Latent Analysis as a Potential Method for Integrating Spatial Data Concepts

Richard A. Wadsworth[1], Alexis J. Comber[2], Peter F. Fisher[2]

[1]CEH Lancaster, Bailrigg, Lancaster, LA1 4AP, UK. E-mail: rawad@ceh.ac.uk
[2]Department of Geography, University of Leicester, Leicester, UK. E-mail: ajc36@le.ac.uk, pff1@le.ac.uk

Abstract

In this paper we explore the use of Probabilistic Latent Analysis and Latent Dirichlet Allocation (LDA) as methods of latent analysis to quantifying semantic differences and similarities between categories. The results are promising, revealing 'hidden' or not easily discernable data concepts. LDA provides a 'bottom up' approach to interoperability problems for users in contrast to the 'top down' solutions provided by formal ontologies. We note the potential for a meta-problem of how to interpret the concepts and the need for further research to reconcile the top-down and bottom-up approaches.

1 Introduction

Many workers have identified differences in data semantics as the major barrier to data integration and interoperability (Frank 2001; Harvey et al. 1999; Pundt & Bishr 2002) and as Frank (2007a) notes, "In order to achieve interoperability in GIS, the meaning of data must be expressed in a compatible description". The crux of the problem is that the same real world features can be represented in many different ways. The suitability (quality) of a data set is therefore not static or absolute but depends on the

G. Navratil (eds.), *Research Trends in Geographic Information Science*,
Lecture Notes in Geoinformation and Cartography, DOI 10.1007/978-3-540-88244-2_9,

appropriateness of the representation in the context of the user's needs (Frank et al. 2004; Frank 2007b).

Many large datasets depend on multi-disciplinary teams whose members have different conceptualizations of the phenomena being recorded, and who are funded by Research Councils, Government Departments and Conservation Agencies etc who bring their own set of policy, scientific, financial and ethical concerns to the process. The difficulty in achieving interoperability in this context has not been helped by it becoming enmeshed in narrow technical issues related to discovery level metadata and metadata reporting standards.

A "top-down" approach to interoperability might start with the formal assertion that Newtonian physics and Euclidian geometry are sufficient (Frank 2003) and proceed to the development of ontologies, taxonomies and controlled vocabularies into which real data may be placed. We adopt a "bottom-up" approach and consider interoperability from the standpoint of a (naive) data user. We want to know; what the data "labels" *mean*, how the categories are related to each other and did the data producer have the same conceptual understanding of the phenomenon as the user? The final, and perhaps most difficult, task of bridging the top-down and bottom-up approaches has yet to be attempted within both formal ontology research activities such as OWL and emerging e-science infrastructures such as INSPIRE.

In particular we are concerned with *consistency* and *similarity* between data objects and how this affects a user's analysis (Comber et al. 2006). This paper proposes using a text mining approach called Latent Dirichlet Allocation (LDA) (Blei et al. 2003) (a development of Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999a,b)) to extract or infer the data concepts contained in written descriptions of spatio-environmental information.

2 Estimating Semantic Consistency

Estimating Semantic consistency can be done in various ways:

- **Declarative Approaches**: Rules (typically *If … Then ... Else ...*), are used to characterize relationships between objects. Generating rules is difficult, time consuming, and error prone. Rules may be inconsistent through error or because non-monotonic logic applies (consider the *if ... then ... else* rules of the children's game rock-paper-scissors).
- **Semantic Look Up tables**: Relationships are encoded in tables (matrices). Comber et al. (2004a,b; 2005a,b) used expert opinion to encode consistency as "expected", "uncertain" and "unexpected"

relationships in a successful attempt to compare two Land Cover Maps – a problem the data producers warned users was intractable. Wadsworth et al. (2005) decomposed land cover attributed into data primitives before re-integrating them to explore inconsistencies between three land cover maps of Siberia. Fritz & See (2005) used fuzzy logic to average the response of a group of experts.

- **Statistical approaches**: Foody (2004), Hagen (2003), Csilliag & Boots (2004) used statistical analysis to compare alternative representations of the same phenomenon in attempts to highlight the locations where variables are incompatible. Kampichler et al. (2000), Maier & Dandy (2000), Guo et al. (2005) and Phillips et al. (2006) made use of Genetic Algorithms and Neural Networks for similar purposes. These approaches are not always robust in the face of "noise".

