ABSTRACT

Title of Document: **SEMANTIC INTEGRATION OF GEOSPATIAL CONCEPTS – A STUDY OF LAND USE LAND COVER CLASSIFICATION SYSTEMS**

Hua Wei, Ph. D., 2011

Directed By: Professor John R. G. Townshend
Department of Geography
University of Maryland, College Park

In GI Science, one of the most important interoperability issues relates to land use and land cover (LULC) data, because it is a key to the evaluation of LULC's many environmental impacts throughout the globe (Foley et al. 2005). Accordingly, this research aims to address the interoperability of LULC information derived by different authorities using different classificatory approaches.

LULC data are described by LULC classification systems. The interoperability of LULC data hinges on the semantic integration of LULC classification systems. Existing works on semantically integrating LULC classification systems has a major drawback in finding comparable semantic representations from textual descriptions. To tackle this problem, we borrowed the method of comparing documents in information retrieval, and applied it to comparing LULC category names and descriptions. The results showed notable improvement compared to previous work.

However, lexical semantic methods are not able fully to solve the semantic heterogeneities in LULC classification systems: the confounding conflict – LULC categories under similar labels and descriptions have different LULC status in reality, resulting in a naming conflict – LULC categories under different labels can represent similar LULC type. Without confirmation of their actual land cover status from remote sensing, lexical semantic method cannot achieve reliable matching.

To discover confounding conflicts and reconcile naming conflicts, we developed an innovative method by applying remote sensing to the integration of LULC classification systems. Remote sensing is a means of observation of actual LULC status of individual parcels. We calculated parcel level statistics from spectral and textural data, and used these statistics to calculate category similarity. The matching results showed this approach fulfilled its goal of overcoming semantic heterogeneities and achieved more reliable and accurate matching between LULC classifications in the majority of cases.

To overcome the limitations of both methods, we combined the two by aggregating their output similarities, and achieved better integration. LULC categories that display noticeable differences between lexical semantics and remote sensing once again remind us of semantic heterogeneities in LULC classification systems that must be overcome before LULC data from different sources become fully interoperable.

SEMANTIC INTEGRATION OF GEOSPATIAL CONCEPTS – A STUDY OF LAND USE LAND COVER CLASSIFICATION SYSTEMS

By

Hua Wei

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Ph. D
2011

Advisory Committee:
Professor John R. G. Townshend, Chair
Professor Shunlin Liang
Dr. Chengquan Huang
Dr. Jianguo Ma
Professor Jimmy Lin

© Copyright by
Hua Wei
2011

Acknowledgements

I want to express deep thanks to my advisor, the committee, the department, and my family and friends. Without their generous help, I could not have accomplished this task.

First, I would like to express my gratitude to my adviser Dr. John Townshend. His keen sense and vast knowledge in scientific research made him aware of the problem of semantic heterogeneity in geography years ago. Inspired by our numerous helpful conversations, I eventually realized the use of remote sensing data can provide an innovative and effective approach to address my research questions, which becomes the backbone of this dissertation. It has been an honor for me to have Dr. Townshend as my adviser.

This research has received valuable guidance and support from the members of my advisory committee: Dr. Jimmy Lin, Dr. Chengquan Huang, Dr. Shunlin Liang and Dr. Jianguo Ma. I am grateful for their knowledge helps.

I thank Dr. Naijun Zhou for his selfless help when I started my PhD research. He is a great person and I learned a lot from him, not only academically but also on how to be a better person.

I always consider myself lucky to study in the Department of Geography at University of Maryland College Park. It is the best geography department in my opinion.

There can never be enough said of the support that I am so fortunate to have from my wife Chun Yi and my parents He Wei and Xiaohong Lu. Pursuing a PhD is not easy. Without their support, I could have never finished it.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Over the past decades, the technology of collecting geospatial data has developing fast (Goodchild 1997, Butenuth et al. 2007), and has led to rapid accumulation of geospatial data. For example, Landsat 5 and Landsat 7 contribute over 400 images per day to the Earth Resources Observation Systems (EROS) data archive, and Landsat 5 alone has gathered more than 700,000 images since its launch in 1984. Also many state administrations collect LU data often through a combination of ground datat collection and interpretation of aerial photographs The abundance of geospatial data collection leads to more distributed and heterogeneous sources.

Meanwhile, in the context of geographic information services (Kuhn 2005), users need to share geospatial data from multiple data sources (Elwood 2008). Interoperability of geospatial data is of decisive importance to answering many fundamental geographical research questions, such as the impact of human activities on global change. Any single data source is not adequate to capture this complexity, and the interoperability of geospatial data is strongly required.

