Brian Baily*

The extraction of digital vector data from historic land use maps of Great Britain using image processing techniques

Keywords: Land utilization; Dudley Stamp; image processing; geo-reference; supervised classification; digitization.

Summary
The mapping of the land utilization of Great Britain has taken place periodically for the whole country since the 1930s. This and other subsequent projects produced a series of paper maps, which potentially hold valuable information on land use change across Britain. The Department of Geography at the University of Portsmouth, has been involved with a series of collaborative projects, which have collated and scanned the maps from the First Land Utilisation survey of Great Britain directed by Professor L. Dudley Stamp of the London School of Economics. These maps have been geo-referenced and made available to the public via the Vision of Britain web site (www.visionofbritain.org.uk). Interest was shown by a number of organisations concerning the possibility of extracting vector data from the maps relating to the various land use categories. As part of the research, two pilot projects were carried out at Portsmouth to examine the various techniques available and to ascertain time and costs estimates for the data extraction from the full set of maps. As well as manual digitization, image-processing techniques were used to classify scanned images of the selected maps. GIS software tools were then used to refine and clean the data. This paper reviews the work carried out thus far and highlights some of the problems that have arisen.

Introduction

Historical land use maps of Great Britain exist for a number of epochs and are potentially invaluable in the analysis of the changing face of the British landscape. Many of these maps exist purely in a hard copy format and as a result a great deal of potentially valuable information has been lost over time. The collation, scanning and geo-referencing of these data sources are crucial if further material is not to be lost. Whilst the digital output of the maps in itself is invaluable, vectorised data of the various land use types would allow a wider, more detailed use of the data. As a result, following funding from the Environment Agency of England and Wales, projects were set up to collate, scan and geo-reference the maps from the First land Utilization Survey of Great Britain (Stamp 1931, 1948). These have subsequently been made available to the general public via the Vision of Britain web site (www.visionofbritian.org.uk). A subsequent project was asked to investigate the possibility of using a series of methods for extracting land use data from the various maps into a vector format. These techniques included manual digitization and image processing classification techniques using Leica Erdas 8.7. These procedures are generally used on satellite imagery to semi-automatically extract data often relating to land use. This paper ex-

* Dr., Senior Research Associate, University of Portsmouth, Department of Geography, Buckingham Building, Lion Terrace, Portsmouth, England, PO1 3HE [brian.baily@port.ac.uk]

amines the potential applicability of this approach alongside in an initial attempt to extract vector data from the maps.

Background to the Land Utilisation Survey

The Land Utilisation Survey of Great Britain (LUSGB) was accomplished during the 1930s and was directed by Professor L. Dudley Stamp (Figure 1) of the London School of Economics. The aim of the survey was to create a series of one inch to the mile (1:63,000) maps of the whole of Britain showing the various land use types.

Figure 1. Sir Dudley Stamp (1898-1966)

The field survey work was organised by administrative county, the first contact usually being with the Director of Education. Arrangements for the survey were in place for most English counties by the summer of 1931, and for most Welsh and Scottish counties a year later (Stamp, 1948). Teams of volunteers including students and schoolchildren carried out the fieldwork identifying the various land use groups. Each supervised group was given a sheet covering six square miles to annotate with the various categories (Figure 2). Six different categories were identified including (a) meadow and permanent grass, (b) arable land including rotation grass, (c) heathland and moorland or rough hill pasture, (d) forests and woodland, (e) gardens and (f) land agriculturally unproductive (Stamp 1931). By the autumn of 1934, 90% of the field survey maps had been returned. However, there were two major problems, firstly it was difficult to find enough volunteers for all areas, especially those in more remote or hazardous locations. As a result, Stamp organized teams of students and staff to go to particular areas. The last area to be surveyed was the Isle of Arran in 1941. The second major problem involved funding with Stamp having to go to a variety of sources to get sufficient income to complete the project.

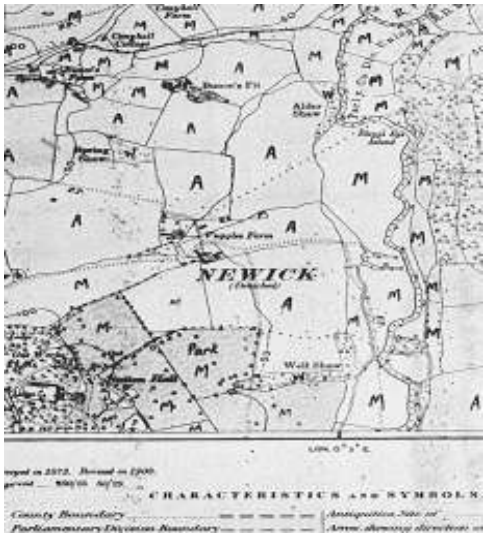

Figure 2. A six-inch to the mile Ordnance Survey map annotated with some of the land use classes.

The first of the resulting composite one-inch to the mile maps was published in January 1933. However, because of staffing and financial problems, the printing and publishing process was an extremely slow and complex programme. The Ordnance Survey (OS) printed the first nineteen sheets for Stamp, but early in 1935 the OS complained that printing the land use maps was straining their resources. Between 1935 and 1949, eight separate printers produced the remaining sheets. In all, the Land Utilization Survey published 135 maps of England and Wales, an additional 34 of Scotland, and 92 County Reports (Stamp 1948, Southall *et al* 2003)

Surviving materials and study sites

The principal output from the Stamp Survey was a set of one hundred and sixty nine one-inch composite maps (all land use classes), over-printing land use information onto reproductions of the Ordnance Survey's *Popular Edition* maps (Figure 3).

