Title: Integrating land cover data with different ontologies: identifying change from inconsistency

Alexis Comber^{1*}, Peter Fisher¹ and Richard Wadsworth²

¹Department of Geography, University of Leicester,
Leicester, LE1 7RH, UK.

² Centre for Ecology and Hydrology,
Monks Wood, Huntingdon, Cambridgeshire, PE28 2LS. UK

email: ajc36@le.ac.uk*, pffl@le.ac.uk, rawad@ceh.ac.uk

*Corresponding author

Abstract

Spatially coincident land cover information frequently varies due to technological and political variations. This is especially problematic for time-series analyses. We present an approach using expert expressions of how the semantics of different datasets relate to integrating temporal time series land cover information where the classification classes have fundamentally changed. We use LCMGB and LCM2000 as example data sets because of the extensive object based meta-data in the LCM2000. Inconsistencies between the two datasets can arise from random, gross and systematic error and from an actual change in land cover. Locales of possible land cover change are inferred by comparing characterisations derived from the semantic relations and meta-data. Field visits showed errors of omission to be 21% and errors of commission to be 28%, despite the accuracy limitations of the land cover information when compared to the field survey component of the Countryside Survey 2000.

1. Introduction

The process of natural resource inventory takes place against a background of change which may result in inconsistency. Human understanding within the scientific field which is the subject of the inventory may well vary as may the policy initiative under which the inventory occurs. The methods by which the inventory could be conducted may be revised and the nature of the resource at a location may change (Comber et al., 2003a). These different changes necessitate the re-inventory of an area after some number of years. In geological mapping and soil survey changes in understanding and methods are the most important reasons for re-inventory. In the study of land use and land cover the main incentive for remapping is that the use or cover itself may have changed, but at the same time there may be alterations to the methods of analysis. Changes in methodology make it difficult to separate changes in the phenomenon being measured (such as land cover) from changes that are the result of the revised methodology; that is to say ontological inconsistency may be a problem for change detection. This discord causes problems for research that seeks to develop time-series of land cover or land use to

monitor environmental change and for initiatives that aim to react to environmental change.

The issue of dataset inconsistency is endemic to resource inventory and should be unsurprising to many involved in the activity. Firstly, different surveys at the same instance in time may record nominally similar features (such as land cover), but may do so in completely different ways due to their particular institutional or, indeed, personal perspectives. Secondly, different surveys at different points in time would not be expected to record objects of interest (such as patches of land cover) in the same way because of scientific developments and new policy objectives (Comber *et al.*, 2002; 2003a).

The effect of changing methodologies is that much of the value of previous land resource inventories is lost with each successive survey; each inventory becomes the new baseline against which future changes are theoretically to be measured, but in reality never are and cannot be. Ideally each successive methodological evolution would be accommodated in a multi-layered derived dataset, presenting the previous and the new approaches alongside each other. Often constraints of time and money prevent this.

In this paper we describe an approach for integrating spatial data that combines expert descriptions of how the semantics of different land cover datasets relate with object level spectral meta-data. This approach is applied to two satellite derived land cover surveys of the UK in order to identify inconsistency between the datasets, a subset of which are locales of actual land cover change.

2. Land Cover and Changing Ontology

2.1 Land Cover Mapping in the UK

One specific example of data dissimilarity caused by revision of methodology is land cover mapping in Great Britain where land cover inventories have been conducted in 1990 and 2000: the Land Cover Map of Great Britain (LCM1990) and the Land Cover Map 2000 (LCM2000) (Fuller *et al.*, 1994; Fuller *et al.*, 2002). Both surveys are based on similar data (composite winter and summer satellite data), but contain significant structural and thematic differences. LCM1990 records 25 Target land cover classes selected by an expert panel and primarily defined using reflectance values with limited ancillary data; classification was carried out using a supervised maximum likelihood algorithm applied to each pixel. The project demonstrated the viability of using satellite imagery to map land cover in the UK. In contrast LCM2000 records 26 Broad Habitats which were designed to fit into new policy obligations as a result of the Rio earth summit, and subsequent national and European Union legislation. The “expert panel” of 1990 was expanded to a steering group that included a range of potential users. The steering group then had to marry what was technically possible with that which was politically suitable. LCM2000 identified land cover using a *per-parcel* supervised maximum likelihood classification algorithm. That is, a segmentation procedure was used to divide the landscape into parcels that were more or less spectrally homogenous. Next the “core” of each parcel was extracted, where the core was assumed to be

homogeneous. The whole parcel was then classified on the mean reflectance of this core. Knowledge-based corrections were applied to the data for certain types of habitat (soil acidity masks to identify Acid, Neutral and Calcareous grassland Broad Habitats; peat drift masks to identify areas of Bog) and to re-classify parcels that were out of context (Fuller *et al.*, 2002). Extensive meta-data is attached to each LCM2000 parcel, including a description of the whole parcel spectral heterogeneity, the spectral subclass percentage, and some processing history information, as described by Smith and Fuller (2002).

In Britain at the time of writing, the only national land cover datasets are the LCM2000 and its predecessor the LCM1990. Yet because of the problems of semantic and methodological difference, the 2000 dataset is accompanied by a “health warning” against comparing it to its predecessor (Fuller *et al.*, 2002). In the research reported here we are interested in identifying those locations where the land covers are inconsistent between the two dates of classification as a first step in identifying change.

