Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

A SEMANTIC STATISTICAL APPROACH FOR IDENTIFYING CHANGE FROM ONTOLOGICALLY DIVERS LAND COVER DATA

Alexis Comber[1], Peter Fisher[1] and Richard Wadsworth[2]

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

[2] Centre for Ecology and Hydrology, Monks Wood, Huntingdon, Cambs. UK

email: ajc36@le.ac.uk, pff1@le.ac.uk, rawad@ceh.ac.uk

1. INTRODUCTION

Repeat inventories of land resources are rarely executed with the same set of objectives and methodology due to technological developments and changes in policy (Comber *et al.*, 2002). One example of this data dissimilarity is the land cover mapping of Great Britain in 1990 (Fuller *et al.*, 1994) and 2000 (Fuller *et al.*, 2002). The significant differences between LCM1990 and LCM2000 are:

- The use of per-parcel rather than per-pixel classification;
- A different minimum mappable unit;
- Different knowledge-based correction procedures and ancillary data;
- Increased information about each object including, crucially, a processing history and a measure of spectral heterogeneity with LCM2000;
- Different meanings or interpretations of the class names, also different class names.

The problem of changing methodologies has implications for data re-use, data sharing and data interoperability as well as for time series monitoring, model development and the identification of loci of change. Different frameworks, protocols, standards and methodologies have been proposed and developed within the GIS community. These include notions of data sharing, access, integration and interoperability as articulated by Ahlqvist *et al.* (2000), Bishr (1998), Devogele *et al.* (1998), Frank (2001), Kavouras and Kokla (2002), and Visser *et al.* (2000), amongst others. A common theme to emerge from these papers is the identification of semantics as the bottleneck to translation between different data ontologies.

In this work we take the perspective of being interested in identifying locales of land cover change between 1990 and 2000. Due to differences outlined above LCM1990 and LCM2000 are not directly comparable: a matrix of correspondence between the two sets of land cover classes reveals a lack of 1:1 mappings between data concepts. We have used a knowledge engineering approach based on Look Up Tables – referred to as LUTs through this paper – to encapsulate and manipulate what is known about the semantic differences. This approach seeks to overcome the semantic stumbling blocks to achieving interoperability by formalising expert descriptions of how the semantics of one dataset relate to those of the other

2. SPATIAL DATA INTEGRATION

The problem of relating information that is coincident in some dimension has been tackled by a number of different research groups. Initial work identified the origins of the problem. An example is the Countryside Survey 1990 (CS90) this combined digital satellite information with field survey data to provide alternative estimates of land cover features. Differences in land cover data were consistently characterised as relating to timing, spatial artefacts, thematic attributes and statistical issues (Wyatt *et al.*, 1993; Cherrill *et al.*, 1995; Fuller *et al.*, 1998).

Towards the end of the 1990's the need to better conceptualise methods for spatial data integration was explicitly acknowledged and a series of different conceptual approaches were proposed. The problem of semantic interoperability is that geographic objects, stored in independent databases, may vary in their geometric or syntactic representation, in their class hierarchies and their semantics, even though they may refer to the same real world feature (Bishr, 1998; Harvey *et al.*, 1999).

The formal semantics of ontologies have been suggested to support the translation process between data sources (Visser *et al.*, 2002). Descriptions of ontologies can be found in Uschold and Gruninger (1996) and Guarino (1995). This approach originates in concepts of data sharing in computing science. The translation "problem" is that an object described in one language may not correspond to the same object in another language because of the way that social systems construct rules for their internal organizations (Frank, 2001). In the "ontologies approach", the capacity for data integration and sharing depends on understanding the way that the data is conceptualised.

To further develop the potential for web based data delivery and integration, sets of standardisation protocols have been proposed. One such is OpenGIS which aims to develop consensus on spatial data integration specifications amongst industry, academia and government agencies. Descriptions of the approach are found in Herring (1999) and at the OpenGIS website (http//:www.opengis.org).

