Journal of Land Use Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t724921301

Using semantics to clarify the conceptual confusion between land cover and land use: the example of 'forest'

A. J. Comber [a]; R. A. Wadsworth [b]; P. F. Fisher [c]
[a] Department of Geography, University of Leicester, Leicester, UK [b] CEH Monks Wood, Abbots Ripton, Cambridgeshire, UK [c] Department of Information Science, City University, London, UK

Online Publication Date: 01 January 2008

To cite this Article Comber, A. J., Wadsworth, R. A. and Fisher, P. F.(2008)'Using semantics to clarify the conceptual confusion between land cover and land use: the example of 'forest'',Journal of Land Use Science,3:2,185 — 198
To link to this Article: DOI: 10.1080/17474230802434187
URL: http://dx.doi.org/10.1080/17474230802434187

Journal of Land Use Science
Vol. 3, No. 2–3, June–September 2008, 185–198

Taylor & Francis
Taylor & Francis Group

Using semantics to clarify the conceptual confusion between *land cover* and *land use*: the example of 'forest'

A.J. Comber[a]*, R.A. Wadsworth[b] and P.F. Fisher[c]

[a]*Department of Geography, University of Leicester, Leicester, UK;* [b]*CEH Monks Wood, Abbots Ripton, Cambridgeshire, UK;* [c]*Department of Information Science, City University, London, UK*

(*Received February 2007*)

This article is concerned with data and classifications that confuse the concepts of land cover and land use. This conceptual confusion is problematic for data integration and has resulted in calls for the separation of land use and land cover from the global land monitoring community (GLP 2005). Text mining is used to unravel the different concepts embedded in land cover and land use semantics and applied to legal definitions of forest cover and use. Whilst the results show the distinct biological dimension to land cover descriptions and the socioeconomic character of land use, they reveal the deep degree of semantic confusion embedded in land cover and land use descriptions. The implications for this lack of internal semantic accuracy and consistency in land resource inventories are discussed and the case made for separating the concepts of land cover from land use.

Keywords: semantics; land cover; land use; text mining

Introduction

There are many ways of representing and describing land-based features. Historically, the overriding trend in land inventory, nationally and locally, has been to record information on *land use* (Fisher, Comber and Wadsworth 2005). Since the 1970s many land inventories have reported on *land cover* driven by the availability and machine processing of satellite imagery compared to the earlier demand and application driven surveys. In the process, land use and land cover have become interchangeable concepts often because of the demands of different agents and actors involved in the commissioning process. Wyatt and Gerard (2001) note that 'Land classifications commonly mix concepts of land cover, use and other environmental attributes such as soil type or climatic zone. This often leads to ambiguity and confusion'.

Confused thinking in the reporting of land information hinders the translation of information from different surveys. In the past this may not have mattered so much as it does today because obtaining land data often involved extensive dialogue with the data producer and the survey memoir provided extensive metadata (see Fisher 2003 for a discussion of this trend in natural resource inventories). Today there is often no memoir and data access is relatively easy and quick via various web-portals, e-Science initiatives and spatial data infrastructures (e.g. the computing grid and the EU's INSPIRE). These factors minimise the interaction between producers and users and therefore reduce the potential for them to clarify any inconsistencies in their shared understandings of, for instance, what they mean by the term 'forest'.

*Corresponding author. Email: ajc36@le.ac.uk

ISSN 1747-423X print/ISSN 1747-4248 online

DOI: 10.1080/17474230802434187
http://www.informaworld.com

The consequences of unavoidable inter-institutional negotiation over data specifications (see Comber, Fisher and Wadsworth 2002, 2003) and the spatial and spectral limitations of satellite imagery compared to field survey are three-fold. First, each individual member or institution on the steering committee of any big mapping project is forced to accept some degree of compromise over the specification of the land features to be identified. Second, the agreed classification is a hybrid of land cover and necessarily *inferred* land use. Third, the data users always have to 're-work' or manipulate the data in some way in order to incorporate the data into *their* analyses and to answer *their* questions. These problems are in part caused by the process of accountable data commissioning but mainly by the lack of 'data primitives' in land information, i.e. those dimensions or measurements that describe at the most fundamental level the processes under investigation. This is especially so in the case of land information derived from the classification of remotely sensed data.

In this article, we argue that that there is a need for clearer thinking in the way that information about the land is recorded and measured. This article presents a rationale for the divorce of the concepts of land cover from land use, particularly because they have been combined in so many survey initiatives over such a long period, their confusion is now the accepted paradigm. However, they represent different constructs for measuring features of interest on the earth's surface. To explicitly differentiate between land use and land cover, we analyse their semantics and concepts, as embedded in descriptions of *land cover* and *land use* definitions of 'forest'.