2.2 Ontologies

LCM2000 is very different from LCM1990, not just in terms of its more readily quantifiable aspects such as minimum mapping unit and its data structures, but also with regard to the whole way that land cover in the UK is conceptualised. The objects identified by each survey, represent different paradigms about how the landscape should be represented: similarly named classes have different semantics, they have different meanings, and they represent different conceptualisations. In short LCM1990 and LCM2000 have different ontologies, where an ontology is defined as an explicit specification of a conceptualisation, and a conceptualisation is an abstract representation of the world (Gruber, 1993; Guarino, 1995).

The problem of incompatible ontologies originates in concepts of data sharing in computing science, where the translation “problem” is that an object described in one vocabulary may not correspond to the same object in another vocabulary. In the ontologies approach, the capacity for data integration and sharing depends on understanding the way that the data is conceptualised (Pundt and Bishr, 2002). Necessarily this approach extends beyond consideration of differences in data structure (such as pixel size or minimum mapping units) and seeks to understand differences in object conceptualisation (such as are found in their semantics; Uschold and Gruninger, 1996; Guarino, 1995).

We have used a knowledge engineering approach based on Look Up Tables (henceforth LUTs) to encapsulate and manipulate what is known about the semantic similarities and differences between the datasets. The aim is to overcome the semantic stumbling blocks to interoperability by formalising descriptions of how the semantics of the 1990 dataset relate to those of 2000. The expert descriptions in a Semantic LUT are compared to metadata descriptions of the LCM2000 spectral heterogeneity from a Spectral LUT. It is worth noting that we use the Level 3 LCM2000 product (that with the full set of published metadata), and the standard (or only) product for 1990.

2.3 Change and Inconsistency: preliminary work

In the work reported here, we are making no evaluation of the accuracy or otherwise of either dataset. Rather we are assessing the consistency of the 2000 classification and attribution, with how the land was classified in 1990. However, inaccuracy in one or other product is obviously one possible reason for inconsistency.

Therefore we define inconsistency in this context as whether the information for a particular land cover object (in this case a LCM2000 parcel) is inconsistent with the cover types within that parcel in 1990 when viewed through the lens of the Spectral and Semantic LUTs. If the semantic definition of the land cover types at a location in the 1990 map are inconsistent with those present in 2000, we can identify two possible causes of the change. Either the cover type at one time or the other is in error, or else the cover type on the ground has changed.

Previous analyses used a Euclidian distance calculated between 2 characterisations of the parcel based on the Spectral and Semantic LUTs. Parcels with the greatest distance (by proportion of the parcel area and in absolute terms) were identified. The pattern of vector directions was found to be related to the level of ontological change. This methodology is reported in full in Comber *et al.* (2003b, 2003c). Field visits showed that 26% of the parcels identified as inconsistent were believed to have actually changed since 1990. We also identified situations with meta-data inconsistencies (e.g. empty attribution fields) and small parcels that bore no relation to landscape objects on the ground. Filters were developed to eliminate such artefacts from analysis (Comber *et al.*, 2003d). A second tranche of analyses considered the filtered data and identified a second set of inconsistent parcels based on absolute vector distance (not proportion). Again a sample of these was visited in the field. The results showed 41% of these parcels were believed to have actually changed since 1990. The remainder (59%) were due to inconsistencies (errors of classification) in either the 1990 or the 2000 dataset. This work is reported in Comber *et al.* (submitted).

The aim here is to take the identification of inconsistency and change a step further. Using the knowledge gained from the field visits about the distribution of different types of inconsistency, the aim is to provide more robust inferences about the distribution of change. Information from the visited parcels from the Spectral and Semantic LUTs was used to predict inconsistency and to provide statistics about the probability of parcels so identified as being locales of actual change.

3. Spatial Data Integration

Early work on data integration (Shepherd, 1991) focussed on spatial characteristics:

- Format - typically raster versus vector representation;
- Granularity - scale, resolution and pixel size or level of detail of the processes or objects under investigation;
- Minimum mapping unit - which defines the lower areal limit for representing areas homogeneous with respect to some parameter such as land cover class. Areas smaller than this threshold are incorporated into adjacent areas of sufficient size.

Examples include the 1990 Countryside Survey in Great Britain (Wyatt *et al.*, 1993; Fuller *et al.*, 1998), the European Union 4th Framework Programme of Research on Environment and Climate (LANES, 1998), Nordic Landscape Monitoring (NordLaM) (Groom and Reed, 2001), and the FAO/Africover Land Cover Classification System (LCCS) (Jansen and Di Gregorio, 2000). Differences in the configuration of landscape elements were attributed to differences in the spatial characteristics of data.

More recent approaches have focussed on how derived land cover information may be integrated. Ahlqvist *et al.* (2000) used rough set theory to manage uncertainty between spatially coincident but semantically and conceptually divergent data. An uncertain set is specified using an upper and a lower approximation to indicate extremes of certitude. Similarly Kavouras and Kokla (2002) applied concept lattices to map between the semantic concepts from different data ontologies. Wyatt and Gerard (2001) proposed a standard set of attribute classifiers to overcome semantic differences between land cover classifications. These solutions were for idealised cases, where the semantics of the problem were well understood. For instance concept lattices require clear, non-overlapping and unambiguous identification of the attribute semantics of each classification scheme to identify subcategory objects. Whilst this works well for the example presented by Kavouras and Kokla (2002; urban / industrial land use) the LCM1990 and LCM2000 classes do not fulfil these criteria. Many classes have ambiguous definitions that overlap in various dimensions such as species composition, biogeography and reflectance (Comber *et al.*, 2001).