Other work has presented methodologies to facilitate translation between data concepts. Ahlqvist *et al.* (2000) explore the use of rough set theory to represent uncertainty between spatially coincident but semantically divergent data. Kavouras and Kokla (2002) apply concept lattices to map between the semantic concepts from different data ontologies. Jones *et al.* (1999) apply Bayesian probabilities based on a variety of metrics of geometric and semantic similarity from overlaid vector maps to identify areas of change. Devogele *et al.* (1998) propose a approaches for conflict resolution when seeking to integrate information from different databases and for formally defining correspondence between the objects of two databases.

Many of these solutions presented are for idealised cases where the semantics of the problem are well understood. For instance the concept lattice approach of Kavouras and Kokla (2002) requires clear, non-overlapping and unambiguous identification of the attribute semantics of each classification scheme. Central to the approach proposed by Devogele *et al.* (1998) (relating semantic concepts) is the investigation of correspondence to precisely identify the correlations. The 5 ontological tiers and their consistency constraints articulated by Frank (2001) provide a framework for integrating different ontologies at design time. They do not provide a *post hoc* solution to the LCM2000-to-LCM1990 problem, where land cover classes are not so unambiguously defined (Comber *et al.*, 2001).

Establishing semantic linkages has consistently been identified as the bottleneck in developing translation or integration ontologies. We believe that the methodology described

in the next section is suitable for *post hoc* application and provides a way to circumvent the need for unambiguous knowledge of how the different semantics relate. Further, it provides a mechanism with which to explore how the semantics between two classification schemes are related.

3. STATISTICAL SEMANTIC METHODOLOGY

3.1 Overview

The statistical semantic approach relates the data concepts of the two datasets through an expert description of relations, a LUT. One dataset is chosen to be the Primary dataset, in our example, LCM2000. Information from the Secondary dataset (LCM1990) is related to the Primary dataset objects (LCM2000 polygons) as a set of attributes. The value of these attributes is determined by the type and number of the intersecting objects of the Secondary dataset (in our example, pixels counts of the LCM1990 land cover classes that intersect with the LCM2000 polygon). The LCM2000 Level 3 product includes meta-data about the spectral heterogeneity of each polygon, and information about the expected spectral overlap between different land cover classes. This publicly available information was used to create a second LUT of the expected relations between the polygon class and the polygon spectral attributes. Thus there are two descriptions of the polygon: semantic attributes created by intersecting LCM1990 and LCM2000 spectral attributes from the meta-data. Using the LUTs, it is possible to compare the polygon positions at two times and to calculate a movement vector in a feature space of Unexpectedness (x) and Expectedness (y). Those polygons with the largest vectors are identified as candidate areas of land cover change. The methodology is described below.

3.2 Extract the LCM2000 and LCM1990 data for each LCM2000 polygon

The 1990 raster dataset was combined with the LCM2000 vector data. The 2000 polygon structures were maintained and the 1990 LCM1990 data formed a set of attributes for the LCM2000 polygons. The result is two independent sets of information about each polygon: the coincident LCM1990 classes and the top five spectral subclasses of that polygon in LCM2000.

3.3 Construct LUTs to identify the expected relationship between the datasets

These were obtained from two different sources. Firstly, an informed user produced a "semantic LUT" based on their interpretation of the relationship between the LCM1990 classes and LCM2000 classes. Secondly, a LUT of the relations between the 2000 land cover classes and their spectral heterogeneity attributes was constructed from information published by the data producers about where the overlap between broad habitat spectral subclasses might occur. Each LUT uses a three valued logic of the relations:

-1 to indicate a relationship that should not occur;
0 to represent a relationship where the expert was ambivalent whether it was expected or unexpected;
+1 to indicate a relationship that should be expected.

3.4 Calculate coordinates for each polygon

These were calculated as follows: For each polygon the number of pixels in each of the 26 LCM1990 classes is recorded, each LCM1990 class is compared in turn with the expert semantic LUT:

- If the LUT has an expected relation (+1) then the Expected score is incremented by the number of pixels in that LCM1990 class;

- If the LUT has an unexpected relation (-1) then the Unexpected score is incremented by the number of pixels in that LCM1990 class;
- If the LUT has no relation (0) then the attribute (class) is ignored.