Background

Origins of the confusion of land cover and land use

The persistence and perpetuation of an inconsistent and counter-intuitive conceptual framework for measuring and monitoring land-based resources can be seen in many national and international programmes (e.g. the International Geosphere-Biosphere Programme (IGBP) Land Use and Cover Change programme, Nunes and Auge 1999). The origins of this illogical paradigm lie in the 1970s when the availability of medium resolution satellite imagery coincided with the wish of governments to better manage their land resource for a range of objectives as exemplified by the most influential work in this area (Anderson, Hardy, Roach and Witmer 1976). Traditionally, agencies concerned with tax, environmental management, planning, etc. had their own data specifications, data collection methodologies and classification schemes for recording land-based features. Anderson *et al.*'s (1976) outline of the US Geological Survey (USGS) land use and land cover classification specified a hybrid land use and land cover classification. In developing a standard national remotely sensed land classification, the confusion of land cover and land use was driven by a number of factors:

- The need to accommodate the existing classifications of different agencies;
- The ability to machine process remotely sensed imagery (i.e. statistical discrimination of land features);
- The need for consistent information that could be compared across time, space and at different levels of aggregation;
- The need to accommodate differing agency interests;
- The need for a 'resource oriented' classification to address the 95% of the national area not covered by previous 'people-oriented' classification of the standard land use coding manual (US Urban Renewal Administration, Housing and Home Finance Agency, and Bureau of Public Roads, Department of Commerce 1965).

Fisher *et al*. (2005) also note that the real cause of the inclusion of land use and land cover as a combined concept in the work of Anderson *et al*. (1976) may have been due to cartographic

objectives: the desire for a spatially even density of information, giving the same level of detail by identifying land cover classes within rangeland use, particularly adjacent to (relatively small) urban areas. Generally, it is cover in rural areas and use in urban areas that are the variables that give a finer grain to the landscape reporting.

Many subsequent inventories and initiatives have copied the land classification confusion of Anderson *et al*. (1976), developing hybrid classifications that confuse land use and land cover. Indeed the 'land cover/land use' couplet has become the *modus operandi* for many initiatives and most surveys where the differences between land cover and land use are frequently noted, but rarely accommodated. These include most national and international projects, for example, the classifiers developed in LCCS (Di Gregorio and Jansen 2000), the CORINE (Coordinating Information on the European Environment) classification (European Environment Agency (EEA) 2008), the GLC2000 (Bartholomé and Belward 2005) and the Land Use and Cover Change project (LUCC) as described by Lambin, Geist and Lepers (2003). In common with Anderson *et al*. (1976), such surveys seek to marry the capabilities of the satellite imagery with the varying objectives of the institutions on their steering committees.

Land use and land cover are often used interchangeably in many studies, surveys, programmes of research and reports. Whilst land use dynamics are the major determinant of land cover changes, they are in essence very different things. The fundamental difference between land cover and land use is that the former describes the physical characteristics of the earth's surface and the latter describes the activities upon it. Their differences are described in the reports of many mapping projects that incorporate a hybrid classification (e.g. Anderson *et al*. 1976; Di Gregorio and Jansen 2000) and many text books on remote sensing (e.g. Campbell 1981; Lillesand and Kiefer 2000). Despite widespread acknowledgement of the differences, the two concepts continue to be intertwined.

Land cover

Land cover is the physical material at the surface of the earth. Land covers include grass, asphalt, trees, bare ground, water, etc. There are two primary methods for capturing information on land cover: field survey and through analysis of remotely sensed imagery.

Field survey involves the detailed recording of land cover features. Typically, surveyors record attributes of floristic and landscape features by annotating base maps (traditionally paper but increasing digital). The emphasis of field based land cover surveys is usually ecological (e.g. the countryside surveys in the UK – Barr *et al*. 1993; Haines-Young *et al*. 2000) capturing information on the distribution of plant species, vegetation communities and phytosociological associations. Field surveys are time consuming and labour intensive but can capture data primitives. The countryside survey is unusual in that the collection of large volumes of species data is central to the philosophy and the reported land cover classes are then created by aggregating that data. This article is primarily concerned with the differentiation of land cover and land use features as recorded in remotely sensed imagery but cites the case of the nature of the information captured by field survey here as a contrast to that captured from remotely sensed data.

Land cover in remote sensing terms is the material which we see and which directly interacts with electromagnetic radiation and causes the level of reflected energy which determines the tone or the digital number at a location in an aerial photograph or satellite image. Tone or digital number in discreet wave bands alone may not be enough to distinguish between land features, but supported by empirical investigation, different land covers are increasingly separable, although context, pattern and texture may also be used (Lillesand and Kiefer 2000). The land cover classes that are discerned are clusters of pixels in the N wave bands that are within some defined statistical tolerance or distance in that feature space. Because of the nature of the information recorded by remotely sensed land cover and the way that the information is reported, the fundamental

dimensions of the data are not explicit. Yet land cover classes are not described in these terms but by their ecological or use characteristics. The reasons for this are many but in part due to the fact that the spectral characteristics of many features of interest (e.g. woodlands, urban areas) are not consistent across different scenes, sensors, landscape contexts and spatial scales (Comber, Law and Lishman 2004). Land cover is essential for environmental models (e.g. climatic and hydrologic), but is not directly useful for most policy and planning purposes (planning of the human or the natural environment), where land use is the relevant phenomenon.

Land use

Land use is a socioeconomic variable describing how people *utilise* the land. Urban and agricultural land uses are two of the most commonly recognised high-level classes of use. Residential land, sports grounds, commercial areas, etc. are also all land uses: land use describes socioeconomic activity. Nunes and Auge (1999, p. 37) describe land use as involving 'considerations of human behaviour, with particularly crucial roles played by decision makers, institutions, initial conditions of land cover'.