Frank (2001) describes a 5-tier ontology for GIS to integrate different ontologies in a unified system. Frank proposes consistency constraints, relating to each ontological tier, to help integrate data from different sources. For instance, he shows how appropriate formulations of reality must be selected according to the process of interest and how observations must be checked using measurement-based approaches. However, the constancy constraints for Frank's tier of socially constructed reality are only applicable for that context – the problem in translating different terms for lakes is given. Whilst the discussion provides a valuable insight into relating data of different ontologies, its applicability to our problem is limited by 3 factors. Firstly, it is difficult to test for consistency of the “knowledge” of the world as collected by different individuals, agencies, organizations or other “cognitive agents”. Secondly, Comber *et al.* (2003a) have shown the LCM2000 classification scheme to be more strongly socially constructed compared to that of LCM1990. Thirdly, the 5 tiers and their consistency constraints provide a framework for integrating different ontologies at design time, but they do not provide a *post hoc* solution for reconciling data sets such as the LCM2000-to-LCM1990 problem.

Devogele *et al.* (1998) formally correlate semantic concepts between the objects of two databases by defining correspondences. Once the correspondences have been found, an integrated schema can be built. However, there is uncertainty about relations between the LCM1990 and LCM2000 classifications and ambiguity in the way in which some LCM2000 land cover classes are defined. The authors note that there is the potential for semantic conflict and, interestingly, the final decision about whether a correspondence

holds or not “is the responsibility of database administrators, based on their knowledge of the semantics of the data” (Devogele *et al.*, 1998, 338). In Devogele’s view the database managers are the domain experts.

Establishing semantic linkages has consistently been identified as the bottleneck in developing translation or integration ontologies. The methodology described in the next two sections, which specifically addresses semantic relations, is suitable for *post hoc* application to a problem where the semantics are not clearly understood.

4. Background

4.1 Pre-ambles

The question we wanted to answer was whether the LCM2000 parcel represented a change compared to how the same area was classified in 1990. Our approach was to exploit the LCM2000 parcel based meta-data and to compare that with an inferred heterogeneity in LCM1990 data.

One of the LCM2000 attributes (‘PerPixList’) provides a description of the parcel spectral heterogeneity, listing the top 5 spectral subclasses within the parcel. This meta-data was the result of a production stream that was completely separate to and independent of the process by which the parcel was identified and then classified. Of greater significance was, firstly, the methodological similarity between LCM1990 and the PerPixList – both were the result of a standard supervised *per-pixel* maximum likelihood classification. Secondly, the PerPixList subclasses are broadly equivalent to the subclasses from which the LCM1990 classes were aggregated. Thirdly, accompanying the release of the standard Level 3 LCM2000 product was information about the expected overlap of the spectral subclasses of PerPixList with the Broad Habitat classification.

Therefore PerPixList meta-data offered a thematic and methodological equivalency with LCM1990. Information was available on how it related to the LCM2000 parcel classes from which it was possible to construct a table of relations between the spectral subclasses and Broad Habitat classes (i.e. LCM2000 meta-data classes to LCM2000 Broad Habitat classes). The semantics of LCM2000 and LCM1990 were related by an expert, familiar with both datasets. They described the pair-wise relations based on their understanding of the links between the semantics of the two classifications. In both cases the pair-wise relations were described in Look Up Tables (LUTs) scored to indicate those relationships that should not occur (i.e. were *unexpected*), those which were ambivalent (i.e. were *uncertain*) and those which should occur (i.e. were *expected*).

In summary, a direct comparison between LCM1990 and LCM2000 was not possible. The LCM2000 meta-data offered an equivalency between the PerPixList attribute (a record of parcel spectral heterogeneity) and the spectral subclasses from which the LCM1990 Target classes were derived. The link between the PerPixList spectral subclasses and the LCM2000 Broad Habitat classes was made by a *Spectral LUT* from published information. The link relating LCM2000 classes to LCM1990 classes was provided by an expert in a *Semantic LUT*. The LUTs allowed the 26 LCM2000 Broad

Habitat classes and the 25 LCM1990 Target classes to be related via the PerPixList attribute. This is shown in Figure 1.

4.2 Hypothesis

The Semantic and Spectral LUTs generated 2 characterisations of the parcel, one showing the extent to which the PerPixList attributes related to the descriptions of expected spectral overlap, the other indicating how many of the LCM1990 pixels within the parcel would be *Expected*, *Unexpected* and *Uncertain* according to an expert description of the semantics. Our hypothesis was that most parcels have not changed, but those which have are those whose characterisations are very different from the mean values for that Broad Habitat.

4.3 Parcel Characterisation

A triple of (Unexpected, Expected, Uncertain) was calculated for each parcel which described how its coincident LCM1990 classes related to its LCM2000 Broad Habitat class. Each intersecting LCM1990 pixel was compared in turn with the Semantic LUT:

- If the LUT recorded an Expected relation then the Expected score was incremented by one;
- If the LUT recorded an Unexpected relation then the Unexpected score was incremented by one;
- If the LUT recorded an ambivalent relation then the Uncertain score was incremented by one.

