCs 1 and 2). The two dimensions relate to the visible and near infrared (NIR) part of the spectrum. Correlation analysis supported these findings, with strong positive correlations within, but not between, the visible and NIR wavebands. In spite of the strong 2-dimensionality of the data-set, it was decided that out of the 12 available wavebands, the three bands which made the strongest contribution to PCs 1-3 and which were the least correlated were Bands 4, 6, and 10. These occupy a point of maximum red absorption by vegetation (670 nm), a point along the so-called ‘red-edge’ (708 nm) between the red absorption trough and NIR reflectance peak, and a point in the NIR vegetation reflectance maximum (780 nm). The segmentation algorithm was tested using PCs 1-3 and CASI Bands 4, 6 and 10, in the four Landscape types. This demonstrated the use of individual wavebands to give a better result, with more ‘meaningful’ parcels created.

The segmentation procedure builds polygons around ‘seedpoints’ that have been selected as within a segment or a land parcel; an edge detector is used to ensure that the appropriate seedpoints are selected away from parcel-edges. There is potential in the software to dictate the degree of region merging by setting segmentation thresholds for each of the spectral bands and by establishing the number of standard deviations expected to contain the majority of the population of a segment. If the first threshold (entered separately for each band) was set low (i.e. 1 SD) then a higher number of segments was generated initially. If the second threshold was then set high (i.e. 6 SDs) a much greater level of region merging took place. This gave a much better end-product than growing bigger parcels initially, as more detail was retained without generating an overly segmented image (Figure 4).

Post-segmentation generalisation simply involved dissolving parcels of 9 or less pixels (i.e. one pixel of data in the raw CASI image) into the surrounding parcels. Sliver parcels greater than 9 pixels in size occurring at boundaries were, however, retained since linear features were very much a part of the CASI data.

It is important to note that this was a low-level segmentation process (Haralick & Shapiro, 1985) in that the parcels created were not necessarily meaningful entities (such as fields) but merely parts of them. The parcels were identified according to spectral variation which may have related, for example, to crop development, wind damage or unplanted field margins.

Once acceptable segmentations were achieved, vector versions were created in a GIS database. This was a simple procedure of raster-to-vector conversion where the boundaries between segments with different values in the raster images were represented by vector lines. These formed the basis of the vector data-base used in the classification procedure.

5. CLASSIFICATION

The classification approach was a per-parcel procedure based on CLEVER-Mapping (Smith & Fuller 1998), using the vector boundaries derived from the segmentation procedure and the full 12-band CASI image. For the purposes of the software package, a 16-bit to 12-bit conversion of the imagery was required. This reduced the dynamic range of the spectral values recorded, but maintained the relative differences between landscape features.

The classification was trained by assigning a class value to selected parcels of known land-cover types (Table 1). For the trial squares, this made use of detailed data from the Field Assessment Booklets and from personal visits to the sites during 1998 and 1999. Training was at the level of individual land-cover types, allowing post-classification aggregation into Broad Habitats (Table 2). For example, in the 1998 imagery of Arable Square 180, eight agricultural types were identified in the training data (bare, barley, wheat, turnips, peas, kale, harvested and set-aside). The basic aim of the training procedure was to identify as much spectral variance within the image as possible, and to achieve this for each land-cover type present (i.e to achieve a full and accurate sub-division of the spectral feature space). Because of the nature of the segmentation process, the parcels available for training varied in size, but were reasonably consistent in spectral variance. The important consideration in creating a training data-set in the CLEVER-Mapping system, was therefore not achieving an equal distribution of parcel size, but achieving an even distribution of training parcels throughout the spectral feature space.

Having identified a series of training parcels, it was then necessary to review the training data to decide on the spectral sub-classes to be used for classification. A refinement built into IGIS operation (as part of the LCM2000) allowed ‘image chips’, representing the remotely sensed data for each training area, to be displayed side-by-side on the screen, almost like a colour-chart. This enabled the training parcels to be compared and labelled to give a series of different spectral sub-classes where necessary (Kershaw & Fuller, 1993). The training areas were reviewed in what was considered to be the two most useful 3-band combinations (Bands 4, 3, 2 and Bands 10, 6, 4) to ensure that the spectral sub-classes were not mixed. When deciding on the aggregation of training parcels, the general rule applied was that the narrower range of spectral variance allowed in each spectral sub-class, the less likely would be confusion in classification at the aggregate level. So, for example, 11 sub-variants of bog were identified in the classification of Upland Square [redacted].