The recording of land use and land use classifications have a number of characteristics that result in the concepts and measurements of land use being more contested than for land cover. First, the relationship between land use and land cover is complex and cannot be directly inferred from remotely sensed data, although it frequently is, as indicated by the quote from Nunes and Auge (1999) above. Fisher *et al*. (2005) noted that land cover and land use have complex *many-to-many* relationships and cited the example of the cover 'grass' which can occur in a number of different land uses: sports grounds, urban parks, residential land, pasture, etc. Likewise, very few areas of homogenous land use have a single land cover. Furthermore, they pointed out that land use classifications do not necessarily fulfil the criteria of allocating features on the land surface uniquely into one class: a single point in space may quite legitimately have a number of different land uses at any given moment. Much land has multiple states of use which may be simultaneous or alternate: the field with cows may be the village football pitch at weekends; the reservoir may provide flood control but also angling, boating and water supply; and plantation forestry may also be used for several forms of recreation, including hunting and hiking, and even for grazing. The specification of any particular land use at any specific point in space is more problematic and contested because of these issues compared to land cover. For example, Hoeschele (2000) revealed serious differences in how land is used and regarded by indigenous commercial and subsistence farmers, on the one hand, and by forestry technocrats, on the other, in the Attappadi district of India.

Method

We suggest that the official definitions and descriptions of forests allow an insight into how the writers visualise a forest and what concepts are important to them. There is an issue as to whether concepts are mentioned (included) to help describe the habitat or to distinguish it from other habitats. It would be possible to manually extract concepts from the text description; however, such a process is likely to be inconsistent and we therefore prefer semi-automated methods.

In this work, a text mining approach was applied to the various types of forest use and cover descriptions. Generating information from text using automated computer techniques ('mining') such as natural language processing (NLP) remains a very difficult problem and is the subject of much current research including the establishment in the UK of worlds first 'National Centre for Text Mining' (Ananiadou, Chruszcz, Keane, Mcnaught and Watny 2005) and the development of sophisticated software (e.g. General Architecture for Text Engineering, http://gate.ac.uk/). The

complexity arises because a word or term can have many meanings depending on the context. Simple text mining is used by many Internet search engines, which rank the documents found by relevance derived from the similarity with phrase entered by the user. Knowledge discovery from text and text mining are data-mining techniques concerned with machine learning, natural language processing, information retrieval, information extraction and knowledge management (Karanikas and Theodoulidis 2002). Following Karanikas and Theodoulidis (2002), we use the term 'text mining' to refer to the process of extracting patterns from textual data and providing approaches to search, query and analyse unstructured textual data that is lacking in metadata.

Data-mining approaches such as genetic algorithms and neural networks have been used to extract association rules for environmental variables (e.g. Fielding 1999; Kampichler, Dzeroski and Wieland 2000; Maier and Dandy 2000; Guo, Kelly and Graham 2005; Phillips, Anderson and Schapire 2006). These methods search for patterns in data input and attempt to develop rules. They are therefore sensitive to data errors and many interesting patterns are the result of 'noise' as natural language processing is a very complex problem (compared to document categorisation which is a much simpler). Comparing multiple descriptions of forest land use and forest land cover is an extension to information retrieval and has been used to explore the semantic relations between different land cover data sets. Wadsworth *et al*. (2005) analysed the conceptual overlaps between different global land cover data, while in subsequent work they applied computer characterisation to the textual descriptions of two UK land cover maps in order to be able to integrate them. They found the integrative approach based on text mining to be more effective than human experts (Wadsworth, Comber and Fisher 2006).

Data

The website 'Definitions of forest, deforestation, afforestation, and reforestation' (Lund 2008) contains hundreds of different descriptions of forest activity and forest cover. These descriptions are organised into different definitional groupings of forest which were 'based upon literal interpretations of the definitions' (Lund 2008). The descriptions in the 'as a land use type' and 'as a land cover type' categories were extracted from the General, National and International groups for analysis (the state and provincial data were not analysed). The descriptions were only edited to get rid of the references to source of the data descriptions.

Initial processing

Each description was converted into a word list. Some words were gathered into terms or phrases, e.g. '25 m', 'per cent'.

Matrices were constructed for use and cover of descriptions (or classes) against words (or terms) used in the different forest descriptions, where the cells in the matrix contained the number of times each term appears in each class. The terms in the matrix were weighted using the 'tf.idf' (total frequency × inverse document frequency) scheme (Robertson and Jones 1976):

$$W_{ij} = \frac{n_i}{\sum n_i} \ln \frac{D}{n_j}.$$

Where

W_{ij} is the weight of the i^{th} word in the j^{th} class
n_i is the number of times the word appears in the j^{th} class
Σn_i is the total length of the j^{th} class description

D is the total number of classes
n_j is the number of classes containing the i^{th} word

The weighting has the effect that a word that appears in all class descriptions has a zero weight, but a word appearing frequently in a few short classes has a high weight.