The second parcel triple was calculated by comparing the PerPixList attributes with the Spectral LUT in the same way. Comparing the *Unexpected* and *Expected* scores for 1990 and 2000 resulted in a 2-D vector for each parcel in that feature space. Note that the Uncertain scores were ignored for the moment.

5. Methods

In all the work (current and previous) the LCM2000 and LCM1990 data was for a 100km by 100km square corresponding to Ordnance Survey National Grid square SK, which includes an area from Leicester in the south to Sheffield in the north.

The results of field visits undertaken between December 2002 and March 2003, which are described in section 2.3, assessed whether the 1990 and 2000 classes could have been correct, and whether there had been any land cover change in the period. In a third programme of field validation we visited 200 further parcels at random and assessed their 2000 Broad Habitat class and whether there was evidence of a change since 1990. In total 361 parcels were visited.

The Semantic LUT was used to partition the visited parcels into subsets of *possible change* and *no change* in the following way:

- Proportional changes in Expected and Unexpected scores between 1990 and 2000 were interpreted as measures of Belief and Disbelief in the hypothesis that the cover types in a parcel had changed;
- A threshold derived from the combined Belief of the visited parcels were applied to the entire dataset to identify other areas of change;

- Uncertain scores (from both LUTs) were used to generate further support for possible change parcels. There was a higher belief in change for parcels with the smallest Uncertain scores.

5.1 Analysis of Visited Parcel Data

Each of the visited parcels was labelled in terms of whether change may have occurred (yes or no) and whether the 1990 and 2000 datasets were thought to be correct (yes or no). Five Broad Habitat classes that had at least 5 instances (parcels) of change were used to determine the combined Belief threshold.

5.2 Data Analysis

Parcels with an area less than 49 pixels (30625m^2) were excluded because this is the smallest sized parcel where the core area can exceed the periphery – parcel classification was made on the basis of the core area, but edge pixels are known to be particularly heterogeneous (Fuller *et al.*, 2002). For each of the five Broad Habitat classes, each parcel with an area greater than 49 pixels was treated in the following way:

- Proportional changes in the Unexpected (ΔU) and Expected (ΔE) scores between 1990 and 2000 were calculated;
- Uncertain scores at both times were derived;
- Therefore two values of Belief and Disbelief in a hypothesis of parcel change were generated from a standard cumulative distribution function using the class mean and standard deviation for ΔU and ΔE ;
- The beliefs and disbeliefs were combined using the Dempster-Shafer theory of evidence;
- Uncertainty information was used to generate further support for belief in the hypothesis that the parcels may have changed.

5.2.1 Calculation of parcel ΔU and ΔE and Uncertain scores

The parcel scores for 1990 described the number of LCM1990 pixels that were considered to be Unexpected, Expected and Uncertain by the expert, as indicated in the Semantic LUT. For example consider a LCM2000 parcel of Acid Grassland Broad Habitat with intersecting LCM1990 pixels as shown in Table 1. The 1990 scores are determined by calculating the proportion of pixels that represent each type of relation. In this case the parcel has an Unexpected score of 0.36 (161/446), an Expected score of 0.40 (178/446) and an Uncertain score of 0.24 (107/446). Scores are calculated from the parcel PerPixList attribute using the relations contained in the Spectral LUT. Table 2 shows the percentage of the top 5 spectral subclasses for the same example parcel. The Unexpected, Expected and Uncertain proportions are 0.32, 0.42 and 0.25 respectively.

The proportional changes in the Unexpected (ΔU) and Expected (ΔE) scores between 1990 and 2000 were calculated. The scores for ΔU and ΔE were converted into Beliefs in a hypothesis of parcel change with a standard cumulative distribution function using the class mean and standard deviation for ΔU and ΔE . For the Broad Habitat class of Acid Grassland the mean was 0.299 and the standard deviation was 0.321. The calculation of the Beliefs in ΔU and ΔE is shown in Table 3.

Uncertain scores were treated slightly differently because some classes had no Uncertain relations indicated in one or both of the LUTs. Where they did Uncertain scores in 2000 and in 1990 were calculated for each parcel via the Semantic and Spectral LUTs. The Uncertain scores were characterised as being ‘Low’ if they were less than the 25th percentile of the distribution for that Broad Habitat class, ‘High’ otherwise. Where no Uncertain relations were indicated in either LUT the parcel uncertainty was characterised as ‘None’. Table 4 describes how the Uncertain characterisations were combined to give qualitative statements of the change in Uncertain scores, ΔQ .

5.3 Calculating Combined Belief

Dempster-Shafer was used to combine beliefs from ΔU and ΔE because it is the only theory of evidence which allows for the explicit representation of uncertainty. Dempster-Shafer requires a pair of functions to assess a proposition: the belief function (Bel) and the plausibility function (Pls). The uncertainty (Unc) is the difference between the plausibility and belief. The belief is therefore a lower estimate of the support for a proposition and the plausibility an upper bound, or confidence band. Disbelief (Dis) is 1-Pls. They are related in the following way:

$$Bel + Dis + Unc = 1$$

Equation 1

The greater the increase in the Expected score (that is a positive ΔE) the higher our Belief that the parcel has changed; similarly the more the Unexpected score has decreased (that is, negative ΔU) the higher our Belief that the parcel has changed. Since values of ΔE and ΔU are both close to a Gaussian distribution we approximate a Belief value from the appropriate cumulative distribution function.