This results in a (Unexpectedness, Expectedness) tuple for each polygon. The tuples for the 2000-2000 relations were calculated in a similar manner from spectral LUT for each polygon. This provides information of how polygon spectral subclasses relate to its LCM2000 broad habitat class, via the published information about expected spectral overlap.

3.5 Determine the vector 1990 and 2000

The vector describes the Euclidian distance and direction of movement in the feature space of "Expectedness" and "Unexpectedness"

3.6 Analysis

The distance can be visualised by plotting the two locations on the Unexpectedness (x) and Expectedness (y) axes, and the movement by linking them. The results can then be analysed by the distance between the two sets of coordinates and the direction. The distance identifies polygons that have apparently changed and the direction allows their movement to be characterised. Subpopulations with similar vectors can be identified and their histories determined.

4. RESULTS

We analysed six broad habitat land cover classes to represent three typical scenarios of ontological difference: minimum change in ontologies between 1990 and 2000, some change and maximum changes. The 100 largest movements between 1990 and 2000 are shown. The direction of movement can be characterised according to the compass quadrants of NE, SE, SW and NW. The directions can be interpreted in the following way:

- NE: the evidence is increasingly conflicting, that is increased Unexpectedness and increased Expectedness;
- SE: towards increased Unexpectedness (2000) and decreased Expectedness (1990);
- SW: the evidence is increasingly indistinct, that is decreased Unexpectedness and decreased Expectedness;
- NW: towards increased Expectedness (2000) and decreased Unexpectedness (1990).

The results for six example classes are shown in Figure 1.

4.1 Scenario 1: minimum ontological change

Broadleaved Woodland.

Movement is NW, for 100/100 polygons (Figure 1 a). The LCM1990 attributes for these polygons contained many Unexpected "open" LCM1990 classes, such as Grass Heath, Moorland Grass, Mown / Grazed Turf and Tilled Land mixed with Expected Deciduous Woodland. In 2000 the spectral subclass attributes contained Expected attributes of Deciduous Woodland. Possible explanations of movement in this direction for this class are changes from sparse areas of broadleaved woodland that included grasslands and heaths, indications of maturation of woodland that were recently planted in 1990 and changes in the mapping unit.

Coniferous Woodland

Movement is mainly NW for 90/100 polygons (Figure 1 b). These polygons have different mixtures of Unexpected LCM1990 classes: Deciduous woodland with Dense Shrub Moor; Tilled Land with Grass heath; and Urban development with Suburban / Rural Development, all with some Pasture / meadow / amenity grass. In 2000 the spectral

subclass attributes contained Expected Coniferous spectral subclasses. Possible explanations include maturation of recently planted woods, spectral confusion with suburban classes and changes in the application of the mapping unit. A smaller set has movement NE for 10/100 polygons. These polygons have Unexpected 1990 attributes of Deciduous Woodland. In 2000 they contain both Expected spectral subclasses of Coniferous Woodland and Unexpected Deciduous Woodland ones. That is they have conflicting attribute information in 2000, relative to 1990, and may be indicative of a change in the ontology in LCM2000, where the coniferous class incorporates deciduous elements, or of the data quality: these may be uncertain coniferous regions in both or either LCM2000 and LCM1990.

4.2 Scenario 2: some ontological change

Arable Cereal

Movement is mainly NW for 81/100 polygons (Figure 1 c). This is due to the Unexpected 1990 classes of Suburban / Rural Development with some Tilled Land, Recently Felled and various grassland classes. The 2000 attributes are Expected spectral subclasses of Arable Cereal. It is uncertain whether this indicates an agricultural change on the ground (for instance increased cereal in place of pasture), a shift in ontology (e.g. mapping unit), or spectral confusion between grasses and cereals. Secondly there is movement NE for 19/100 polygons. All of these have Unexpected LCM1990 grass classes (Grass Heath, Mown Grass / Turf, Pasture / meadow / amenity grass) and Expected LCM2000 arable spectral subclasses with Unexpected Improved Grassland ones. That is they have conflicting attribute information in 2000, relative to 1990 which may be due to the spectral similarity of, for instance, juvenile cereal crops and grass or may be indicative of the changed ontology.