The first two approaches (declarative and semantic) rely on the interaction with domain experts (knowledge engineering). The third method requires the user to already have a significant amount of both data sets, while we assume the user may want to perform an assessment before obtaining the data. As experts are not always available we want to try and extract the knowledge that they have "stored" in written descriptions. NLP (natural language processing) (Jurafsky and Martin 2008), especially of scientific texts, is a very complex problem but document categorization and information retrieval making the "bag-of-words" assumption is a much simpler problem. We adapted the work of Lin (1997) and Honkela (1997) to look at the similarity between categories rather than documents (Wadsworth et al. 2006). In an attempt to understand why two categories might be considered similar we are now investigating the potential of two Latent Semantic Analysis techniques (PLSA and LDA) (Hofmann 1999a,b; Blei et al. 2003).

3 Methods

In Latent Analysis the assumption is that there are underlying and unobserved variables (the latent variables) that can be used to explain an observed pattern. In Latent Semantic Analysis the pattern is the frequency of words in documents and the latent variables are concepts (ideas) described in the documents. We can observe the relationship between the documents and words and we want to uncover the latent concepts that can explain the distribution of words in documents. Probabilistic Latent Semantic Analysis (PLSA) was proposed by Hofmann (1999a,b) as a "generative" model of latent analysis; the joint probability that a word (w) and document (d) co-occur (P(d,w)) is a function of two conditional probabilities; that the

document contains a concept (z) (P(z|d)) and that the word is associated with that concept (P(w|z) (equation 1)

$$P(d,w) = P(d)\sum_{z \in Z} P(w \mid z)P(z \mid d) \tag{1}$$

Because we know the frequency of the words in documents (*n(d,w)*) it is possible to rearrange the probabilities to develop an iterative expectation maximization scheme to estimate all the probabilities. The expectation step generates P(z|d,w) while the maximization step calculates P(w|z), P(d|z) and P(z).

When using PLSA there can be problems with "over-fitting" so Hofmann (1999a) proposes using a variation on simulated annealing (called tempered annealing) to prevent this. Unfortunately the tempered annealing requires a "hold out" of test data and most of our data sets are too short to allow this. An alternative approach is to assume that the very skewed frequency distribution of words in documents follow a known distribution; such an assumption leads to a technique called Latent Dirichlet Allocation (Blei et al 2003). Implementations of LDA are available for free in both C and Java (http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation).

Deciding how many latent variables exist is analogous to determining how many classes exist in a fuzzy classification scheme (like c-means). Making the assumption that the probabilities are like membership functions then the indices proposed by Roubens (1982) can be applied. In our implementation of PLSA we used both the Fuzziness Performance Index (FPI) and the Modified Partition Entropy (MPE) to estimate the "best" number of classes from where the sum of the two indices is at a minimum. For consistency we used this "best" number in our explorations of LDA.

Because of restrictions on space we present the results of LDA for only three data sets; the Land Cover Map of Great Britain (LCMGB; Fuller et al. 1994) class descriptions, USDA Soil Orders (Soil Survey Staff 1999) and the abstracts of 677 refereed papers in the International Journal of Geographic Information Science (IJGIS).

4 Results

4.1 Number of Latent Variables in a Data Set

With the PLSA the optimum number of latent variables in the LCMGB land cover example is about 12 (there are 25 categories); this is the minimum of the combined FPI and MPE (Roubens 1982). Because the process may converge to a local minima several trials need to be conducted; Fig. 1 shows the results of five trials. Seven latent variables were found in 12

categories of soil orders; while ten themes were specified (not estimated) for the journal abstract example.

Estimating the "correct" number of latent variables in larger data sets and with LDA is more problematic. With large data sets like the abstracts from IJGIS a hierarchical approach might be preferred, with higher levels of the hierarchy showing the broad trends and lower levels breaking down the finer details and changes over time.