Evidence from ΔU and ΔE was combined using the Dempster-Shafer theorem:

$$Bel_{\Delta U \Delta E} = \frac{Bel_{\Delta U} Bel_{\Delta E} + Bel_{\Delta U} Unc_{\Delta E} + Bel_{\Delta E} Unc_{\Delta U}}{\beta}$$

Equation 2

where β is a normalising factor that ensures $Bel + Dis + Unc = 1$ (Equation 1) and

$$\beta = 1 - Bel_{\Delta U} Dis_{\Delta E} - Bel_{\Delta E} Dis_{\Delta U}$$

Equation 3

Our hypothesis was that the parcel has changed. The only alternative was that it had not. Therefore in this step there was no uncertainty to be allocated to the *frame of discernment*, or set of other *competing* hypotheses. Thus in this case Disbelief is 1-Belief and Uncertainty is zero. Because the uncertainty is zero this is equivalent to negation in standard probability theory.

Equations 2 and 3 allow a measure of combined belief and disbelief in a hypothesis of change to be calculated, Equation 4, for the data in Table 3, for the example parcel.

$$Bel_{\Delta U \Delta E} = \frac{0.210 * 0.581 + 0.210 * 0 + 0.581 * 0}{1 - 0.210 * 0.419 - 0.581 * 0.79} = 0.269$$

5.4 Abstract from Visited Parcels to Population

The combined Belief (from ΔU and ΔE) was calculated for all parcels in SK. We extracted the values for the parcels that have been visited. Visited parcels were identified as *correct change* or correct *no change*. Placing each parcel into histogram bins of combined belief values provided a simple way to partition the data and identify a threshold for change. The threshold was applied to the entire dataset to identify areas of possible change.

5.5 Filter using changes in Uncertain scores

Parcels of possible change identified from the combined evidence of ΔE and ΔU were given further ‘Strong’ support where ‘Low’ or ‘None’ Uncertain scores occurred at both times. In these cases most of the weight of evidence from the Semantic and Spectral LUTs was partitioned into Unexpected and Expected scores and where no uncertainty was indicated in either (or both) of the LUTs, ΔU and ΔE were exclusive.

6. Results

6.1 Analysis of Field Data

The distribution of the combined Belief (from ΔU and ΔE) for change for the visited parcels is shown in Table 5. The distributions indicate that the data can be partitioned by selecting a threshold for *change* and from this errors of commission and omission determined. Selecting a threshold of ≥ 0.9 Belief in a hypothesis of change results in error of omission of 21% and error of commission of 28%. That is the threshold of change of 0.9 correctly partitions 75% of the (error free) visited sites (26+26 out of 69).

6.2 Abstracting from Field Data – identifying change

The thresholds were applied to the LCM2000 data to identify possible change parcels (i.e. those with Belief above the threshold). An area to the north of Sheffield is shown by way of illustration which has a mixture of rural and urban land cover types and changes to woodland Broad Habitats (‘Coniferous Woodland’ and ‘Broadleaved Woodland’) and to the Broad Habitat of ‘Suburban Rural Development’. This area has been subject to de-industrialisation, road building with concomitant increased roadside verges, numbers of nature reserves and country parks.

6.2.1 Woodland Broad Habitats

Parcels identified as possible changes are shown in Figure 2. No uncertain relations for Coniferous Woodland and Broadleaved Woodlands Broad Habitats were indicated in the LUTs and no filtering due to differences in Uncertain scores was possible. The parcels in Figures 2a and 2b are those identified purely on the threshold of belief described above, Figure 2c shows the distribution of 1990 grass classes in red and Figure 2d shows heath classes in purple. There are two patterns:

- 1) The major area of existing woodland identified in 1990 in the south-eastern corner is now a nature reserve. Possible change parcels have been identified at its fringes.

- 2) New woodland areas have been established in a country park, running northwest to southeast, just to the north of the suburban area.

In both cases change to woodland in 2000 is from grass and heath classes as local recreational woodland activity increased. Confirmation of this change is provided by the Ordnance Survey backdrop, showing areas of woodland.

6.2.2 Suburban Broad Habitat

Parcels identified as changes are shown in Figures 3. The main urban area is the villages of Chapeltown and High Green. Without the filtering a large number of Suburban and Rural Development Broad Habitat parcels are identified as having changed (Figures 3a and 3b). Filtering for Low Uncertain scores at both times eliminated many of the parcels (Figure 3c). Most of the changes are from the LCM1990 arable class of 'Tilled Land' (Figure 3d) which is intuitively correct as urban developments in small towns are likely to be at the fringes of existing urban areas, where arable areas are likely to be found. However the pattern of distribution of the LCM1990 arable class in the urban area would suggest that there has been some spectral confusion in 1990 between arable and suburban classes.

7. Discussion

7.1 Discussion of the Results

Previous work has shown the use of semantic and spectral LUTs to be moderately successful in identifying change (Comber *et al.* submitted). Major failures were found to be due to error in either LCM1990 and/or LCM2000, typically due to common spectral confusions. In this work we have described an approach that extends the application of these expressions of expert opinion in order to predict specific locales of land cover change as a component of inconsistency between LCM1990 and LCM2000.