Suburban/ Rural Development

Movement is mainly NW for 89/100 polygons (Figure 1 d). This is due to Unexpected LCM1990 attributes of Tilled Land and Possible Urban Development, or Unexpected grass and Recently Felled classes, combined with Expected spectral subclasses in 2000. This could indicate urban expansion but may also suggest some spectral confusion between LCM1990 classes. A second set of 11/100 polygons shows movement SE. This is as a result of Expected classes (Suburban / Rural Development) with some Possible Urban Development in 1990, and Unexpected spectral subclasses in 2000 (Urban, Arable and grass). Whilst this might be due to change in the 1990 to 2000 interval, given the nature of the land covers and land uses it may also be indicative of spectral confusion or changes in the mapping unit. The class Suburban / Rural Development indicates built areas with some green space, and this is reflected in the spectral subclass attribution.

4.3 Scenario 3: changed ontologies

Acid Grass

The movements for this broad habitat class is very different to the previous classes with movement in all 4 directions (Figure 1 e). The 41/100 polygons that show movement NE have Unexpected semi-natural LCM1990 attributes (Moorland Grass, Open Shrub Moor, Dense Shrub Moor and Bracken) The 2000 attributes are conflicting, with Expected acid grassland spectral subclasses with Unexpected ones (Bog, Heath) and Possible ones (Other grass spectral subclasses).