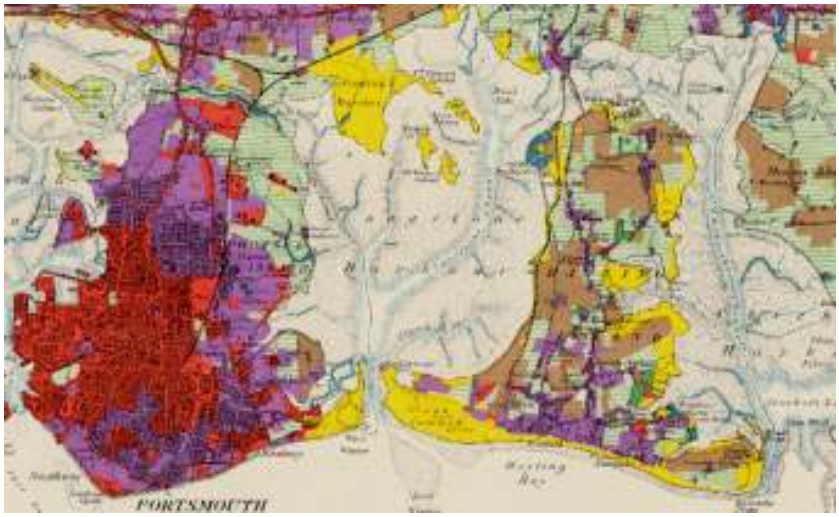

Figure 3. An example of one of the one-inch to the mile land utilization maps showing the Portsmouth and Hayling Island area of England.

The Stamp survey also produced summary (generalised) sheets at ten miles to the inch (1:625,000) (Figure 4). Four distinct summary maps were published, each covering Great Britain in two sheets:

- Land Utilisation
- Land Classification
- Types of Farming
- Grasslands (of England and Wales) and Vegetation (of Scotland)

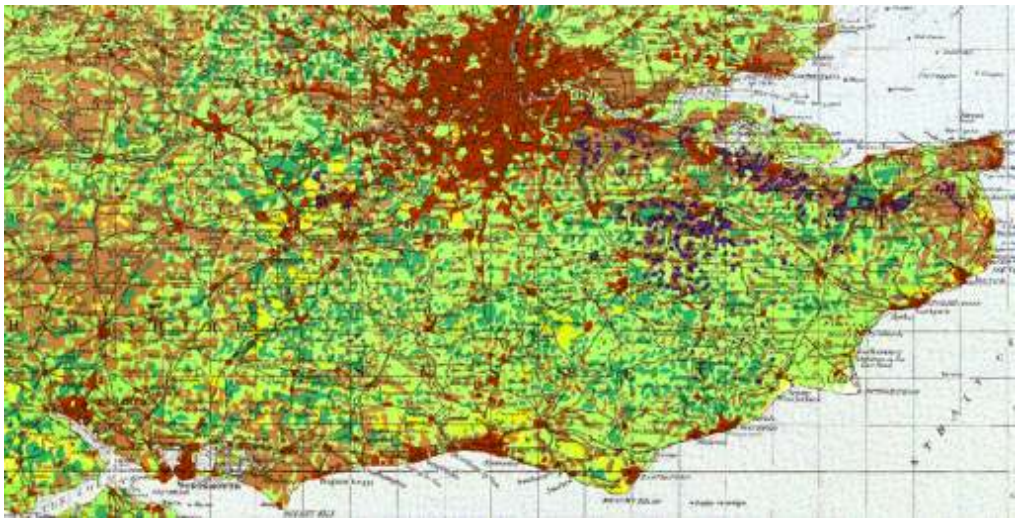

Figure 4. An example of one of the UK summary sheets showing south-east England.

One further output available from the Stamp survey was essentially a by-product of the printing process. Separate colour sheets were available for the individual colours used in the printing process. The separate 'pull' sheets have the advantage of not being contaminated by other data (e.g. text and contour lines) and contain only one land use class per sheet (Figure 5). However, it should be noted that some of these are in a very poor condition and that the majority have been lost or destroyed (Figure 6).

Figure 5. An example of a separate colour pull sheet from the Stamp survey.

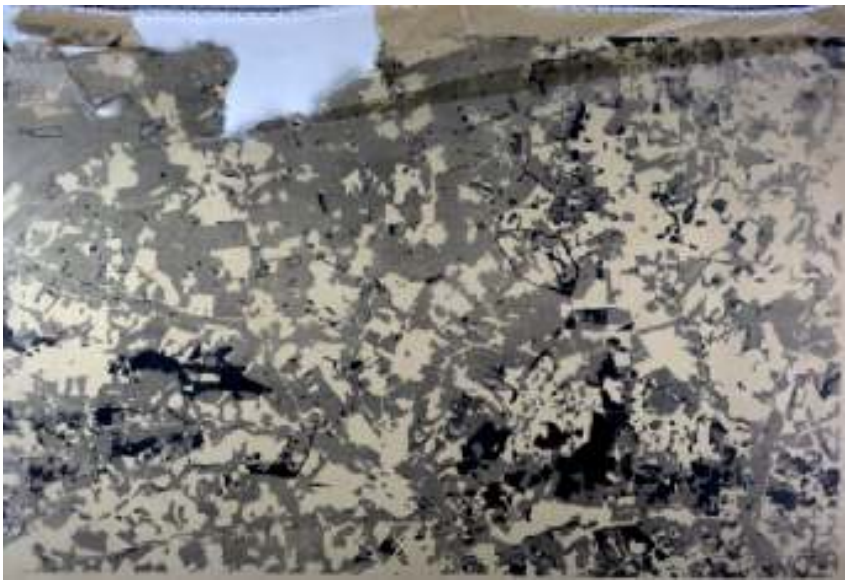

Figure 6. A damaged separate layer sheet.

Extraction of vector data from the maps: potential methodologies

One aspect of the research undertaken at Portsmouth, concerned the extraction of data from the maps covering the various land use classes. This work builds on the approach outlined by Southall *et al* (2003). The aim of this process is to design a semi-automated approach to extract the relevant data using Leica Erdas 8.7 and ESRI ArcGIS tools. Before any digital data extraction could occur the maps had to be scanned and geo-referenced. These digital files not only allowed manual digitising of the various land use polygons, but also offered the possibility of using a semi-automatic data extraction technique similar to the methodologies used in satellite image processing (see for example, Lillesand and Kiefer, 2004, Chapter 7 in particular).