Training was carried out separately for the following 1 km CASI data-sets: Arable Square [redacted] (1998 data); Arable Square [redacted] / Pastoral Square [redacted] (1999 data); Marginal Square [redacted] (1999 data); and Upland Square [redacted] (1999 data). Only in the case of the Arable and Pastoral squares in 1999 CASI data, was land-cover distribution considered similar enough to enable between-site spectral normalisation. The total array of spectral sub-classes identified across the 1 km squares is shown in Table 2. These can be readily amalgamated into Broad Habitats, with the one exception of BH 3 (Boundary and linear features) which was trained for classification into its constituent parts of hedges and built surfaces.

The classification procedure used the Maximum Likelihood algorithm applied to the parcel, using mean statistics to select the most likely class in statistical terms. The parcel statistics were extracted from a shrunken area (by a margin of 3 pixels) to avoid edge pixels with a mixed signature.

A substantial refinement of the CLEVER-Mapping system in IGIS has been to record the probabilities for the top five sub-class options, usually covering >95% of the probability distribution (Fuller *et al.* 2000). This information proved to be useful in later knowledge-based corrections.

6. KNOWLEDGE-BASED CORRECTION

A degree of mis-classification of parcels was expected due to spectral similarities between certain land-cover types. Likely inter-class confusion could be estimated prior to classification from the review of training data. For example, the three grassland types in the four trial sites (improved, neutral and acid) showed spectral overlap with each other, and with sunlit aspects of deciduous woodlands / hedges, and with certain crop types (e.g oilseed rape, peas, maize, barley) depending on crop maturity. The shaded aspects of deciduous woodlands / hedges showed spectral overlap with mature arable wheat, marsh / swamp, water, and shadow; whilst built surfaces showed spectral overlap with the arable classes of bare, harvested, and set-aside. Since shadows can be cast over any land-cover type present within a square, this class had a wider spectral range and showed overlap with more land-cover classes than the other spectral sub-classes.

Knowledge-based correction (KBC) procedures were required to address these mis-classification errors, and have been developed using a combination of context, LIDAR height data, CS 1990 codes, and class probabilities. Because the correction procedures operated per-parcel, more subtle internal context rules could be used (e.g. assigning parcels to adjoining or nearby classes).

6.1 Phase-1 KBC procedure

The simplest KBC rules devised were contextual, based on a parcel being surrounded by an unlikely land-cover type (Table 3). To give some examples, an arable parcel surrounded by built surfaces was relabelled as built, whilst a shade parcel surrounded by deciduous woodland was coded as deciduous. It must be remembered that, although the parcels reflect

Land-cover Class	Arable		Pastoral	Marginal	Upland
	1998 data	1999 data	1999 data	1999 data	1999 data
Arable bare	X	X	X	X	
Arable barley	X	X			
Arable harvested	X	X		X	
Arable kale	X				
Arable linseed		X			
Arable maize		X	X		
Arable peas	X				
Arable rape		X			
Arable set-aside	X				
Arable turnips	X				
Arable wheat	X	X	X		
Grassland – improved	X	X	X	X	
Grassland – neutral	X	X	X	X	
Grassland – acid				X	X
Coniferous woodland	X	X			
Deciduous woodland	X	X	X	X	
Deciduous hedge	X	X	X	X	
Dwarf shrub heath					X
Fen, marsh, swamp				X	
Bog					X
Built surface	X	X	X	X	
Water	X	X		X	
Shadow	X	X	X	X	

Table 1 Land-cover types identified in each trial 1 km square. (Note each of these land-cover types may be composed of several spectral sub-classes)

Landscape type	No. of parcels	No. of spectral sub-classes	No. of land-cover types	No. of Broad Habitats
Arable (1998 data)	103	41	16	7
Arable & Pastoral	200	59	15	7
Marginal	139	40	11	8
Upland	63	17	3	3

Table 2 Breakdown of the training data used for the classification of the trial squares. (Note that the Arable and Pastoral Squares in 1999 CASI data were trained and classified together.)