Analysis of terms and semantics

The significance of the terms in the weighted matrices was evaluated using a standard principal components analysis (PCA) technique based on a correlation matrix. For both matrices relating to forest use and to forest cover descriptions, the PCA identified the following:

- The number of components that explained the variation;
- The amount of variation explained by each component;
- The terms with the greatest loading for each component.

The PCA was set to identify the components with eigenvalues greater than one. Within each component, the terms associated with highest component loadings were identified as those within 10% of the highest loading value. The weight of the component loading indicated relative strength of correlation to each principal component.

Results

As a preliminary analysis, the full set of forest descriptions (cover and use) were compared to determine the amount of overlap between the two concepts. Figure 1 shows that ordination plot of the first two components from this analysis. The descriptions of cover and use can be seen to

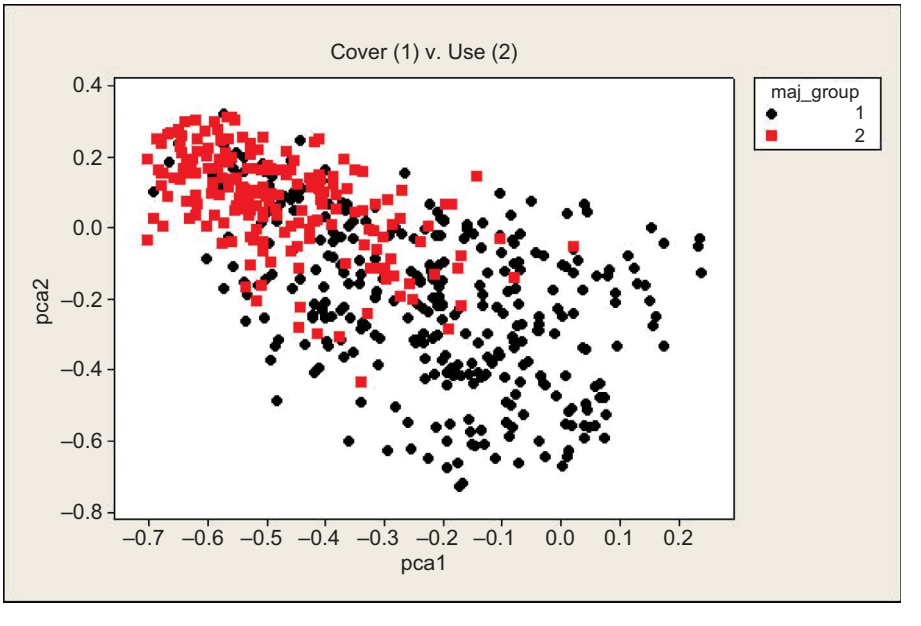

Figure 1. The overlap between use and cover in the first two principal components.

overlap, and in fact, land use descriptions seem to be a subset of cover descriptions. The remainder of this section will try to unpick the dimensions of this overlap.

The analyses describe the terms in each component (for each type of use and cover description) with the highest loading, i.e. they are within 10% of the maximum loading. The results describe the analysis of these *significant* terms. The numbers of principal components with eigenvalues greater than 1, terms with high loadings and the amount of variation in the weighted matrices explained by them are shown in Table 1 for each type of forest description.

Differences between forest cover and forest use

The overlapping terms shared by forest cover and forest use are shown in Figure 2. Examination of the terms shows that there are many terms unrelated to the general form or function of forests. The

Table 1. The original data, the number of terms and components with eigenvalues greater than 1 and significant terms for each type of forest description.

	Use	Cover
Total number of terms	2256	2379
Number of forest descriptions	194	320
Number of components with eigenvalues > 1	73	81
Percentage of variation explained	85	77
Significant Terms	149	170

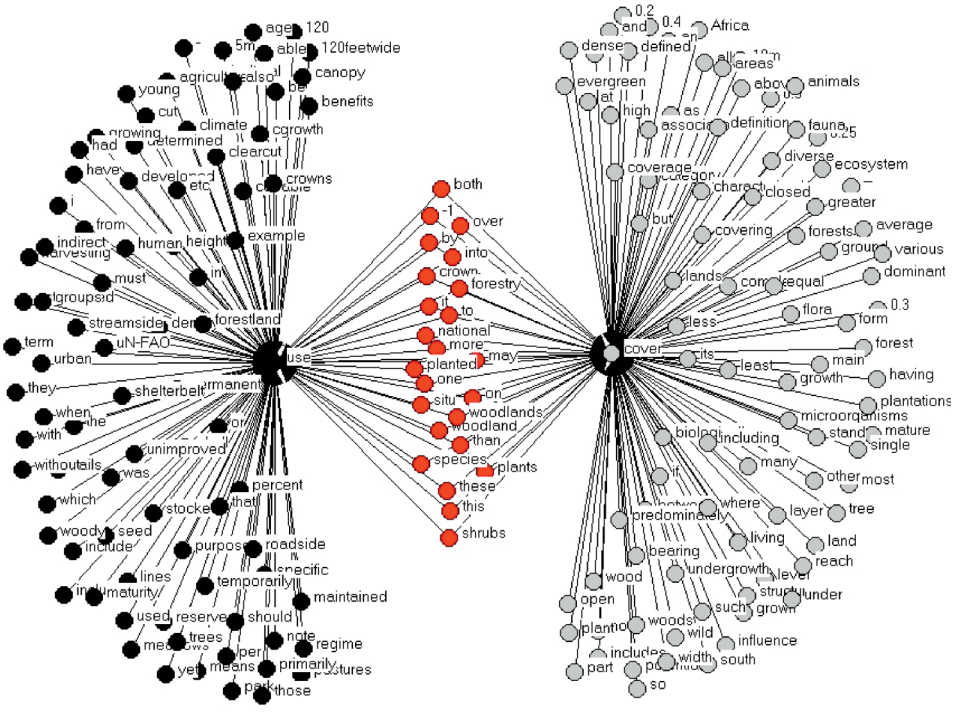

Figure 2. The overlapping terms shared by forest cover and forest use.