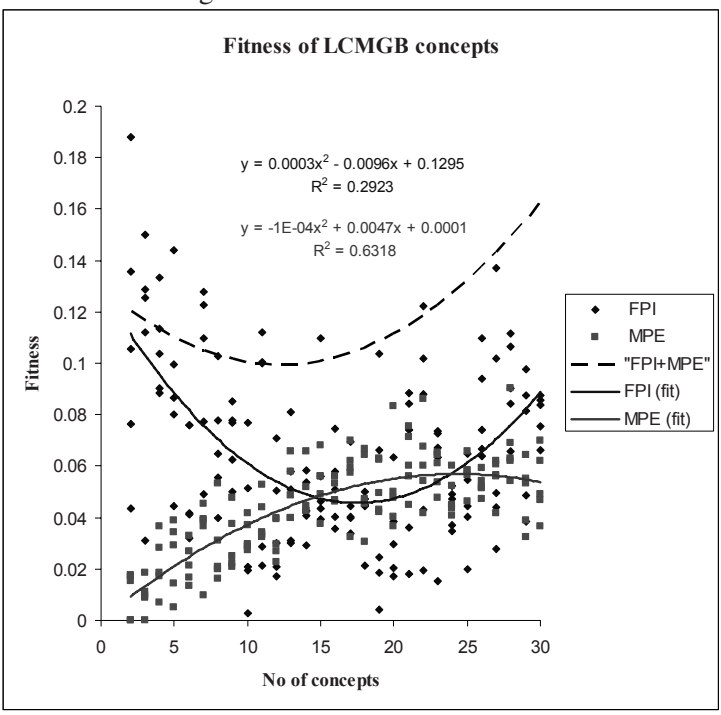

Fig. 1. Fitness measures used to determine optimum number of latent variables (concepts) in the LCMGB categories using PLSA.

4.2 Latent Variables Uncovered by Latent Analysis

Table 1 shows the relationship between the 25 LCMGB target classes and the twelve themes (latent variables) found by LDA.

Table 1. LDA of LCMGB class descriptions into 12 Topics

Topic	LCMGB class(es)	Distinctive Terms (the five words with the highest probabilities)

A	6 Mown Grazed Turf	Grazed Mown turf swards amenity
B	5 Grass Heath 8 Rough Marsh Grass 12 Bracken	Grassland Species Grass Lowland winter
C	20 Suburban & rural development 21 urban development	Rural Permanent vegetation development developments
D	1 sea estuary 2 inland water	Waters inland sea point water
E	3 coastal bare 4 Saltmarsh	High tides tide water lower
F	19 Ruderal weed 23 Felled Forest	Felled ruderal rough bare ground
G	7 Meadow Verge Semi natural swards	Swards grasslands semi-natural agrostis hay
H	15 Deciduous woodland 16 Coniferous evergreen woodland	Bare deciduous coastal evergreen woodland
I	10 Open shrub moor 11 Dense shrub moor 13 Dense Shrub Heath 25 Open Shrub Heath	Shrub grass dense heath moorland
J	0 unclassified	Cover types some data 25m
K	18 Tilled land (arable crops) 22 Inland bare ground	Bare ground land soil natural
L	14 Scrub Orchard 17 Upland bog 24 Lowland bog	Scrub upland grass lowland species

Table 2 shows the results of specifying 7 latent variables for the USDA Soils orders data sets, both the most probable terms and those with the highest fidelity are listed.

Table 2. Latent variables in the USDA Soil Orders.

Latent variable	Soil orders	10 most probable terms	10 most specific terms
A	Vertisols	vertisols, cracks, open, cropland, united, drainage, low, temperature, vegetation, xererts	vertisols, cracks, open, xererts, system, conductivity, hydraulic, installed, permeability, presents
B	Mollisols	mollisols, temperature,	ustolls, xerolls, cryolls, rich,

		vegetation, moisture, united, forest, grass, epipedon, cropland, plains	tall, addition, albolls, aquolls, drought, limestone,
C	Aridisols Entisols	united, aridisols, diagnostic, temperature, habitat, wildlife, moisture, rangeland, 100_cm, entisols,	psamments, aquents, orthents, recent, boundary, salts, arents, fluvents, sorted, weathering,
D	Gelisols Spodosols	spodic, organic, materials, spodosols, gelisols, united, permafrost, matter, under, alaska,	gelisols, aquods, cryoturbation, orthods, cryods, ice, gelic, iron, applied, good,
E	Histosols Ultisols	organic, united, vegetation, materials, forest, ultisols, cropland, moisture, drained, argillic,	histosols, bulk, density, udults, ultisol, ustults, botanic, decomposed, fiber, fingers,
F	Andisols Inceptisols	epipedon, forest, united, temperature, vegetation, inceptisols, cambic, moisture, ochric, deposits,	almost, tightly, ustands, xerands, aquepts, plaggen, torrands, udands, udepts, xerepts,
G	Alfisols Oxisols	united, moisture, oxisols, vegetation, forest, temperature, crops, association, udalfs, cropland,	oxisols, association, udalfs, xeralfs, alfisols, aqualfs, deciduous, rare, ustalfs, believed

Table 3 shows the results of applying the LDA to 677 abstracts of refereed papers in IJGIS.