Any data comparison will always be subject to the limitations of data accuracy. If 2 maps are 75% accurate then *a priori* analysis reliability will be limited to just over 55% success. Analysis of maps that record very different land cover features in very different ways and have very different conceptualisations of the objects they seek to record will be further limited by such ontological differences. Because of these problems of semantic and methodological difference, the 2000 dataset is accompanied with a "health warning" against comparing it to its predecessor for change detection (Fuller *et al.*, 2002). Whilst we expect many of the parcels identified as having changed by this methodology to be false positives (for a hypothesis of change) from the field visits we are able to provide some measure of confidence in the form of errors of omission and commission. For instance we would expect 21% parcels that had actually changed to be missed by this approach, and 28% of the parcels identified as having changed, to not have done so. Given the stated accuracy of the data when compared to a sample field survey, these results are very encouraging.

Overcoming the data dissimilarity of LCM1990 and LCM2000 has raised a number of other issues:

Firstly, given the possibility of data error at either time it is important to consider the individual context within which a possible change is identified. Apparent changes from arable to urban in the middle of a town are likely to be a data artefact originating in spectral similarity of the two classes, whereas changes from grass to woodland represent a much more spectrally distinct change. It is not difficult to see how some simple rules could be developed and applied to filter possible spurious changes.

Secondly, where there are such spectral confusions we might be able to identify the *type* of evidence or data that would enable more robust inferences of urban change (for example) to be made. Such ‘Task Oriented’ approaches (i.e. not relying solely on the remotely sensed data to extract as much information as possible, but considering the nature of the problem in hand) to land cover change have been recommended by workers in remote sensing such as Foody (2002) and implemented in the work of Skelsey *et al.* (in press) and Comber *et al.* (in press). Remotely sensed data (and its derived products) are considered as only one of a number of useful datasets that can contribute to solving the problem.

A third issue is the different degrees of ontological change for different classes between LCM1990 and LCM2000. Woodland Broad Habitat classes are defined in broadly similar terms to their LCM1990 counterparts, whereas there are subtle but significant differences between the LCM1990 definition of the ‘Suburban / Rural Development’ Target class and the LCM2000 definition of the ‘Suburban / Rural Development’ Broad Habitat. These differences are reflected in the Semantic and Spectral LUTs which contain no pair-wise relations that were Uncertain for the woodland Broad Habitats, and some that were for the Suburban ones.

7.2 Discussion of the method

LCM1990 and LCM2000 provide very different land cover information regardless of any change on the ground. They have different minimum mapping units, have different accuracy distributions and record landscape features through land cover in very different ways. In most cases direct comparison is precluded as the impacts of structural difference means that attempts to relate any signal contained in each dataset gets lost in the noise of dissimilarity. Fuller *et al.* (2003) comment on one way forward as being that: “the vector structure of LCM2000 could be used to interrogate LCMGB (LCM1990) raster data and determine the previous cover dominance for each LCM2000 land parcel” (Fuller *et al.* 2003, 251). In this context both the LCM1990 pixels that intersect with the LCM2000 parcel and the parcel spectral heterogeneity provide comparable, independent descriptions of the parcel.

The method described here adopts the strategy outlined by Fuller *et al.* (2003) and extends it by incorporating some of the extensive LCM2000 object-based meta-data. In so doing it obviates extensive and sometimes tortuous probabilistic analyses of data characteristics combined to account for structural differences between the datasets. Such integration models do not exist. Whilst the results and analysis may be seen to rely on conjecture or anecdote in the form of the Semantic LUT, it is precisely because the

representational issues are both irreconcilable and impenetrable that we have developed this method.

7.3 Future Work

We have applied a Semantic LUT of an informed User under a scenario of idealised “semantic relations”. That is, one based on an understanding of how consistent the definitions (or concepts) behind the land cover class labels are. However other LUTs are possible from different perspectives:

- Semantic LUTs can be provided by different experts;
- a “change” LUT can indicate the transitions that an expert might expect to see happen to a land cover class over time;
- a “technical” LUT can describe some of the common technical issues of satellite image classification such as spectral confusion.

Future work will consider how further evidence about change from these sources can be combined.

8. Conclusions

We have shown that it is possible to reason about change using expressions of expert opinion of how the dataset semantics relate. We have robustly overcome ontological discord and identified locales of change with confidence limits. In the process some of the wider issue of using remotely sensed data to identify land cover have been highlighted, confirming the need for contextual information to be incorporated into remote sensing analyses.

The UK land cover mapping example shows that despite many and varied differences between the datasets it is possible combine data of different ontological pedigrees for specific applications. Any approach working with satellite derived land cover information that records moderate ecological detail such as LCM2000 and LCM1990, however, will be limited by the accuracy of the data as reported by many papers in the remote sensing literature.

Future land cover mapping exercises in the UK and elsewhere will record the landscape in different ways due to technological and political developments (Comber *et al.* 2003a). It is only possible to develop solutions to overcome these data limitations where data producers communicate some of their understanding and knowledge of the data in the form of metadata. The Centre for Ecology and Hydrology are to be lauded for including detailed object-based process history and measures of thematic heterogeneity as part of their standard products.

Acknowledgements

This paper would not have been possible without the assistance those involved in creating LCM2000, especially Geoff Smith. This paper describes work done within the REVIGIS project funded by the European Commission, Project Number IST-1999-14189. We wish to thank our partners in the project, especially Andrew Frank, Robert Jeansoulin, Alfred Stein, Nic Wilson, Mike Worboys and Barry Wyatt. The paper was greatly improved in addressing the comments of anonymous reviewers.