Table 2. Characterisation of the significant terms for forest use and cover (major differences between use and cover in bold).

	Use (%)	Cover (%)
Biological	9.4	**20.0**
Socioeconomic	**28.9**	6.5
Spatial/structural	15.4	**28.2**
Non-specific	46.3	45.3
Total	100.0	100.0

second part of the analysis was to separate the cover and use sets of descriptions and to examine the characteristics. The results of this analysis are presented in Table 2.

The significant terms identified during the PCA from the weighted matrices described in Section 3.2 were placed into three groups related to the general nature of land use and land cover descriptions: activity and surface, respectively. The object was to start to draw out the fundamental concepts from the data semantics associated with forest use and cover. Each term was characterised as being

- 'Biological' – those relating to vegetation, the environment and plants. This was expected to be more clearly associated with forest cover;
- 'Socioeconomic' – those relating to commercial activities, maintenance and management. This was expected to be more clearly associated with forest use;
- 'Spatial/Structural' – those relating to measurements specifications such as height, spatial extent and area as well as structural aspects such as crown closure. This was expected to be an important aspect within both sets of descriptions;

Other terms such as prepositions and common verbs (such as to be and to have) that could not be placed into the three categories were ignored. The distributions for the different categories of forest description are shown in Table 2. From the results in Table 2, the following statements can be made:

- Cover descriptions have a higher proportion of *Biological* terms than use ones;
- *Socioeconomic* are more frequent in use descriptions;
- Use has a lower proportion of *Spatial/Structural* terms than cover does;

Comparing the semantics of forest cover and use

The significant terms between forest use and forest cover were explored. The aim was to reveal the nature of the concepts and terms that were unique to cover and use and those that were shared. For each of the three characterisation groups (biological, socioeconomic, spatial), two analyses were performed.

(1) Land use and land cover were compared by looking at the overlap between the terms extracted at different scales;
(2) Land use and land cover were compared by looking at the overlapping terms when all use and all cover descriptions were compared as two groups.

Biological terms

Table 3 shows the set of significant biological terms unique to and shared between use and cover semantics. Cover has many more unique significant biological terms relating to the fact that it is the biology which defines the cover more than the use.

Table 3. The significant biological terms unique to and shared between forest cover and forest use descriptions.

Cover	Cover and use	Use
Animals	Growth	Climate
Bearing	Plants	Growing
Biologically	Shrubs	Maturity
Coverage	Species	Seed
Dominant		Trees
Ecosystem		Woody
Evergreen		
Fauna		
Flora		
Form		
Grown		
Living		
Mature		
Microorganisms		
Plant		
Tree		
Undergrowth		

Socioeconomic terms

Comparing significant use and cover terms shows that most of them are unique to use descriptions, few associated with cover and few that are shared, reflecting the socioeconomic basis of land use (Table 4).

Spatial/structural terms

When the spatial/structural terms are compared (Table 5), cover descriptions have many more significant spatial/structural terms associated with their semantics and descriptions than use. In particular the cover terms are associated with more quantitative terms and threshold values. Very few terms are shared.

Fuzzy cover and fuzzy use semantics

A membership function of the fuzziness of each term was calculated by considering the weightings generated in creating the inverse distance matrix and multiplying them by the average fuzzy

Table 4. The significant socioeconomic terms unique to and shared between forest cover and forest use descriptions

Cover	Cover and use	Use			
Plantations	Forestry	Able	Established	Permanent	Unimproved
Potential	Planted	Agricultural	Forestland	Purposes	Urban
Wild	Use	Agriculture	Harvesting	Regime	Used
	Woodland	Benefits	Human	Reserves	Young
	Woodlands	Capable	Maintained	Shelterbelt	
		Clearcut	Meadows	Stocked	
		Cut	Parks	Temporarily	
		Developed	Pastures	Trails	

Table 5. The significant spatial/structural terms unique to and shared between forest use and forest cover

	Cover	Cover and use	Use
0.2	Greater	−1	120
0.25	Ground	Crown	120 feet wide
0.3	High	One	5 m
0.4	Includes		Canopy
0.5	Including		Crowns
10 m	Land		Density
Above	Lands		Height
Area	Layer		Include
Areas	Open		Included
Association	Part		Per cent
Average	Predominately		Primarily
Characterised	Reach		Roadside
Closed	Stand		Streamside
Complex	Structure		
Covering	Under		
Dense	Width		
Diverse			

membership of the classes it appears in. The class-by-term matrix (containing the tf.idf weightings) was multiplied by the membership to land cover (and land use) of the classes. The weights for each term were summed to allow terms to be ranked according to how important they are for cover or use descriptions.