Land-cover class	Surrounded by:	Convert to:	In Square(s):
Arable bare	Arable barley	Arable barley	Ar (98)
Arable bare	Arable set-aside	Arable set-aside	Ar (98)
Arable bare	Built surface	Built surface	Ar /Pa (99)
Arable barley	Arable harvested	Arable harvested	Ar (98)
Arable barley	Arable set-aside	Arable set-aside	Ar (98)
Arable barley	Grassland - improved	Grassland – improved	Ar /Pa (99)
Arable harvested	Arable wheat	Arable wheat	Ar (98)
Arable harvested	Built surface	Built surface	Ar /Pa (99)
Arable maize	Built surface	Built surface	Ar /Pa (99)
Arable peas	Arable bare	Arable bare	Ar (98)
Arable peas	Arable harvested	Arable harvested	Ar (98)
Arable peas	Arable wheat	Arable wheat	Ar (98)
Arable rape	Grassland - improved	Grassland – improved	Ar /Pa (99)
Arable turnips	Arable barley	Arable barley	Ar (98)
Arable wheat	Arable harvested	Arable harvested	Ar (98)
Arable wheat	Arable rape	Arable rape	Ar /Pa (99)
Grassland – improved	Arable peas	Arable peas	Ar (98)
Grassland – improved	Arable set-aside	Arable set-aside	Ar (98)
Grassland – improved	Arable rape	Arable rape	Ar /Pa (99)
Grassland – improved	Arable maize	Arable maize	Ar /Pa (99)
Grassland – improved	Grassland – neutral	Grassland – neutral	Ar /Pa (99)
Grassland – improved	Grassland - acid	Grassland – acid	Ma (99)
Grassland – neutral	Grassland – improved	Grassland – improved	Ar /Pa (99)
Grassland - acid	Grassland - improved	Grassland – improved	Ma (99)
Fen, marsh, swamp	Grassland – acid	Grassland – acid	Ma (99)
Fen, marsh, swamp	Grassland - improved	Grassland – improved	Ma (99)
Deciduous hedge	Arable harvested	Arable harvested	Ar (98)
Deciduous hedge	Arable peas	Arable peas	Ar (98)
Deciduous hedge	Arable set-aside	Arable set-aside	Ar (98)
Built surface	Arable set-aside	Arable set-aside	Ar (98)
Built surface	Grassland – improved	Grassland – improved	Ar (98)
Built surface	Grassland	Arable bare	Ar /Pa (99), Ma (99)
Built surface	Arable	Arable bare	Ar /Pa (99), Ma (99)
Shadow	Deciduous woodland	Deciduous woodland	Ar (98), Ar /Pa (99)
Shadow	Arable wheat	Arable wheat	Ar /Pa (99)

Table 3. Contextual knowledge-based correction rules as applied to the Arable, Pastoral and Marginal trial squares.
Ar (98) = Arable Square [redacted] in 1998 data, Ar /Pa (99) = Arable Square [redacted] and Pastoral Square [redacted] (1999 data), Ma (99) = Marginal Square [redacted] (1999 data).

genuine spectral variance from ground features, they do not necessarily represent whole objects. Thus, fields were composed of many parcels, and so the KBC rules operated at the within-field level. Changes to parcel class assignment through the KBC process were applied at the level of land-cover types within the Broad Habitats. Thus in an arable setting, class re-assignment would be to an individual crop type.

The LIDAR height data was invaluable at addressing mis-classification between the deciduous woodland / hedge classes and certain grassland and arable classes. It was possible to identify a height threshold, which all parcels classified as hedge or deciduous must exceed, and all other parcels (except for built surface) must be less than. Conversion to deciduous woodland / hedge classes was simple, but conversion from deciduous woodland / hedge to neighbouring land-cover classes was according to a series of priority rules.

Classification errors between shadow and both arable and built spectral sub-classes was also addressed using the LIDAR height data, since a shadow could only be cast if neighbouring a feature with height. A mask, based on a threshold distance to a surface feature with height, was applied to convert erroneous shadow parcels to the surrounding land-cover classes, again according to a series of class priority rules.

The CS 1990 codes (and obviously the CS 2000 codes in any repeat exercise) represent an important data source that could be used in the KBC process. However, using these data for full knowledge-based correction would remove any ability to identify change by the classification of airborne imagery. Exceptions to this are the more stable classes such as roads, railways and built up areas, which are highly unlikely to be converted into agricultural use, grassland, forestry or semi-natural vegetation. Thus, a mask of CS 1990 reporting classes 51-52 (Railway and Road) and classes 53-55 (Built on land) was applied to identify and re-assign parcels mis-classified as arable (bare, maize, harvested), shadow, or water; and to distinguish between Broad Habitats 3 (Boundary and linear features) and 17 (Built up areas and gardens). From an operational standpoint, this correction can only be applied to Countryside Survey squares for which previous field survey data exist.