References

- Ahlqvist, O., Keukelaar, J. and Oukbir, K., 2000. Rough classification and accuracy assessment. *International Journal of Geographical Information Science*, **14**, 475-496.
- Comber, A.J., Fisher, P.F. and Wadsworth, R.A., 2002. Creating Spatial Information: Commissioning the UK Land Cover Map 2000. In *Advances in Spatial Data*, edited by Dianne Richardson and Peter van Oosterom (Berlin: Springer-Verlag), pp. 351-362.
- Comber, A. Fisher, P. and Wadsworth, R., 2003a. Actor Network Theory: a suitable framework to understand how land cover mapping projects develop? *Land Use Policy*, **20**: 299-309.
- Comber, A.J., Fisher, P.F. and Wadsworth, R.A., 2003b. A semantic statistical approach for identifying change from ontologically diverse land cover data. In *AGILE 2003, 5th AGILE conference on Geographic Information Science* 24th-26th April 2003. Edited by Michael Gould, Robert Laurini, Stephane Coulondre, (PPUR, Lausanne), pp. 123-131.
- Comber, A.J., Fisher, P.F., Wadsworth, R.A., 2003c. A Semantic Statistical Approach to Negotiating Heterogeneous Ontologies. In *Proceedings of the 2nd International Symposium on Spatial Data Quality* 19th-20th March 2003. Edited by Wenzong Shi, Michael Goodchild, Peter Fisher, (Hong Kong Polytechnic University, Hong Kong), pp. 349-358.
- Comber, A.J., Fisher, P.F., Wadsworth, R.A., 2003d. Identifying Land Cover Change Using a Semantic Statistical Approach: First Results. In *Proceedings of the 7th International Conference on GeoComputation*, 8th -10th September 2003, University of Southampton, UK.
- Comber, A.J., Law, A.N.R. and Lishman, J.R., 2001. Methodologies and Approaches for Automated Land Cover Change Detection. In *Innovations in GIS 8: Spatial Information and the Environment*, edited by P. Halls (London: Taylor and Francis), pp. 37-51.
- Comber, A.J., Law, A.N.R. and Lishman, J.R., in press. Application of knowledge for automated land cover change monitoring. *International Journal of Remote Sensing*.
- Comber, A.J., Fisher, P.F., Wadsworth, R.A., (submitted). Assessment of a Semantic Statistical Approach to Detecting Land Cover Change Using Inconsistent Data Sets. Submitted to *Photogrammetric Engineering and Remote Sensing*, July 2003.
- Devogele, T., Parent, C. and Spaccapietra, S., 1998. On spatial database integration *International Journal of Geographical Information Science*, **12**, 335-352.

Frank, A.U., 2001. Tiers of ontology and consistency constraints in geographical information systems, *International Journal of Geographical Information Science*, **15**, 667-678

Foody, G.M., 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**: 185-201.

Fuller, R.M., Groom, G.B. and Jones, A.R., 1994. The Land Cover Map of Great Britain: an automated classification of Landsat Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing*, **60**, 553-562.

Fuller, R.M., Wyatt, B.K. and Barr, C.J., 1998. Countryside Survey from ground and space: different perspectives, complementary results. *Journal of Environmental Management*, **54**, 101-126.

Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, R.A. and Thomson, A.G., 2002. Land Cover Map 2000: construction of a parcel-based vector map from satellite images. *Cartographic Journal*, **39**, 15-25.

Fuller, R.M., Smith, G.M., and Devereux, B.J., 2003. The characterisation and measurement of land cover change through remote sensing: problems in operational applications? *International Journal of Applied Earth Observation and Geoinformation*, **4**, 243-253.

Groom, G. and Reed, T., 2001. *Strategic Landscape Monitoring for the Nordic Countries*. Nordic Council of Ministers, TemaNord:523.

Gruber, T.R., 1993. A translation approach to portable ontologies. *Knowledge Acquisition*, **5**, 199-220.

Guarino, N., 1995. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, **43**, 625-640.

Jansen, L.J.M. and Di Gregorio, A., 2002. Parametric land cover and land-use classifications as tools for environmental change detection. *Agriculture Ecosystems and Environment*, **91**, 89-100.

Kavouras, M. and Kokla, M., 2002. A method for the formalization and integration of geographical categorizations. *International Journal of Geographical Information Science*, **16**, 439-453.

LANES 1998. *Development of a harmonised framework for multi-purpose land cover / land use information systems derived from earth observation data*. URL: <http://www.cesd.lu/LU/LU-Index-Frame-Set.htm> [available 22 June 2000].

Pundt, H. and Bishr, Y., 2002. Domain ontologies for data sharing-an example from environmental monitoring using field GIS. *Computers and Geosciences*, **28**, 95-102.

Shepherd, I.D.H., 1991. Information integration and GIS. In *Geographical Information Systems: Volume 1: Principles*. Edited by D.J.Maguire, M.F.Goodchild and D.W.Rhind (London: Longman) pp 337-360.

Skelsey, C., Law, A.N.R. and Lishman, J.R., in press. A system for monitoring land cover. *International Journal of Remote Sensing*, **24**(23).