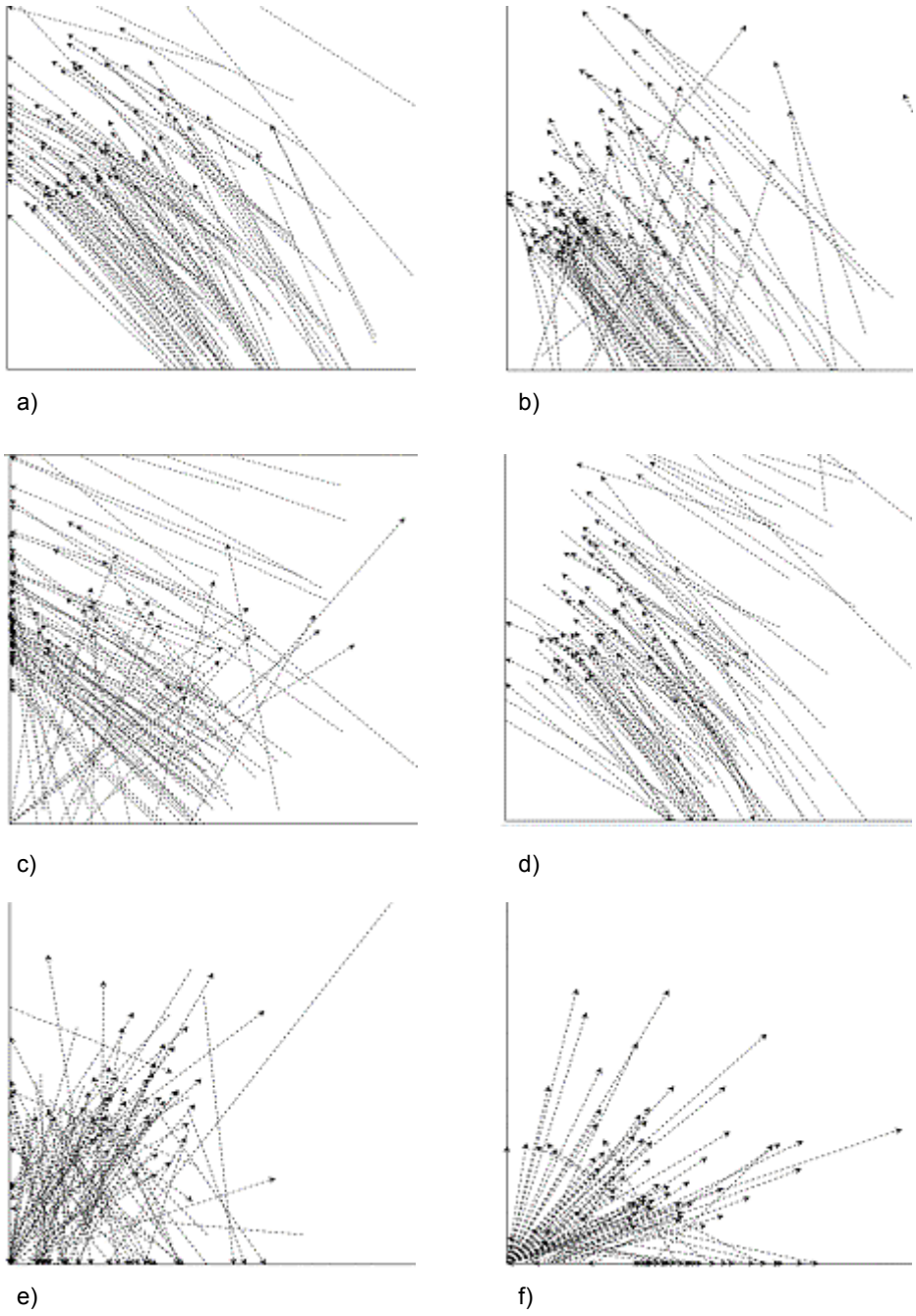

Fig. 1 The movement of the 100 largest vectors polygon between 1990 and 2000 in the dimensions of Expectedness (y), and Unexpectedness (x) a) Broadleaved Woodland, b) Coniferous Woodland, c) Arable Cereal, d) Suburban/ Rural Development, e) Acid Grass and f) Bog

Some 26/100 polygons show movement SE have grass classes in 1990 that were Expected by the Expert LUT (Grass Heath, Mown / Grazed Turf, Pasture / meadow / amenity grass) and in 2000 spectral subclasses that were Unexpected (Arable, Deciduous woodland, Inland Bare Ground) with some Possible (other grass classes) and some Expected (acid Grassland spectral subclasses)

Movement SW was shown by 17/100 polygons. The LCM1990 attributes (Mown Grass / Turf, Pasture / meadow / amenity grass) were Expected by the expert LUT and the 2000 spectral subclass attributes were indicated as being "Possible" by the Spectral subclass LUT (other grass classes).

Some 16/100 polygons moved NW because their 1990 attribution was of Moorland Grass which was Unexpected by the Expert LUT. Their 2000 spectral subclass attributes were Expected Acid Grassland subclasses.

It is difficult to interpret the meaning of these movements with regard to determining land cover change because of the heterogeneity of movement direction (the trends are not as strong as they are in the other classes), the very different ontology where areas of Acid Grassland were identified using a soil acidity mask in 2000 (meaning that the 2000 spectral subclasses do not relate to the final classification, movement NE and SE), and perhaps a failure on the part of the expert to understand the nature of the upland semi-natural land cover (movements SW and NW).

Bog

The major movement (87/100 polygons) for the Bog broad habitat is NE (Figure 1 f), due to Possible 1990 Possible (Moorland Grass, Open Shrub Moor, Dense Shrub Moor), and Unexpected 2000 attributes (grass, woodland, urban) combined with Expected and Possible ones (Bog, Heather). A second set (13/100 polygons) shows movement NW because of Unexpected attributes of Moorland Grass and Open Shrub Moor in 1990 and Possible Moorland spectral subclass attributes, with some Expected Bog one in 2000.