Scanning

The first stage in the procedure involved the scanning of the maps. A common resolution for scanning large maps is between 300 – 400 dots per inch at a colour depth of 8 bits per channel (24 bit colour). Obviously, this resolution can be increased but is in part, determined by the storage capacity available to the project, the physical size of the material, and the printing and dissemination techniques applied to the subsequent imagery. There is also a threshold to resolution where any further increase will not yield a noticeable improvement in quality.

One hundred and forty six individual composite sheets of England and Wales had already been scanned and geo-referenced during a previous stage of the research (Southall *et al.* 2003). However, as part of the ongoing project the colour separations of a number of sheets existed and were scanned by King's College London. The areas scanned for analysis of the separate layers were the Salisbury and Bulford sheet, the Birmingham sheet and the Dartmoor, Tavistock and Launceston sheet. All of the colour separations were scanned as one image. Also selected for analysis, were the UK Summary sheets separate layers, which because of their size were scanned in two sections.

Geo-referencing

The digital files of the scanned maps do not contain any information relating the area represented on the map to its location on the ground. This means that it is not possible to view, query or analyse the data with other geographic data, or indeed with any other of the scanned maps. In order to create this functionality, it is necessary to align, or geo-reference, the image to a map coordinate system, in this case the OS National Grid. Maps containing a printed grid are simple to geo-reference as it is possible to click on an intersection of the grid and type in the coordinates for that point. However, the LUSGB maps have no grid and therefore prominent landmarks must be used whose coordinates could be surveyed or that could be identified on another, already geo-referenced, source. Using the latter approach it is possible to geo-reference a map by clicking on landmarks within the LUSGB image, such as churches or road junctions and then click on the same features within an already geo-referenced map such as those that can be obtained from the Ordnance Survey.

One significant disadvantage of using a product from the Ordnance Survey to geo-reference the LUSGB images, is that the resulting combination of information would probably be regarded by the OS as a 'derived work' in which they held a copyright, and could control dissemination. Fortunately, the Great Britain Historical GIS project at Portsmouth had already created a complete set of geo-referenced one-inch to the mile maps that contain grid lines but were published more than fifty years ago, and are therefore free from OS copyright. These New Popular Edition maps from the 1940s have therefore been used as the source of coordinate information for geo-referencing. The maps were then geo-referenced using the Leica Imagine Geometric Correction tool (Figure 7).

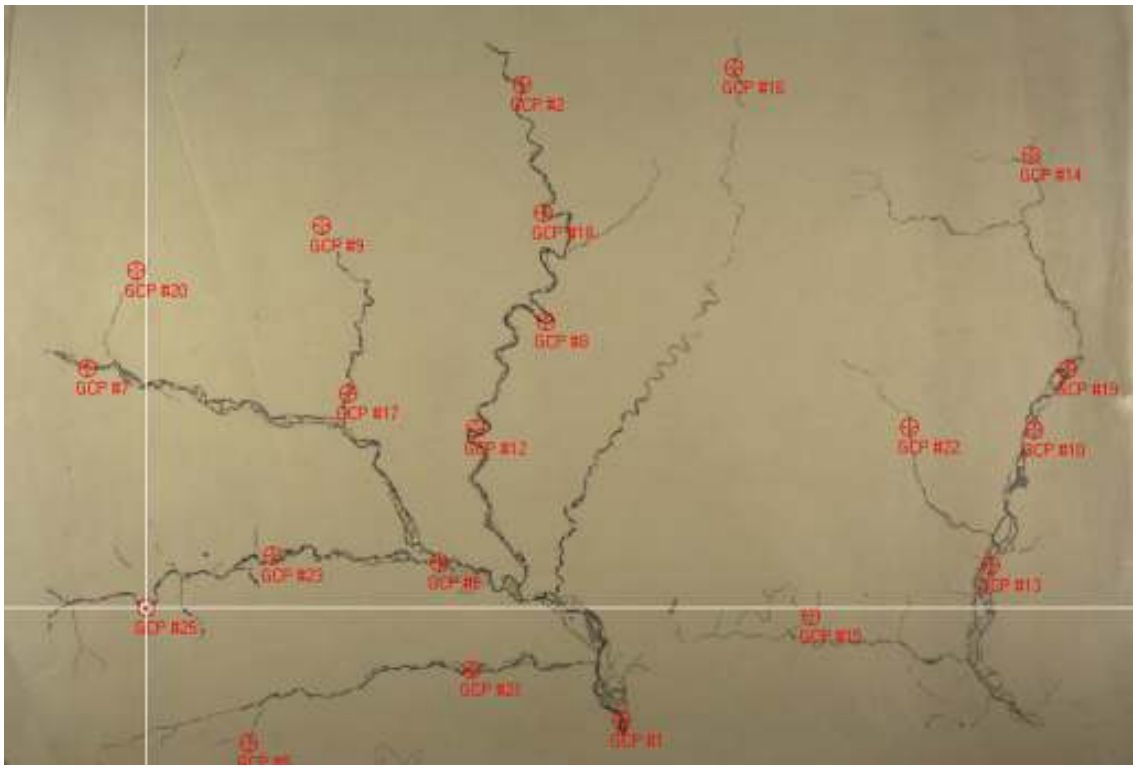

Figure 7. Control points being measured on the water (blue) separate layer sheets.

Extraction of digital vector data from the Stamp land use data

Two main methods were compared for the extraction of digital data from the maps. Although several others were tested and discarded.