Table 3. Abstracts from IJGIS grouped into 10 “themes” by LDA

Cluster	Typical title of a papers	Distinctive words (words with a high probability and high fidelity)
A	Accuracy assessment of digital elevation models using a non parametric approach	Error accuracy errors dem elevation flow propagation estimates interpolation mean input monte carlo terrain cent source positional average square confidence
B	Assessing farmland dynamics and land degradation on Sahelian landscapes using remotely sensed and socioeconomic data	Cover urban neural sdi imagery aggregation pixel agricultural metrics indices agent suitability social sds class city artificial nitrogen landscapes trend

C	A Voronoi based 9 intersection model for spatial relations	Relations topological voronoi tree query join boundary indexing intersection topology diagram relation metric hierarchical graph structures index building formal indices
D	A general model of watershed extraction and representation using globally optimal flow paths and up slope contributing areas	Terrain elevation visibility topographic parallel triangulation dems dem interpolation delaunay surfaces tin parameters variable triangulated irregular aspect channel radiation paths
E	Comparing area and shape distortion on polyhedral based recursive partitions of the sphere	Fuzzy query vague operators uncertain crisp precision arc text gps views membership info geo dbms position insurance soft view shell
F	Analysis of land use drivers at the watershed and household level: Linking two paradigms at the Philippine forest fringe	Soil forest crime erosion regression units risk fuzzy index factors vegetation kappa class moisture predictions loss landslide expert cover membership
G	TERRA VISION the integration of scientific analysis into the decision making process	Urban support criteria ca growth cellular automata factors suitability programming economic vulnerability sdss sensitivity group parameters making transition integrated sustainable
H	A proposed framework for feature level geospatial data sharing: a case study for transportation network data	Temporal spatio phenomena geography generic dynamics current census distributed events event internet matrix agents geospatial individuals behaviour spatiotemporal relationship transportation
I	Data gathering strategies for social behavioral research about participatory geographical information system use	Technology project national government state science community technical technologies current activities social benefits article discussion role countries support education efforts
J	Colour coded pixel based highly interactive Web mapping for georeferenced data exploration	Cartographic generalization interactive path datasets search rough mining solution task ontology categorical original discovery categories interpretation display exploration facility base

5 Discussion and Conclusions

Although the description of the LCMGB classes are rather short Latent Analysis has managed to identify some reasonable concepts (reasonable in the eyes of a domain expert). Unfortunately, some of the concepts are rather more difficult to interpret and may reflect statistical artifacts or the lack of words to process. A problem with stochastic approaches like the PLSA is that repeated “runs” on the same data set do not always result in the same groups being “discovered”; in this respect LDA is much more stable. When applying the approach to other data sets we have had mixed results. In descriptions of soil orders the main problem facing the non-expert is being unfamiliar with the terms used, but, by using latent analysis first it helps the user to identify which of these unfamiliar terms are likely to be the most important and therefore should be “de-coded” first; for an

expert it may help understand the subtleties of the categorization process used. When applying the approach to the abstracts from the IJGIS the method produced apparently interpretable "clusters", however, the size of the data set and the difficulty in deciding what constitutes an appropriate number of clusters suggests that a more hierarchical approach might be better.

Where human domain experts exist then knowledge engineering methods can codify their expertise in ways that make inter-operability a practical proposition. Domain experts may not exist or may not be accessible (through time constraints or geography) in those cases where domain experts have expressed their expertise through *long* textual descriptions text mining can produce acceptable estimates of semantic similarity. A "reconnaissance" assessment of PLSA and LDA suggests that LDA may go someway to explain why concepts are considered to the similar.

As yet the task of reconciling the top-down and bottom-up approaches to interoperability remain unexplored but the latent analysis approaches can be applied to more than one dataset to identify classes (i.e., documents) with shared concepts to facilitate data integration.

References

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. Journal of machine Learning Research 3: 993–1022

Comber A, Fisher P, Wadsworth R (2004a) Integrating land cover data with different ontologies: identifying change from inconsistency. International Journal of Geographical Information Science (IJGIS) 18(7): 691–708

Comber AJ, Fisher PF, Wadsworth RA (2004b) Assessment of a Semantic Statistical Approach to Detecting Land Cover Change Using Inconsistent Data Sets, Photogrammetric Engineering and Remote Sensing, 70(8), pp 931–938

Comber AJ, Fisher PF, Wadsworth RA (2005a) A comparison of statistical and expert approaches to data integration. Journal of Environmental Management 77, pp 47–55

Comber AJ, Fisher PF, Wadsworth RA (2005b) Combining expert relations of how land cover ontologies relate. International Journal of Applied Earth Observation and Geoinformation 7(3): 163–182

Comber AJ, Fisher PF, Harvey F, Gahegan, M, Wadsworth RA (2006) Using metadata to link uncertainty and data quality assessments. In: Riedl A, Kainz W, Elmes G (eds) Progress in Spatial Data Handling, Proceedings of SDH 2006, Springer, Berlin Heidelberg, New York, pp 279–292