Smith, G.M. and Fuller, R.M, 2002. Land Cover Map 2000 and meta-data at the land parcel level. In *Uncertainty in Remote Sensing and GIS*. Edited by G.M. Foody and P.M. Atkinson (London: John Wiley and Sons), pp 143-153.

Uschold, M. and Gruninger, M., (1996). Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, **11**(2): 93-136.

Wyatt, B.K. & Gerard, F.F.G. (2001). What's in a name? Approaches to the inter-comparison of land use and land cover classifications. In *Strategic landscape monitoring for the Nordic countries*. Edited by G.B. Groom and T. Reed (Nordic Council of Ministers, TemaNord:523), pp113-121.

Wyatt, B.K., Greatorex-Davies, N.G., Bunce, R.G.H., Fuller, R.M. and Hill, M.O., 1993. *The comparison of land cover definitions. Countryside 1990 Series: Volume 3*. (London: Department of the Environment).

Figures

Figure 1. Schema showing how the Spectral and Semantic LUTs link LCM2000 and LCM1990.

Figure 2. Woodland changes identified in blue and LCM1990 woodland pixels in green

- a) With some context from OS 1:50000 Raster scanned maps (© Crown Copyright Ordnance Survey. An EDINA Digimap / JISC supplied service);
- b) With the Belief threshold of 0.90;
- c) With LCM1990 woodlands and grass classes (red);
- d) With LCM1990 woodlands, grass and heath classes (purple);

Figure 3. Suburban changes identified in blue and LCM1990 suburban pixels in grey

- a) With some context from OS 1:50000 Raster scanned maps (© Crown Copyright Ordnance Survey. An EDINA Digimap / JISC supplied service);
- b) With the Belief threshold of 0.90;
- c) Filtered for parcels with Low Uncertain scores;
- d) With the LCM1990 arable class ('Tilled Land') in red;

Tables

Table 1. The number and type of intersecting LCM1990 pixels for a LCM2000 Acid Grassland parcel and their relation to that Broad Habitat from the Semantic LUT.

Table 2. The proportions of different spectral subclasses for a LCM2000 Acid Grassland parcel and their relation to that Broad Habitat from the Spectral LUT.

Table 3. An example of how individual Beliefs and Disbeliefs were calculated

Table 4. Belief in a hypothesis of change from simple characterisations of Uncertain scores in 1990 and 2000.

Table 5. Distributions of combined Belief for "correct" change and no change parcels

Figures

Figure 1. Schema showing how the Spectral and Semantic LUTs link LCM2000 and LCM1990.

Figure 2. Woodland changes identified in blue and LCM1990 woodland pixels in green
a) With some context from OS 1:50000 Raster scanned maps (© Crown Copyright Ordnance Survey. An EDINA Digimap / JISC supplied service);
b) With the Belief threshold of 0.90;
c) With LCM1990 woodlands and grass classes (red);
d) With LCM1990 woodlands, grass and heath classes (purple);

Figure 3. Suburban changes identified in blue and LCM1990 suburban pixels in grey

- a) With some context from OS 1:50000 Raster scanned maps (© Crown Copyright Ordnance Survey. An EDINA Digimap / JISC supplied service);
- b) With the Belief threshold of 0.90;
- c) Filtered for parcels with Low Uncertain scores;
- d) With the LCM1990 arable class ('Tilled Land') in red;

Tables

LCM1990 Target class	Number of Intersecting pixels	Semantic LUT relation
Blank	4	U
Mown / Grazed Turf	18	U
Pasture / Meadow / Amenity Grass	107	Q
Moorland Grass	178	E
Open Shrub Moor	22	U
Dense Shrub Moor	22	U
Bracken	31	U
Scrub / Orchard	85	U
Deciduous Woodland	8	U
Suburban / Rural Development	2	U
Total	446	

Table 1. The number and type of intersecting LCM1990 pixels for a LCM2000 Acid Grassland parcel and their relation to that Broad Habitat from the Semantic LUT.

PerPixList Spectral subclasses	Proportion	Spectral LUT relation
Ga_e	.42	E
Gbr_c	.25	Q
H_c	.14	U
Bo_b	.11	U
Hga_b	.07	U

Table 2. The proportions of different spectral subclasses for a LCM2000 Acid Grassland parcel and their relation to that Broad Habitat from the Spectral LUT.

	1990	2000	Change	Class mean	Class s.d.	Belief	Disbelief
ΔU	0.36	0.32	0.04	0.299	0.321	0.210	0.790
ΔE	0.40	0.42	-0.02	-0.037	0.280	0.581	0.419

Table 3. An example of how individual Beliefs and Disbeliefs were calculated

Support for Inconsistency		2000 Q	
		Low/None	High
1990 Q	Low/None	<i>Strong</i>	<i>Weak</i>
	High	<i>Moderate</i>	<i>Weak</i>

Table 4. Belief in a hypothesis of parcel inconsistency from simple characterisations of Uncertain scores in 1990 and 2000.

<i>Belief</i>	<i>Change & Correct</i>	<i>No Change & Correct</i>
0	0	0
0.1	3	11
0.2	0	7
0.3	1	2
0.4	0	3
0.5	0	2
0.6	1	1
0.7	1	0
0.8	1	0
0.9	0	4
1	26	6
<i>Total</i>	33	36

Table 5. Distributions of combined Belief for “correct” change and no change parcels