1. Manual digitizing

The first method used was the traditional technique of manual digitization of the map. This method has advantages over techniques as it yields highly accurate results, as well as requiring no further editing, other than edge-matching sheets together. However, given the complexity of the maps this could prove economically unacceptable. A small section of a selection of maps was digitised using this method, which gave an indication of the amount of time required for the digitization of larger areas. With this method, the image was uploaded to a screen and a head up digitising technique used to capture the data. This method does have the advantage of allowing sub-groups within the classes to be separated and digitised (e.g. the class forest has 3 sub-groups).
2. Supervised classification
The second method adopted was to use a supervised classification technique to extract data from the various land use classes. This approach is taken from standard methodologies used for classifying remotely sensed imagery such as that gathered using the SPOT and Landsat satellites. Although the LUSGB maps contain a far greater amount of 'clutter' (such as text and contours, see for example Figure 8) than satellite imagery, it was considered that there would still be merit in experimenting with remote sensing classification techniques. This method has the advantage of potentially being quicker than manual digitization and therefore cheaper economically.

2a. Sharpening the image

One technique, which helps to improve the accuracy of the later stages of this method, is to sharpen the image. For this project, the Leica Imagine Crisp tool was used. The Crisp filter sharpens the overall scene luminance without distorting the interband variance content of the image. This is a useful enhancement if the image is blurred or fuzzy at any point. The crisp tool helps to sharpen up the edges of the various LUSGB classes.

Figure 8. Showing examples of 'clutter' on the composite maps.
The left image shows the text and the building outlines whilst the right image shows the contour lines.

2b. Classification

The main method which was used to classify the map is supervised classification, which utilises supervised training and is closely controlled by the analyst (Leica Geosystems, 2003). In this process, groups of pixels are selected which represent patterns or land cover features from the various land use classes. This approach requires knowledge of the data and of the classes desired before classification. By identifying patterns, it is then possible to instruct the computer system to identify pixels with similar characteristics. If the classification is accurate, the resulting classes represent the categories within the data that were originally identified. The result of training is a set of signatures that defines a training sample. Each signature corresponds to a class, and is used with a decision rule to assign the pixels in the image file to a class. Some of the classes within the LUSGB maps have a series of sub-categories defined by different symbols on the maps. However, these are not easily distinguishable from each other and classification techniques were unable to separate these sub-groups. Several training classes were identified for the one-inch sheets representing the various land use classes, which could be separated (Table 1).
Several training classes were identified for the one-inch sheets representing the various land use classes (Table 1).

Initial map class	Colour / detail
Black topological detail and text	Black - To be removed
Forest and woodland	Green with black symbols - combined from 3 subclasses
Meadowland and permanent grass	Light green (hatched line symbol)
Arable land	Brown
Water	Blue, sometimes with white lines
Heath and moorland	Yellow
Land agriculturally unproductive (e.g. Urban core)	Red
Gardens etc (e.g. suburban)	Purple

Table 1. The various land use classes extracted from the whole LUSGB sheets.

For the separate colour layers (pull sheets) only one land class is represented for each colour group identified in Table 1. Therefore, the training for this involved the identifcation of the class represented in the separate map and a class representing the background. The colour separations were easier to classify, as these are free from the clutter, which is on the composite LUSGB sheets (Figure 8).
Figure 9 shows a typical area identified for training purposes from the composite maps. Rather than identifying a very clean area of colour, ‘cluttered’ areas were selected in the hope that the software would learn to ignore the clutter during the classification process; thereby reducing the time spent cleaning up the data afterwards. Several training areas were selected for each map, which were then merged into one class. This technique was also copied for the separate colour map sheets to compensate for colour variation. Once the training areas have been selected the software then produces a new image which is grouped into the various classes which have been

identified in the training. This new image is still geo-referenced and has attribute data relating to its identified land use.

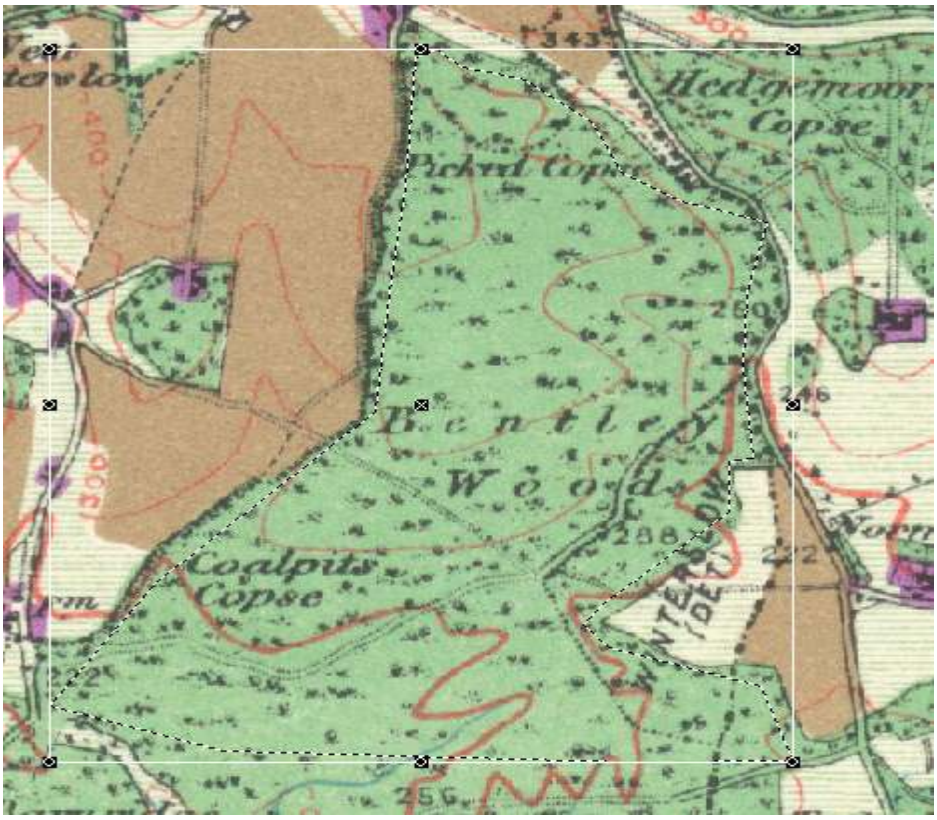

Figure 9. Showing an area of forest including clutter, extracted for training purposes.