Interoperability may be depicted by six levels of heterogeneity (Figure 1.1) (Sheth 1999). Among them, semantic heterogeneity is widely considered to be the main challenge in achieving interoperability (Rodriguez et al. 1999, Sboui et al. 2007). A consideration of ontology provides a possible solution to this problem.

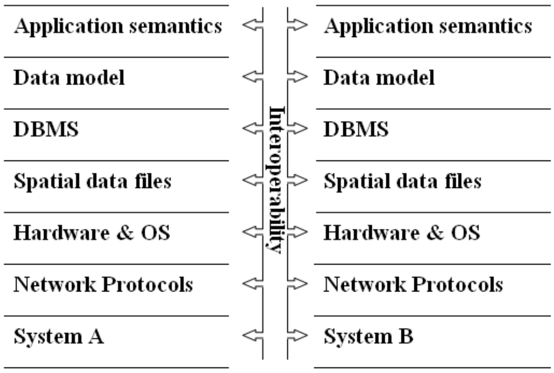


Figure 2.1 Levels of heterogeneity / interoperability

(Bishr 1998)

1.1 Ontological View on Semantic Integration

Semantics refers to the meanings of symbols (e.g. words) (Wood 1975). Jackendoff (1983) commented that semantics, "bridges the theory of language and the theories of other cognitive capacities". A similar definition but in simpler words, by Agarwal (2005) pointed out that semantics implies the meaning attached to concepts.

Semantic heterogeneity originated from the different conceptualization of the physical existence. Ontology, which is the theory of physical existence, should be introduced to solve heterogeneity problems. Ironically, ontology, proposed as a solution to semantic heterogeneities, has semantic heterogeneities in itself. We will first give a brief review of different understandings of ontology, and then clarify the use of ontology in this research.

1.1.1 Defining Ontology

The term ontology has different definitions and usages in different contexts (Agarwal 2005, Guarino 1995, 1998). As a branch of philosophy, ontology is the theory about the nature of being, including the categorization of being and their relations. There are many research works defining ontology. For example, ontologies attempt to clarify and set the explicit knowledge of the domain they describe (Kavouras 2005); ontology is an explicit specification of a shared conceptualization (Torres *et al.* 2009); ontology is a particular knowledge base that describes facts that are always true for a community of users (Guarino 1998); ontology can be a simple taxonomy, a lexicon, or a thesaurus, or even a fully axiomatized theory (Fonseca *et al.* 2002); ontology is the method to extract a catalogue of things or entities that exist in a domain (Sowa 2000).

Despite the wording variance, these definitions of ontology mainly differentiate on whether human conceptualization and physical existence are detached. From one viewpoint, ontology does not only recognize existence, but also specifies the conceptualization (Torres *et al.* 2009) shared by a group of people, such as geographic information community (Bishr *et al.* 1999). Here ontology incorporates epistemology, and this so-called epistemological ontology is defined as a theory about how a given individual, group, language, or science conceptualizes a given domain (Fonseca *et al.* 2002).

Engineering-oriented ontology is developed from the epistemological view of ontology. From engineering perspective, ontology is a strictly pragmatic enterprise (Smith and Mark 2001). In AI-related contexts, for example, ontology is a classification with a rich set of semantic relationships (among terms) that support reasoning (Soergel 2005).

From a philosophical perspective, Agarwal (2005) believes, if the ontology is recognized as THE concept of a being, then the use of plural 'ontologies' is irrelevant as there can be only one ontology. Hence, ontology alignment is irrelevant. However, knowledge engineers and artificial intelligence scientists (e.g. Euzenat and Shvaiko) think engineering-oriented ontology is not the representative of a singular overriding truth, but corresponds to individual and inevitably different conceptualization processes. The use of plural ontologies is relevant, and ontology alignment becomes necessary.

From the other point of view, ontology should solely represent physical existence that is independent from human conceptualization (Smith 2004). In scientific endeavor, where realism–materialism is adopted as the doctrine, physical existence and human conceptualization are detached, and this detachment enables ontology's independence from human conceptualization. By representing existence rather than concept, ontology has the power to solve semantic heterogeneity problem. This study as a semantic integration attempt uses this power through embracing a unique view of ontology (section 1.1.3).