6.2 Phase-2 KBC procedure

The Phase-1 KBC procedures were applied to the spectrally determined parcels. Additional KBC procedures could be performed on a per-pixel basis or after aggregating all contiguous parcels of the same land-cover class. At the aggregate level, a repeat of the above contextual KBC rules enabled additional cleaning to take place. For example, a patch mis-classified as grass in the middle of an arable field, would not have been converted in the Phase-1 KBC procedure if composed of more than one parcel. In addition, at the aggregate level, it was possible to add a suburban label to parcels of grass or woodland land-cover within an urban setting, thereby placing them into BH 17 (Built up areas and gardens).

At the pixel level, a more spatially detailed knowledge-based conversion was performed to correct deciduous woodland / hedge classification and to remove shadow. Per-pixel KBC was particularly useful for correcting between deciduous woodland / hedge and other classes, since the height data was averaged across spectrally defined parcels in the Phase-1 KBC. The greater spatial detail of per-pixel KBC also enabled the attempted conversion of shade parcels into the likely underlying land-cover types, according to a series of decision rules based on context and class priorities.

Although not used in the above KBC procedures, there is no reason why the elevation, slope and aspect information derived from the LIDAR data could not be used to identify parcels assigned to classes outside their natural context. This may prove particularly useful in the Upland Landscape type, for which no KBC rules have yet been developed.

Later work will investigate the use of LIDAR height data to distinguish between spectrally inseparable deciduous woodlands and hedges, and to identify hedgerow trees from scattered trees. This will, however, require a more overt object-oriented form of analysis.

7. VALIDATION

Validation of the classification output (Figure 5) has been performed as a measure of correspondence with the digitised CS2000 Broad Habitat data. Although both data-sets (Field Survey and classified airborne imagery) have land-cover data in more detailed classes than the Broad Habitats, this is the only level at which automated validation can be performed readily.

Three potential methods of calculating correspondence have been investigated:

(i) per-pixel correspondence between the two data-sets at 1 m spatial resolution,

(ii) labelling the CS2000 parcels with the dominant CASI-LIDAR class and comparing the result with the CS2000 field survey data, and

(iii) labelling the segmented CASI parcels with CS2000 data and comparing the result with the CASI-LIDAR classification.

Investigations, as part of the LCM 2000 validation work, have shown that rasterising the CS2000 vector data to a 1 m grid alters the spatial area estimates of Broad Habitat classes by an average of just 0.1%. Comparison between the two 1 m spatial resolution grids thus gives a direct correspondence per-pixel between the field survey and airborne image classifications. Attaching the classification of one data-set into the vector boundaries of the other for validation purposes (as was tested for Arable Square [redacted]), gave a higher measure of correspondence, and so was considered inappropriate unless this is to done routinely in the production phase.

Subtle differences exist between the CS 2000 Broad Habitat data and the classified CASI-LIDAR data:

- a one year time difference exists between the 1998 field survey and 1999 airborne imagery;
- a mis-alignment occurs between the two data-sets as the field survey linework, digitised from OS mapsheets, does not meet the 15 cm *x*-, *y*-accuracy of the LIDAR data (which, here, is considered the baseline for inter-comparisons);
- a distinction occurs between land-use mapped in the field survey and land-cover mapped in the airborne imagery (e.g. BH 3 (Boundary and linear features) is an amalgamation of hedges, roads and railways);
- the field survey does not identify hedges as features having an area, but as boundary features in a separate layer of the GIS database.

To make the two data-sets more comparable, it was necessary to edit the field survey polygon data (Figure 6). Firstly, 1999 field reconnaissance data were used to update the distribution of arable fields in the field survey data. Secondly, all boundaries identified in the field survey GIS database as hedges have been ‘burnt into’ the 1 m rasterised grid data. These have been assigned a width in accordance with a minimum mappable unit for the CASI data of 3 m (the original image spatial resolution). The inserted hedges were assigned to BH 1 (Broadleaved, mixed and yew woodland). Individual trees, identified in the field survey, were also inserted into the 1 m rasterized grid data. Thirdly, the CS2000 field survey data have been registered to the classified airborne data by a third order polynomial transformation. This process affected the area estimates of the land-cover types present in each 1 km square.