The membership function for each term described the degree of fuzzy use and of fuzzy cover. The two functions sum to unity (i.e. Fuzzy use = 1 – Fuzzy cover). Table 6 shows the 20 terms with the highest memberships for each. The use terms relate to management (e.g. 'harvest', 'ownership') and the cover terms describe biological aspects (e.g. 'biome', 'under-storey').

Summary of results

The semantics of forest use and forest cover overlap: the variation explained by the terms when they are weighted by a relative frequency measure ('tf.idf') shows that many significant terms are unique to either forest or use, and some are shared. When the nature of the semantics are explored, forest *cover* descriptions are more strongly associated with biological and structural or spatial terms and forest *use* semantics with socioeconomic terms. The fuzzy measures also show this pattern as expected.

Discussion

Data primitives

The notion of data primitives is to identify the fundamental building blocks or foundations that underpin the concepts of the phenomenon under investigation, such as land use and land cover. Identifying data primitives – the underlying data concepts, what the data mean and represent – allows data to be better integrated into analyses alternative to the original purpose of the data. It facilitates better data re-aggregation, data re-use and sharing, and enables the uncertainties of data integration for specific analysis to be quantified.

Table 6. The 20 terms with the highest membership functions (MFs) for both use and forest.

Terms	MF (cover)	Terms	MF (use)
Particularly	0.95	Non-timber	0.98
Multi-layered	0.90	Becoming	0.97
Historic	0.90	Harvest	0.97
Over-storey	0.90	Rough	0.96
Perimeter-vertical	0.90	Subdivided	0.96
Under-storey	0.88	Ownership	0.96
Stages	0.88	Maple	0.96
Biome	0.88	1992	0.96
Concentration	0.87	Property	0.96
Drainage	0.87	Mentioned	0.95
Non-urban	0.87	Pinus	0.95
Forest cover	0.86	Overgrown	0.95
Extensive	0.85	Groves	0.95
Predominately	0.84	395 million	0.95
Standpoint	0.84	Almost	0.95
Historically	0.83	Provided	0.95
Characterised	0.83	Light	0.95
Covering	0.83	Green	0.95
Predominate	0.82	Further	0.95
Dense	0.82	Avenues	0.94

In order to be able to effectively integrate data sets, data need to be consistent in terms of what they are reporting. Fisher *et al*. (2005) have described the internal data inconsistencies that may exist if concepts of land use and land cover are combined in Boolean classifications. Land use and land cover do not have a one-to-one relationship. Different covers may be subject to the same use and vice versa. Importantly, land uses may not be temporally consistent – alternative uses are possible for the same piece of land. Because the classification of land use describes social systems, it is much more open to contention (e.g. Hoeschele 2000). But that is not to say that the classification of land cover is in any way natural or predetermined. It is also socially constructed by the institutions and participants concerned with the mapping, and so there is still a level of contest in the schema to be used.

Integration activities incorporating land data that confuse and combine the concepts of cover and use have to overcome the internal data set inconsistency. This is problematic for models that incorporate land cover or land use data (e.g. evaluation of the impact of climate change, of the interaction between terrestrial and atmospheric environments, etc). For these reasons, the IGBP have called for the explicit separation of the concepts of land use and land cover. For example, the Global Land Project (GLP) science plan uses the conventional association of use with socio-economic systems and cover with biophysical systems (GLP 2005).

Results and method

The PCA of forest cover and use semantics indicates that use may be a subset of cover implying that in any shared vocabulary cover attributes may characterise use. The results of analysing individual terms show the association of land use with socioeconomic aspects of land management and of land cover with biological, structural and spatial ones. They also show the confusion between land cover and land use descriptions, although they both can be characterised by the terms

in their semantics and descriptions. However, this only serves to illustrate the difficulty of separating the concepts of forest use from forest cover in different classifications.

The results also show the current lack of primitives in land data. Whilst land cover or land use derived from remotely sensed imagery have spectral primitives – their position in spectral feature space – this is not how they are described. Their semantics describe their supposed characteristics on the ground – not the way that they were actually defined in the data. This is in contrast to information captured during field surveys describing the number and types of different species in plant communities. Land use data derived from reflectance is an anachronism unless that use is consistent in terms of cover, which it is not (Fisher *et al*. 2005). Instead of land use being related to unique positions in spectral feature space, land use can only be inferred from land cover due to the many-to-many relationships between use and cover.

Text mining with frequency and document size weighting has proved to be a useful tool for extracting the terms that contribute to the variation in class descriptions. This approach has been shown to be effective in separating differences in data semantics in many other applications (e.g. changes in soil classifications, land cover and vegetation communities; Wadsworth *et al*. (2008)). Applying text mining to the problem of extracting the fundamental dimensions associated forest use and forest cover has shown the extent to which forest use and forest cover *are* distinct in their semantics but also the considerable overlap or confusion that exists in the way that these features are conceived, or at least in the way that they are described.