Csillag F, Boots B (2004) Toward comparing maps as spatial processes. In: Fisher P (ed) Developments in Spatial Data Handling, Springer, Berlin Heidelberg New York, pp 641–652

Foody GM (2004) Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. Photogrammetric Engineering and Remote Sensing 70 (5): 627–633

Frank AU (2001) Tiers of ontology and consistency constraints in geographical information systems. International Journal of Geographical Information Science (IJGIS) 15(7): 667–678

Frank AU (2003) A linguistically justified proposal for a spatio-temporal ontology. In: COSIT'03**,** Conference on Spatial Information Theory 24-28 September, Ittingen, Switzerland

Frank AU (2007a) Towards a Mathematical Theory for Snapshot and Temporal Formal Ontologies. In: The European Information Society, Lecture Notes in Geoinformation and Cartography, Springer, Berlin Heidelberg New York, pp 317–334

Frank AU (2007b) Incompleteness, error, approximation, and uncertainty: An ontological approach to data quality. In: Morris A, Kokhan S (eds) Geographic Uncertainty in Environmental Security, Proceedings of NATO Advanced Research Workshop on Fuzziness and Uncertainty in GIS for Environmental Security and Protection, Kyiv, Ukraine, JUN 28-JUL 01, 2006, pp 107–131

Frank AU, Grum E, Vasseur R (2004) Procedure to select the best dataset for a task. In: Proceedings of Geographic Information Science, Lecture Notes in Computer Science Vol 3234, pp 81–93

Fritz S, See L (2005) Comparison of land cover maps using fuzzy agreement. International Journal of Geographical Information Science (IJGIS) 19(7): 787–807

Fuller RM, Groom GB, Jones AR (1994) The Land Cover Map of Great Britain: an automated classification of Landsat Thematic Mapper data. Photogrammetric Engineering and Remote Sensing 60: 553–562

Guo QH, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of sudden oak death in California. Ecological Modelling 182(1): 75–90

Hagen A (2003) Fuzzy set approach to assessing similarity of categorical maps. International Journal of Geographical Information Science (IJGIS) 17(3): 235–249

Harvey F, Kuhn W, Pundt H, Bishr Y, Riedemann C (1999) Semantic interoperability: A central issue for sharing geographic information. Annals of Regional Science 33(2): 213–232

Hofmann T (1999a) Probabilistic latent semantic indexing. In: Hearst M, Gey F, Tong R (eds) Proceedings of 22nd International Conference on Research and Development in Information Retrieval University of California, Berkeley, California, Aug, 1999, pp 50-57

Hofmann T (1999b) Probabilistic latent semantic analysis. In: Laskey KB, Prade H (eds) Proceedings of 15th Conference on Uncertainty in Artificial Intelligence, Royal Inst Technol, Stockholm, Sweden, Jul 30-Aug 01, 1999, pp 289–296

Honkela T (1997) Self-Organising maps in natural language processing. PhD thesis, Helsinki University of Technology, Department of Computer Science and Engineering, http://www.cis.hut.fi/~tho/thesis/

Jurafsky D, Martin JH (2008) Speech and Language Processing (Second edition), Prentice Hall Series in Artificial Intelligence

Kampichler C, Dzeroski S, 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

Lin X (1997) Map displays for information retrieval. Journal of the American Society for Information Science 48: 40–54

Maier HR, Dandy GC (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

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling 190(3-4): 231–259

Pundt H, Bishr Y (2002) Domain ontologies for data sharing-an example from environmental monitoring using field GIS. Computers and Geosciences 28(1): 95–102

Roubens M (1982) Fuzzy clustering algorithms and their cluster validity. Eur. J. Oper. Res. 10: 294–301

Soil Survey Staff (1999) Soil Taxonomy A Basic System of Soil Classification for Making and Interpreting Soil Surveys, 2nd Edition, Natural Resources Conservation Service Number 436. U.S. Government Printing Office Washington, DC 20402 (available at ftp://ftpfc.sc.egov.usda.gov/NSSC/Soil_Taxonomy/tax.pdf, accessed 7 October 2007)

Wadsworth RA, Comber AJ, Fisher PF (2006) Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. In: Riedl A, Kainz W, Elmes G (eds) Progress in Spatial Data Handling, Proceedings of SDH 2006, Springer, Berlin Heidelberg New York, pp 197–213

Wadsworth RA, Fisher PF, Comber A, George C, Gerard F, Baltzer H (2005) Use of Quantified Conceptual Overlaps to Reconcile Inconsistent Data Sets. In: Proceedings of GIS Planet 2005, Session 13 Conceptual and cognitive representation, Estoril Portugal 30th May - 2nd June 2005. ISBN 972-97367-5-8. 13pp