Figure 10 shows a classified image against the original colour separation. Unlike the whole sheets only minimal further editing will be required.

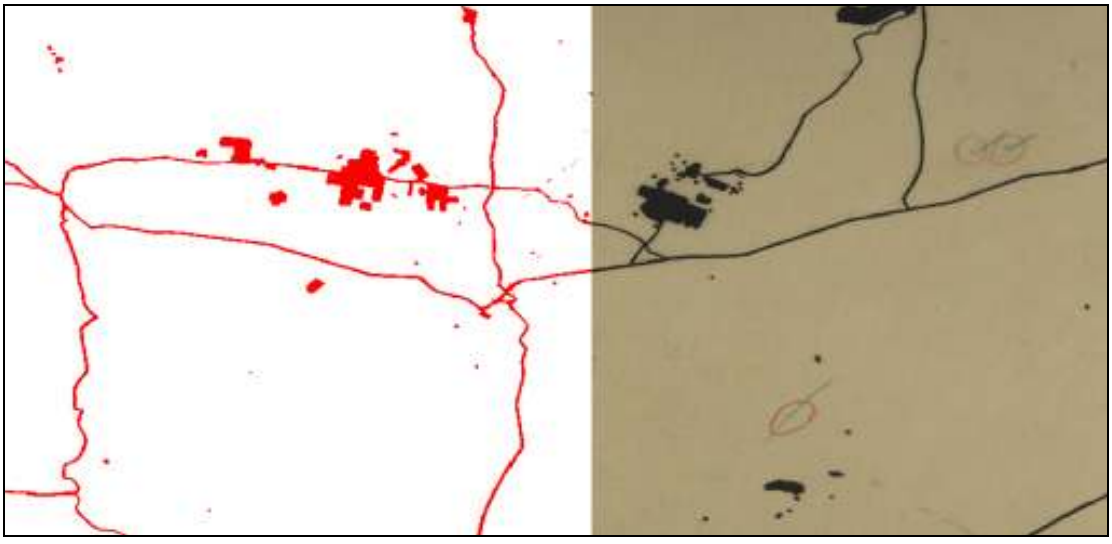

Figure 10. Showing a classified image (left) and the original red colour separation (pull sheet) for the Salisbury and Bulford map.

2c. Further processing

Although the supervised classification techniques is successful in ignoring much of the black detail and other unwanted data in the composite one-inch sheets, it was still necessary to undertake further cleaning of the resultant image in order to create clean land use polygons. It was noted, that the land use types represented by more intense colours classified with greater success than others. In particular, striped coloured areas (e.g. meadowland) produced mixed results, which require further refinement. Figure 11 shows an area of the classified image still exhibiting some clutter from the original map.

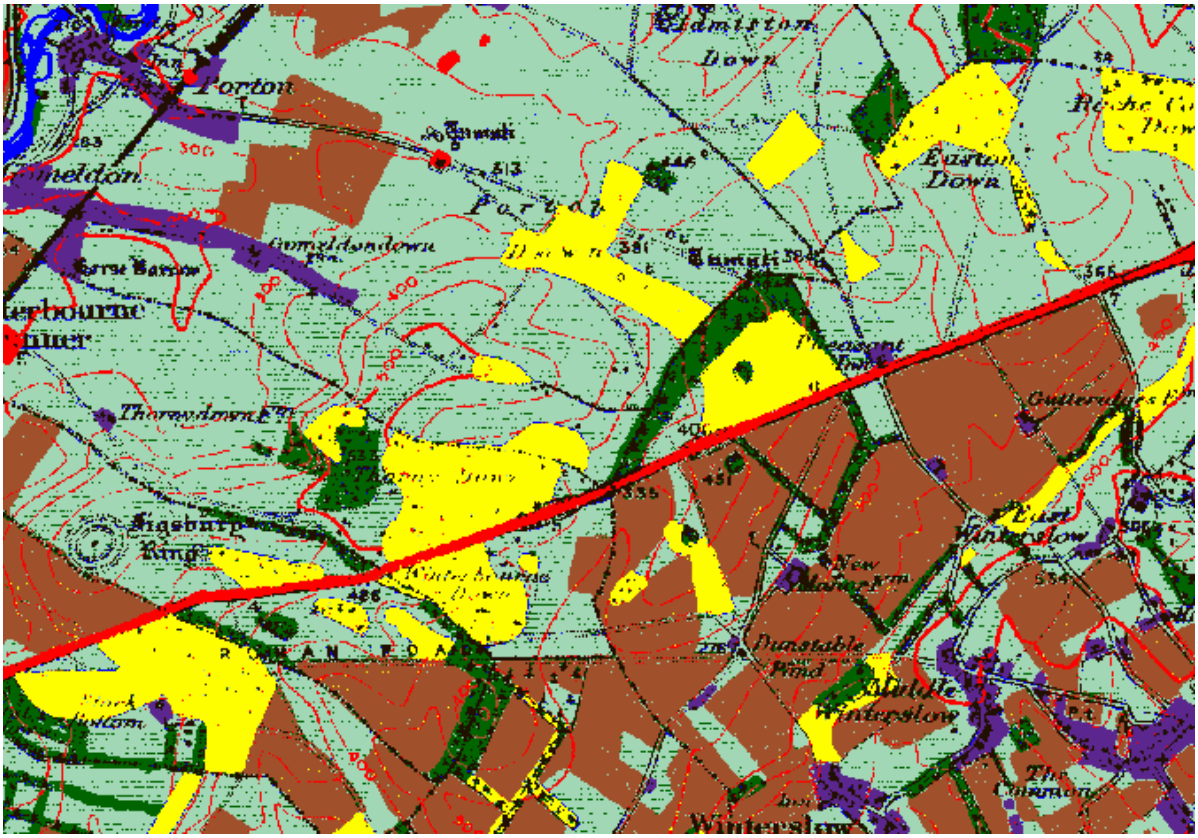

Figure 11. An initial classified image of part of the Salisbury and Bulford sheet; some of the text and contour lines are still clearly visible.