Because of the two-stage KBC procedure and field survey editing, there were two key airborne image classification products and two field survey products that may be of interest for the validation exercise in each square:

(i) image classification with Phase-1 KBC and with the extra Phase-2 KBC;
(ii) field survey data *with* and *without* the hedges/trees inserted and geometric registration.

The validation exercise was thus carried out as a per-pixel correspondence between the four possible data-set combinations, for all four trial squares classified in 1999 CASI data (Squares [redacted], [redacted], [redacted], [redacted]), and the test site classified in 1998 CASI data (Square [redacted]). For comparative purposes, per-polygon correspondence values were calculated for the 1998 CASI classification (Figure 7).

8. RESULTS OF VALIDATION EXERCISE

Correspondence between the field survey data and the classified airborne images at the Broad Habitat level was extremely high for the trial Arable ([redacted]), Pastoral ([redacted]), and Marginal ([redacted]) squares (Table 4). The correspondence level increased with the secondary KBC procedure, and with the registration and editing of the field survey data. These two editing processes applied to the field survey and airborne image classification raised the correspondence for the sites as follows: Arable Square (1998 data) 82-86 %, Arable Square (1999 data) 82-89%, Pastoral Square (1999 data) 82-87%, and Marginal Square (1999 data) 78-81%.

Correspondence statistics in the Upland Square [redacted] were somewhat complicated by the the complex mosaic of semi-natural vegetation types. The field surveyors showed the impossibility of mapping such mosaics as individual parcels, but characterised the complexity of the mosaic components by recording plant quadrats. One-third of the field survey data for this site was labelled as mosaic at the Broad Habitat level. Excluding these areas from the correspondence measures gave a value of 67.8%. Treating a parcel as correct if assigned to one of the Broad Habitat classes constituting a mosaic gave a correspondence for the 1 km square of 72%. This was notably lower than the correspondence for the Arable, Pastoral and Marginal sites, and largely reflected a finer spatial scale of vegetation mosaics than was depicted in the field survey parcel boundaries (Figure 8).

The per-polygon correspondence values calculated for Arable Square [redacted] (1998 CASI data) were consistently higher than the per-pixel values (Table 5). The difference between per-pixel and per-polygon correspondence was only slight (1-2 percentage points) using the CASI segments labelled with the dominant field survey Broad Habitat class. However, using the field survey polygons coded with the dominant CASI-LIDAR class, the correspondence value increased by up to 10 percentage points. So for example, the correspondence between the CASI-LIDAR classification with a 2-stage KBC and the edited and registered field survey data was 85.8 %, calculated per-pixel, 88.3% calculated using the CASI segment boundaries labelled with the field survey data, and 93.6% using the field survey boundaries labelled with the CASI-LIDAR classification. This later product would not be dissimilar to the result of using the field survey boundaries as the vector outlines for CASI classification (Figure 8).

Landscape type	Phase-1 KBC *vs* unedited CS data	Phase-1 KBC *vs* edited CS data	Phase-2 KBC *vs* unedited CS data	Phase-2 KBC *vs* edited CS data
Arable (1998 data)	82.0	83.1	83.2	85.8
Arable (1999 data)	82.3	84.1	86.6	89.0
Pastoral (1999 data)	82.2	82.7	83.5	86.7
Marginal (1999 data)	78.0	79.3	79.7	80.9
Upland (1999 data)	71.0*	-	-	-

Table 4. Correspondence (in %) between CASI-LIDAR image classification and CS 2000 field survey data. (NB Upland square has no KBC.)

Comparison made	Per-pixel	Per-parcel (CASI-LIDAR)	Per-parcel (field survey)
Phase-1 KBC *vs* unedited CS data	82.0	82.6	92.0
Phase-1 KBC *vs* edited CS data	83.1	84.3	93.8
Phase-2 KBC *vs* unedited CS data	83.2	84.6	92.6
Phase-2 KBC *vs* edited CS data	85.8	88.3	93.6

Table 5. Comparison between per-pixel and per-parcel correspondence for Arable Square 180 in 1998 CASI data. Correspondence values are given (in %) per-pixel, and per-polygon using the CASI segment boundaries labelled with the field survey data, and using the field survey boundaries labelled with the CASI-LIDAR classification.