Other considerations

In other work we have recommend that at least 100 words for each class description are needed for effective text mining of data semantics and concepts (Wadsworth *et al*. 2006). It is possible that for some of the definitions in Lund (2008), there were too few terms in use. A second issue relates to the way that the class descriptions were ordered: it is impossible to check in Lund (2008) and we have assumed in this work that his allocation to use and cover groups was consistent. A third consideration is that the forest definitions analysed in this work were the product of many authors in contrast to Wadsworth *et al*. (2006), who analysed descriptions created by a few. Methods of text analysis can be sensitive to writing style and to vocabulary in common use within an organisation. A final consideration is that there is genuine confusion over the concepts and descriptions of forest use and forest cover which may be so intertwined as to be conceptually inseparable.

Further work

The calculation of fuzzy weights raises a number of interesting areas of further work. The process of relating average word weight of a description to the fuzzy membership of the concept would allow fuzzy membership of the any new descriptions to be calculated. For instance, calculation of a regression equation of the average word ‘weight’ of a description against the fuzzy membership to the concept (description) to ‘cover’. This would offer an alternative to the use of thesauruses and controlled vocabularies, which impose a further level conceptual confusion. Developing such fuzzy measures using the outputs from simple text mining removes the problem of term context and local meaning that is problematic in natural language processing applications. An appropriate domain to investigate would be one which is global but has very different manifestations in different parts of the world such as ‘grazing land’. This includes land types such as forest (e.g. in India as above) to store fed beef cattle. There are also issues relating the use of ‘fuzzy regression’ and to the nature of the authors of the class descriptions that need to be investigated further to reveal any national patterns or native language group patterns. The authors are currently developing these analyses.

Conclusions

Lund (2008) quite admirably has catalogued many classifications of forest use and cover groups and included the class descriptions (semantics) of each data set with a reference. The hanging question from this analysis is whether these terms and these semantics are truly data primitives or are they simply words? It is difficult to state that this work has identified *the* unique terms associated with cover and use. However, strong differences in flavour may be discernable. The major findings of this work are two-fold:

First, the confusion between land use and cover is so embedded in many land data sets that the misunderstanding between these two concepts is perpetuated via their descriptions: the concepts of land use and land cover are misused everywhere.

Second, for consistency use descriptions should be concerned with the socioeconomic dimensions of land and cover with the biological ones. Through analysis of the data semantics for forest classifications, we have shown these dimensions to explain most of the variation in use and cover, respectively.

There is no disagreement amongst practitioners that a separation of use and cover is desirable: land use ought to describe the activities on the earth's surface and cover the material at the surface. However, in practice these concepts are frequently or usually confused, not only within the same classification or database but also in the way that individual use and cover classes are described. This is in part due to the legacy of Anderson *et al*. (1976) which admitted its confusion of use and cover in order to satisfy and reach consensus amongst multiple agencies, and in part due to the nature of classifying remotely sensed imagery. Such classification identifies areas that have similar statistical characteristics, as determined by their values in spectral space. This identifies areas of homogenous land cover. However, historically policy makers have been interested in activities and land use. The confusion between use and cover can therefore also be seen to be data driven: the cheap, frequent, extensive and easy availability of satellite imagery has resulted in the headlong rush for applications and the fudging of internal data consistency. Separation of the concepts of land use and land cover is needed to foster a culture of consistency in data recording in order to facilitate data integration and interoperability.

References

Ananiadou, S., Chruszcz, J., Keane, J., Mcnaught, J., and Watny, P. (2005), "The National Centre for Text Mining: Aims and Objectives," *Ariadne* 42, June 2005. *www.ariadne.ac.uk/issue42*

Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E. (1976), "A Land Use and Land Cover Classification System for Use with Remote Sensor Data," U.S. Geological Survey, Professional Paper 964, p. 28, Reston, VA.

Barr, C.J., Bunce, R.G.H., Clarke, R.T., Fuller, R.M., Furze, M.T., Gillespie, M.K., Groom, G.B., Hallam, C.J., Hornung, M., Howard, D.C., and Ness, M.J. (1993), *Countryside Survey 1990: Main Report. Countryside 1990 Series: Volume 2*, London: Department of the Environment.

Bartholomé, E., and Belward, A.S. (2005), "GLC2000: A New Approach to Global Land Cover Mapping from Earth Observation Data," *International Journal of Remote Sensing*, 26, 1959–1977.

Campbell, J.B. (1981), "Spatial Correlation-Effects upon Accuracy of Supervised Classification of Land Cover," *Photogrammetric Engineering and Remote Sensing*, 47, 355–363.

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*, eds. D. Richardson and P. van Oosterom, Berlin: Springer-Verlag, pp. 351–362.

——— (2003), "Actor Network Theory: A Suitable Framework to Understand How Land Cover Mapping Projects Develop?" *Land Use Policy*, 20, 299–309.

Comber, A.J., Law, A.N.R., and Lishman, J.R. (2004), "Application of Knowledge for Automated Land Cover Change Monitoring," *International Journal of Remote Sensing*, 25, 3177–3192.

Di Gregorio, A., and Jansen, L.J.M. (2000), *Land Cover Classification System* (*LCCS*): *Classification Concepts and user Manual*, Rome: Environment and Natural Resources Service (SDRN), FAO.