One potential method for reducing some of the noise in the image is to run a neighbourhood function in Erdas 8.7. This process allows you to perform one of several analyses on class values on an image using a process similar to convolution filtering. Neighbourhood functions are specialized filtering functions that are designed for use on thematic layers. Each pixel is analysed with the pixels in its neighbourhood. The number and location of the pixels in the neighbourhood are determined by the size and shape of the filter, which you define. A 7X7 neighbourhood function was used which appeared to reduce the black detail and greatly improved the quality of the meadowland classified areas (Figure 12).

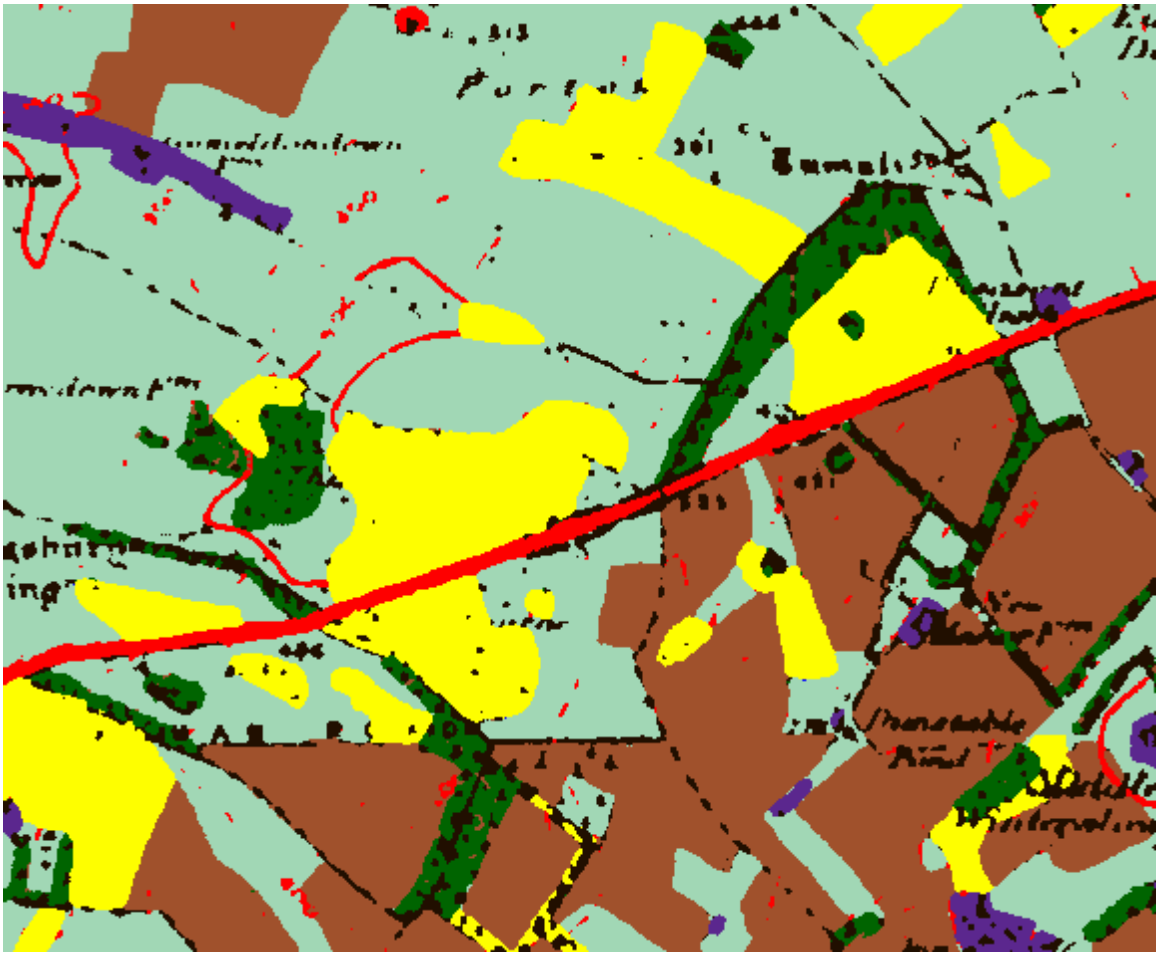

Figure 12. A classified image after the neighbourhood filter has been used in Leica Erdas 8.7.

Further refinement of the image was still required to remove the remaining detail. It can be seen in the first classification example shown above that although the basic class structure appears to be correct, there are still some problems with unwanted data. As expected there is still some black detail which needs to be removed. There are also some other problems including the remaining contour lines etc. Many of these features will be removed during either the dissolving process or in the later removal of small parcels. Some will need final interactive removal.

2.d Removal of anomalies in the classified data

After the classification and initial filtering tools interactive editing is required to remove further unwanted detail. The large majority of the unwanted data is left over from the black text. The removal of the unwanted black detail was removed using ESRI ArcGIS tools as follows:

- Use Arc Grid 'focalmajority' function which removes most linear features, such as road casings, and narrow text. Some black detail is left especially where the black text or detail was thick o the original maps.
- Other data was removed using the ArcGrid 'nibble' function. This allows all other wanted classes to eat into the black' nodata' areas, completely removing them.

At this stage of the process further editing and cleaning has to be done manually. The most time consuming part of the 'clean up' procedure occurs when interactive editing and analysis is required. The data at this stage has had most of the unwanted data removed. However, there are still are two main types of anomaly to be examined. The first group are small, unwanted parcels which have kept in the data that have not been removed by the previous processes. One other group of data which requires editing are those which have been incorrectly coded. One example of this is road and contour line parcels as they are red on the maps. Some automatic removal of anomalies is possible, however larger road sections would need to be manually edited. At this point in the process the maps were converted to a vector format for the final stages of the editing.
The vector versions of the maps were examined in detail in relation to the original paper map information. The smallest individual parcels that were depicted on the original maps were measured and were just over a quarter of a hectare in size. This allowed a threshold of 0.28 hectares to be chosen, below which features have been dissolved into the background, using the 'eliminate' function of the GIS. This function dissolves parcels away, and replaces them with the attribute possessed by the adjoining parcel with the longest shared boundary line. This appeared to work well for the majority of the features however, in some instances, especially in relation to road detail, it is necessary to interactively, select an alternative adjoining value, in order to retain the cartographic integrity of the source map detail.
A final interactive check is essential to ensure correct conversion from the source paper map to the final digital vector map has been achieved. Checks on the quality of the final data can be made by comparing polygons on the original map with the polygons on the classified version. For a number of test areas the agreement was generally around 90-95% for all land use categories. It was generally concluded that the longer the time spent editing the unwanted detail the better the results obtained in the final classified image (Figure 13).