The residual differences in correspondence between the field survey data and the classified airborne data, after cleaning, editing and co-registration can be accounted for by:

(i) residual mis-registration between the two data-sets;
(ii) distinctions between land-cover mapped from the airborne images and land-use mapped in the field survey (e.g. ley grass is an agricultural land use but an improved grass land cover);
(iii) the field survey has been shown to have a repeatability level of 88% in identifying primary land cover codes, which are used objectively to generate BH data;
(iv) remaining mis-classification not corrected in the KBC procedure.

9. PROGRESS MEASURED AGAINST SCHEDULE

Work is currently on schedule and should be completed by March 2001 (see up-dated GANNT). Emphasis has focussed on completing the work for the trial squares, which has been finished ahead of schedule, whilst the pre-processing of the CASI data for the check squares has yet to begin. LIDAR pre-processing for the check squares is however complete. The pre-processing of the 3 x 3 km test area of Square [redacted] (1998 data) is underway.

Methods development is now essentially complete, and it should be a matter of 'handle-turning' to apply the methods to the check squares. Once complete, comparison work for Square [redacted] can be instigated, examining the differences between 1998 and 1999 classifications and between the central 1 km square and the surrounding countryside of a wider 3 x 3 km area. Also, investigations into more object-oriented analyses can begin in the remaining time, which should allow a more sophisticated generation of landscape statistics.

10. REFERENCES

CASI-LIDAR Reports:

Fuller, R.M., Hill, R.A. and Veitch, N. (1998). Countryside Survey 2000 – Part I. Module 8: Airborne Scanner Applications: Classification of airborne casi and LIDAR data of selected CS2000 sample squares. *First Interim Report, CSCL/Int.1. Unpublished report to the DETR.*

Hill, R.A., Fuller, R.M., and Veitch, N. (1998). Countryside Survey 2000 – Part I. Module 8: Airborne Scanner Applications: Classification of airborne casi and LIDAR data of selected CS2000 sample squares. *Second Interim Report, CSCL/Int.2. Unpublished report to the DETR.*

Hill, R.A., Fuller, R.M., & Veitch, N. (1999). Countryside Survey 2000 - Part I, Module 8: Airborne Scanner Applications. Classification of airborne casi and LIDAR data of selected CS2000 sample squares. Third Interim Report. Unpublished ITE report to the DETR.

Hill, R.A., Fuller, R.M., & Veitch N., (2000). *CS2000 Project Specification Module 8: Airborne Scanner Applications - classification of airborne casi and LIDAR data of selected CS2000 sample squares. Quarterly Progress Report*. Unpublished ITE report to the DETR.

Other references:

Barr C.J., Bunce R.G.H., Clarke R.T., Fuller R.M., Furse M.T., Gillespie M.K., Groom G.B., Hallam C.J., Hornung M., & Howard D.C. (1993) *Countryside Survey 1990: Main Report.* London: Department of the Environment.

Fuller R.M., Smith, G.M., Sanderson, J.M., Hill, R.A., Thomson, A.G., (1999). *Countryside Survey 2000 - Part III. Module 7. Land Cover Map 2000. Second Interim Report. incorporating the Fourth Quarterly Progress Report CSLCM/Interim2.* Unpublished ITE report to the LCM2000 Consortium.

Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, R.A., Hall, M., Thomson A.G. & Brown N.J. (2000). *Countryside Survey 2000 Module 7. Land Cover Map 2000. Fourth Interim Report incorporating the Eighth Quarterly Progress Report. CSLCM/Int4/Prog8.* Unpublished CEH report to the DETR.

Haralick, R.M., & Shapiro L.G. (1985) Image segmentation techniques. *Computer Vision, Graphics and Image Processing,* **29**, 100-132.

Jaafar J., Priestnall G., Mather P.M., & Vieira C.A (1999) Construction of DEM from LiDAR DSM using morphological filtering, conventional statistical approaches and Artificial

Neural Networks. *Proceedings of the 25th Annual Conference of the Remote Sensing Society,* 8-10 September 1999, Cardiff.

Kershaw, C.J. & Fuller R.M. (1992). Statistical problems in the discrimination of land cover from satellite images: a case study in lowland Britain. *International Journal of Remote Sensing*, 16, 3085-3104.

Smith G.M, & Fuller R.M (1998) *CLEVER-Mapping of land cover in Jersey. Final Report.* ITE report to the States of Jersey.