EEA, 2008. CORINE Land cover, a technical guide. http://reports.eea.europa.eu/CORØ_landcover/en/land_cover.pdf

Fielding, A.H. (1999), *Machine Learning Methods for Ecological Applications*, Norwell, MA: Kluwer.

Fisher, P.F. (2003), "Multimedia Reporting of the Results of Natural Resource Surveys," *Transactions in GIS*, 7, 309–324.

Fisher, P.F., Comber, A.J., and Wadsworth, R.A. (2005), "Land use and Land cover: Contradiction or Complement," in *Re-Presenting GIS*, eds. P. Fisher and D. Unwin, Chichester: Wiley, pp. 85–98.

GLP (2005), "Science Plan and Implementation Strategy," IGBP Report No. 53/IHDP Report No. 19. IGBP Secretariat, Stockholm, 64 pp.

Guo, Q.H., Kelly, M., and Graham, C.H. (2005), "Support Vector Machines for Predicting Distribution of Sudden Oak Death in California," *Ecological Modelling*, 182, 75–90.

Haines-Young, R.H., Barr, C.J., Black, H.I.J., Briggs, D.J., Bunce, R.G.H., Clarke, R.T., Cooper, A., Dawson, F.H., Firbank, L.G., Fuller, R.M., Furse, M.T., Gillespie, M.K., Hill, R., Hornung, M., Howard, D.C., Mccann, T., Morecroft, M.D., Petit, S., Sier, A.R.J., Smart, S.M., Smith, G.M., Stott, A.P., Stuart, R.C., and Watkins, J.W. (2000), *Accounting for Nature: Assessing Habitats in the UK Countryside*, London: DETR.

Hoeschele, W. (2000), "Geographic Information Engineering and Social Ground Truth in Attappadi, Kerala State, India," *Annals of the Association of American Geographers*, 90, 293–321.

Kampichler, C., Dzeroski, S., and Wieland, R. (2000), "The Application of Machine Learning Techniques to the Analysis of Soil Ecological Data Bases: Relationships between Habitat Features and Collembola Community Characteristics," *Soil Biology and Biochemistry*, 32, 197–209.

Karanikas, H., and Theodoulidis, B. (2002), "Knowledge Discovery in Text and Text Mining Software," Technical Report, Manchester: UMIST – CRIM, *http://www.crim.co.umist.ac.uk/parmenides/internal/docs/Karanikas_NLDB2002%20.pdf*

Lambin, E.F., Geist, H.J., and Lepers, E. (2003), "Dynamics of Land-use and Land-cover Change in Tropical Regions," *Annual Review of Environment and Resources*, 28, 205–241.

Lillesand, T.M., and Kiefer, R.W. (2000), *Remote Sensing and Image Interpretation* (4th ed.), Chichester: Wiley and Sons.

Lund, H.G. (2008), *Definitions of Forest, Deforestation, Afforestation, and Reforestation* [Online] Gainesville, VA: Forest Information Services. *www.home.comcast.net/~gyde/DEFpaper.htm*. Misc. pagination.

Maier, H.R., and Dandy, G.C. (2000), "Neural Networks for the Prediction and Forecasting of Water Resources Variables: A Review of Modelling Issues and Applications," *Environmental Modelling and Software*, 15, 101–124.

Nunes, C., and Auge, J.I. (1999), *International Geosphere-Biosphere Programme: A Study of Global Change of the International Council of Scientific Unions*, Stockholm: IGBP.

Phillips, S.J., Anderson, R.P., and Schapire, R.E. (2006), "Maximum Entropy Modeling of Species Geographic Distributions," *Ecological Modelling*, 190, 231–259.

Robertson, S.E., and Jones, S.K. (1976), "Relevance Weighting of Search Terms," *Journal of the American Society for Information Science*, 27, 129–146.

U.S. Urban Renewal Administration, Housing and Home Finance Agency, and Bureau of Public Roads, Department of Commerce (1965), *Standard Land Use Coding Manual: A Standard System for Identifying and Coding Land Use Activities* (SLUCM), Washington, D.C.: Government Printing Office.

Wadsworth, R.A, Comber, A.J., and Fisher, P.F. (2006), "Expert Knowledge and Embedded Knowledge: Or Why Long Rambling Class Descriptions are Useful," in *Progress in Sptial Data Handling, Proceedings of SDH 2006*, eds. A. Riedl, W. Kainz, and G. Elmes, Berlin: Springer, pp. 197–213.

Wadsworth, R.A., Fisher, P.F., Comber, A., George, C., Gerard, F., and Baltzer, H. (2005), "Use of Quantified Conceptual Overlaps to Reconcile Inconsistent Data Sets. Session 13 Conceptual and Cognitive Representation," in *Proceedings of GIS Planet 2005*, Estoril Portugal 30th May to 2nd June 2005. ISBN 972–97367-5–8. 13 pp.

Wadsworth, R.A., Comber, A.J., and Fisher, P.F., (2008), Probabalistic latent semantic analysis as a potential method for integrating spatial data concepts in Proceedings of the Colloquium for Andrew U. Frank's 60th birthday, (ed. Gerhard Navratil), Geo Info Series 39, Vienna, pp. 99–108.

Wyatt, B.K., and Gerard, F.F. (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*, ed. G. Groom, Copenhagen: Nordlam.