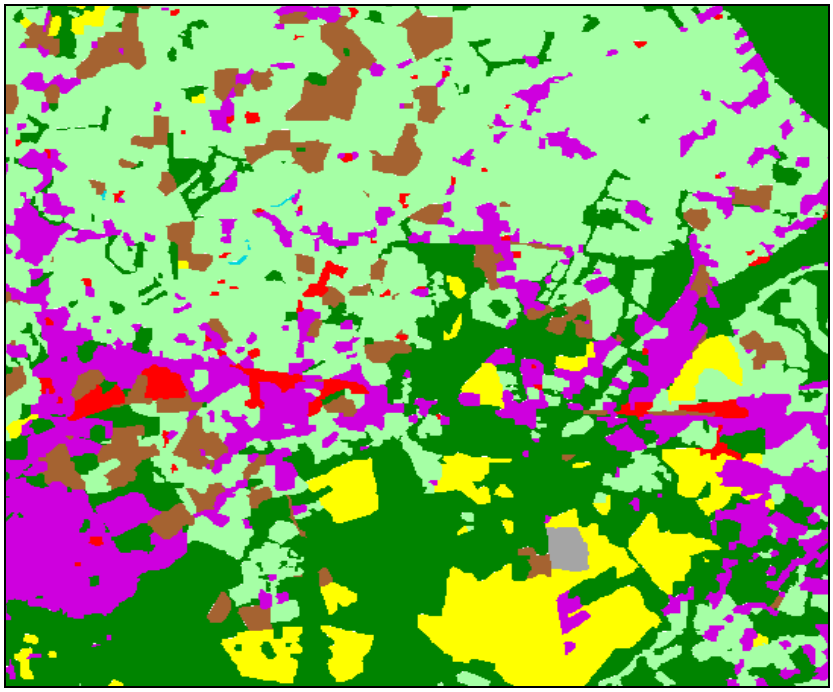

Figure 13. A final digital classified map section produced by supervised classification.

Assessment of the value of the colour separation maps

One element of the research was a comparison of the results for the classification of the composite sheets (all land use layers) obtained via supervised classification, with similar results obtained from the original colour separations (see Figure 10) used in preparing the same sheet. The working assumption was that the colour separations would provide, a less problematic data set and therefore a more accurate representation of the same areas covered on the original printed maps. This is because they are free of the text and other miscellaneous data seen on the composite maps. The vector data extracted from the composite maps was compared with the vector data extracted from the separate layer sheets using the Erdas 8.7 swipe and blend tools and by correlation of the areas covered by the selected land use classes. Although the results for the two largest land uses (meadowland and arable) are satisfactory (matching areas of 97.34% and 95.54% respectively), the worst result is for "agriculturally unproductive land" (69.01% correlation). The land use classes with the lowest spatial extent on the maps tended to be less reliable than the larger groups of data. Although the results look visually acceptable with the matching areas correlating to 93%, unacceptable differences were still apparent between the relevant data sets. Closer analysis revealed that the colour separations could not be viewed as a secure benchmark. The largest problem is that the separate coloured layers were clearly designed to overlap slightly, to prevent any gaps in the final published map. Figure 14 shows how several layers can overlap, only one of which would be picked up in the classification of the whole sheet. Figure 15 demonstrates how even a few polygons have a spatial difference. The cumulative effect of these will be large over the whole sheet and will vary for different categories depending on which ones were used as the top layer in the printing process. Similarly, in any geo-referencing of an image to another image there will always be an error between the two sheets. This will be shown in differences around the edges of the individual polygons. It is suggested, that if all the LUSGB maps were to be classified, that a random sample area on a few maps is hand digitised and compared to the automated classification technique to give an indication of similarity across the whole survey.

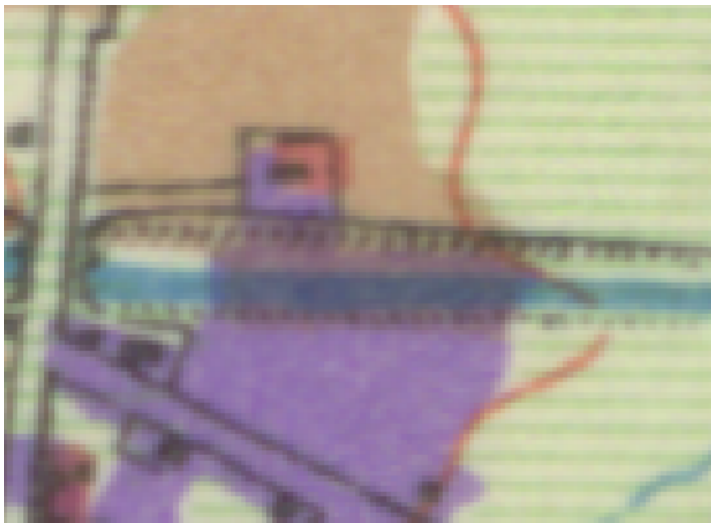

Figure 14. Blue, brown and purple overlap on a whole sheet there is also overlap from the striped green layer.