Any meaning for these directions of movement is difficult to discern. This is because all Bog polygons were subject to knowledge based corrections in LCM2000 and no Bog being identified in the SK area in 1990. As Bog was not classified on spectral characteristics, the spectral properties of the data are not related to the LCM2000 classification

5. DISCUSSION

We discuss the meaning of the results and show how the semantic statistical approach can be used to characterise relations between different ontologies. This step is a precursor to being able to combine data of different ontological pedigrees for a specific application, such as the identification of locales of change. The LUT provides an expression of the expected relations between the different data concepts: a linkage ontology. The different nature of these linkages and the different character of the large movements in each LCM2000 broad habitat class, can be characterised by plotting the movement vectors in a feature space of Expectedness and Unexpectedness.

Our hypothesis for the change application was that polygons that have moved a lot between the two dates were areas of actual land cover change. Analysis of the direction of movements allowed some inferences about change to be made where the ontological links between the data concepts were strong: movements that were towards increased Expectedness and decreased Unexpectedness (NW) are indicative of change. Using the LUT to calculate the distance and direction of movement gives an indication of the ontological links between data concepts. It highlights some of the artefacts specific to each of the datasets especially where the ontological links were weaker.

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.

References

[1] Ahlqvist O., Keukelaar, J. and Oukbir, K., 2000. Rough classification and accuracy assessment. *International Journal of Geographical Information Science*, 145, pp.475-496.

[2] Bishr, Y., 1998. Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science*, 124, pp.299-314.

[3] Cherrill, A.J., McClean, C., Fuller, R.M., 1995. A comparison of land cover types recognised in an ecological field survey of Northern England and in the first remotely sensed Land Cover Map of Great Britain. *Biological Conservation*, 71, pp. 313-323.

[4] Comber, A.J., Fisher, P.F. and Wadsworth, R.A., 2002. Creating Spatial Information, pp. Commissioning the UK Land Cover Map 2000. pp. 351-362 in *Advances in Spatial Data*, eds. Dianne Richardson and Peter van Oosterom. Springer-Verlag, Berlin.

[5] Comber, A.J., Law, A.N.R. and Lishman, J.R., 2001. Methodologies and Approaches for Automated Land Cover Change Detection, pp. 37-51 in *Innovations in GIS 8, pp. Spatial Information and the Environment*, ed P. Halls, Taylor and Francis, London.

[6] Devogele, T., Parent, C. and Spaccapietra, S., 1998. On spatial database integration *International Journal of Geographical Information Science*, 12 4, pp. 335-352.

[7] Frank, A.U., 2001. Tiers of ontology and consistency constraints in geographical information systems, *International Journal of Geographical Information Science*, 15 7, pp. 667-678.

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

[9] Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, R.A. and Thomson, A.G., 2002. Land Cover Map 2000, pp. construction of a parcel-based vector map from satellite images. *Cartographic Journal*, 391, pp. 15-25.

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

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

[12] Harvey, F., Kuhn, W., Pundt, H., Bishr, Y. and Riedemann, C., 1999. Semantic interoperability, pp. A central issue for sharing geographic information. *Annals of Regional Science* 332, pp. 213-232.

[13] Herring, J.R., 1999. The OpenGIS data model. *Photogrammetric Engineering and Remote Sensing*, 65 5, pp. 585-588.

[14] Jones, C.B., Ware, J.M. and Miller, D.R., 1999. A probabilistic approach to environmental change detection with area-class map data. *Integrated Spatial Databases, pp. Digital Images and GIS, pp. Lecture Notes in Computer Science*, 1737, pp. 122-136.

[15] Kavouras, M. and Kokla, M., 2002. A method for the formalization and integration of geographical categorizations. *International Journal of Geographical Information Science*, 16 5, pp. 439-453.

[16] Uschold, M. and Gruninger, M., 1996. Ontologies, pp. principles, methods and applications. The Knowledge Engineering Review, 112, pp. 93-136.

[17] Visser, U., Stuckenschmidt, H., Schuster G. and Vogele, T., 2002. Ontologies for geographic information processing . *Computers and Geosciences*, 28 1, pp. 103-117.

[18] 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, pp. Volume 3*. Department of the Environment, London.