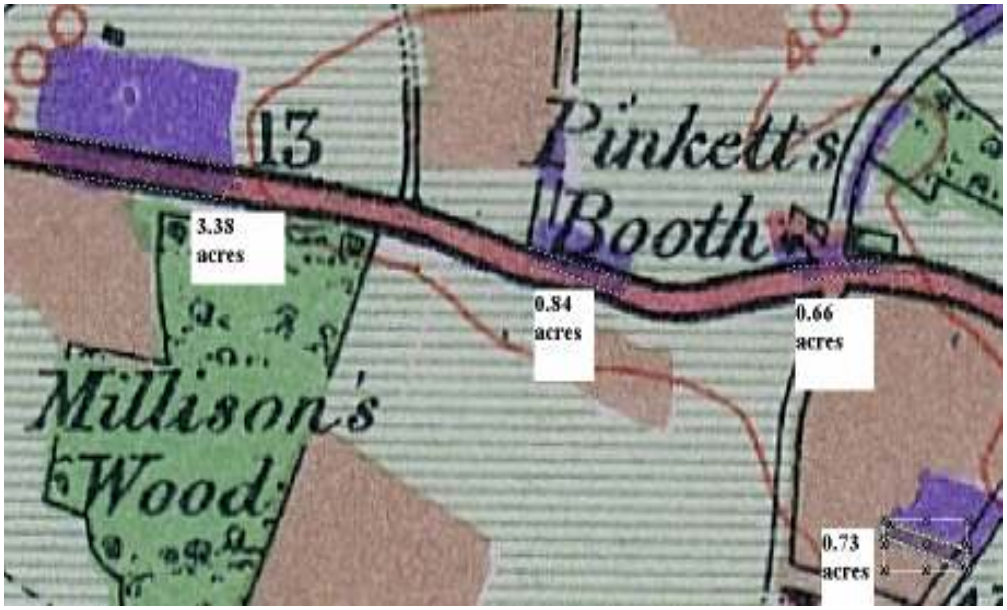

Figure 15. Area calculations for a small section of purple overlap.

Other factors also cause poor matches between the printed maps and the colour separations. The particularly poor result for the red colour, i.e. 'agriculturally unproductive', is partly because the published maps also include red lines coming through from the Ordnance Survey base maps, showing roads. All of these factors combined suggest that the most useful and accurate digital vector data would be obtained from the composite not the separate layer maps.

Time estimates for the vectorization of the complete LUSGB map collection

Having established that with further refinements there are at least two successful methods to vectorise the LUSGB maps, estimates were calculated for the time periods required to perform this process on the whole set. Manual digitising although giving potentially the best results would take longer. It was calculated that on average this would take around 93 hours per whole one-inch sheet, with only a small amount of editing required afterwards. This would amount to 1874.5 working days for the full set of maps (Table 2). The semi-automated method whilst not being to separate the sub-categories would take around 11.5 hours per map (Table 3). Most of this time is concerned with the GIS editing which is the key to getting good results. The total amount of time required for all the maps would be a total of around 247 working days.

	Number of maps	Time per map	Total Hours
English maps	118	93	10,974
Welsh maps	17	93	1,581
Scottish maps	37	93	3,441

Table 2. Hours required for maps of the various regions using manual digitising.

	Number of maps	Time per map	Total Hours
English maps	118	11.5	1357
Welsh maps	17	11.5	195.5
Scottish maps	37	11.5	425.5

Table 3. Hours required for maps of the various regions using semi-automatic classification digitising.

Discussion

The analysis produced so far is the result of the two pilot projects designed to investigate methods for the vectorization of the Stamp land use maps. It is clear, as would be expected that the manual digitising techniques offer the most accurate and thorough solution. However, it is also clear that the economic costs of this approach would probably be too high. The image processing technique using supervised classification on the composite sheets does appear to offer good results and a reasonably quick alternative. This process would be considerably cheaper in economic terms and could then be applied to other land use data sets e.g. Alice Coleman's Second Land Utilisation Surveys of the 1960s (Coleman and Shaw 1980 and Coleman *et al*. 1992). Further tests should be carried out on a few sheets to further determine the replicability of the method and it's potential accuracy for the various land use classes. Whilst it is true that some complicated sheets will require more editing this would be balanced by those sheets with less detail e.g. Large areas of moorland. The semi-automatic supervised classification does offer a relatively cheap and quick alternative to manual digitization and further refinements in the GIS editing stages may enhance the results obtained.

The analysis of the individual colour separations showed that they are much easier to classify than the whole maps. Unfortunately, the printing process required polygons to overlap to prevent gaps in the final maps. As a result, the output cannot be seen as reliable and therefore must be rejected as a potential data source for any future studies.

The results form the pilot projects suggest that semi-automatic classification using image processing software alongside GIS editing tools does allow the vectorization of the Stamp land use maps. Further analysis and investigation is required to refine these techniques still further to improve the data output and streamline the editing stages. Any project, which classifies the maps, should also digitise a selected number of polygons from the various categories to ensure that the classified data is honouring the data in the original map.

References and Sources

Coleman, A., and Shaw, J.E. (1980). *Land Utilisation Survey: Field Mapping Manual* .Second Land Utilisation Survey, London.

Coleman A., England E., Latymer Y. and Shaw J.E. (1992). Scapes and Fringes 1:400,000 "*Environmental Territories of England and Wales". Second Land Utilisation Survey,* London; 2 maps, each 139×93 cm., pLUSGB 98 pp. booklet.

Leica Geosystems (2003). *Erdas Field Guide 7th Edition*. GIS and Mapping LLC, Atlanta, Georgia.

Lillesand T.M. and Kiefer R.W. (2004). Remote Sensing and Image Interpretation. Wiley.

Southall H, Brown N, Burton N, Williamson A. (2003). Digitising the Inter-War Land Use Survey of Great Britain: A Pilot Project. Environment Agency, Environmental Policy, Risk and Forecasting Report. Number 44, July 2003.

Stamp L.D. (1931). The Land Utilization Survey of Britain. The Geographical Journal, 78, pp. 40-47.

Stamp L.D. (1948). *The Land of Britain: its use and misuse* (Longman, London), esp. pp. 3-20 (later editions in 1950 and 1962).