1.1.2 Ontology Enabling Semantic Integration

Based on ontology, semantics are expressed as symbols, and difference in the expressing process leads to semantic heterogeneity; semantic integration aims to eliminate semantic heterogeneities. Semantic heterogeneity is formally defined as the variance of semantically related objects in different data sources (Kashyap and Sheth 1998). Different types of semantic heterogeneity include (Goh 1997): 1) *confounding conflicts* occur when information items seem to have the same meaning, but differ in reality (Figure 1.2 (a)); 2) *naming conflicts* occur when the naming schemes of the information differ largely (Figure 1.2 (b)); 3) *scaling and units conflicts* occur when different reference systems are used to measure a value. The scaling and units conflict is straightforward and less challenging. To overcome the first and second type of heterogeneity, a consideration of ontology is indispensible.

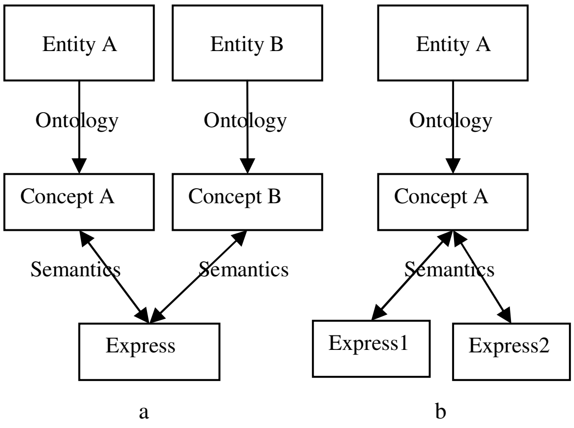


Figure 1.2 Semantic heterogeneity

In reality, two groups of physical existences A and B are conceptualized to concepts A and B, then semantically expressed as texts.

Philosophical ontology and engineering-oriented ontology solve semantic heterogeneity in theory and in practice respectively. In theory, we define ontology as a representation of existence, and it provides the philosophical foundation of semantic integration. Based on ontology (independent of conceptualization), semantics are expressed differently from person to person. This difference in expression leads to the semantic heterogeneity problem. Therefore, when semantic heterogeneity has to be overcome, there always is ontology serving as a base that heterogeneous semantic expressions can return to. To sum up, philosophical ontology makes semantic integration theoretically feasible.

In practice, ontology is used as an explicit specification of a conceptualization namely "a theory of a given domain which can be accepted and reused by all information gatherers in that domain" (Smith and Mark 2001). Based on this use of ontology, ontological engineering was developed. An ontology, in this sense, concerns itself not at all with the question of ontological realism, that is, the question whether its conceptualizations are true of some independently existing reality. Rather, it starts with conceptualizations, and goes from there to a description of the corresponding domain of objects (Smith and Mark 2001). With this function, ontologies (plural only meaningful in this sense) can be used in an integration task to describe the semantics of the information sources (Wache *et al.* 2001). Developing an information system always relies on ontology implicitly or explicitly; making it explicit avoids conflicts between the ontological concepts and implementation (Fonseca *et al.* 2002). Having (engineering-oriented) ontologies defined explicitly, semantic integration is accomplished based on ontological matching (or rebuilding) (Uitermark *et al.* 1999, Fonseca *et al.* 2002, Lin and Ludascher 2003, Kavouras 2005, Durbha *et al.* 2009).

1.1.3 Ontology Built on Universals

In this semantic integration effort, we adopt Barry Smith's definition of ontology as the representation of universals (Smith 2004). The unique advantage of this definition lies in its key – **universal**, and we will start explaining it by differentiating category and concept.

Category and concept are different. Concept is the information item that linked to cognitive semantics. Category, on the other hand, is a grouping of existences in the real world. A concept may have no instances in the real world (i.e. abstract concept), and a category may be a random selection of existences and links to no semantics. Smith and Mark (2001) pointed out that each 'valid' scientific concept must have instances; otherwise it is worthless in terms of scientific research as scientific research is meant to find rules. Based on this point of view, Smith (2004) introduced universal as the invariant in reality deduced from the commonality of instances, and defined ontology as the representation of universals. By his definition, universal must link to semantics as concept does. At the same time, universal must have instances as a category does, and each universal stands for the commonality of its instances.

The main advantage of considering concepts as universals is that we could incorporate instance level information in semantic integration. Semantic integration aims to determine the relations between concepts, and this is based on the measurement of their semantic similarities (Euzenat and Shvaiko 2007). Acknowledging concept as universal, we could measure the similarity between concepts not only by comparing their semantic expression (e.g. text, semantic taxonomy) (illustrated as approach 1 in Figure 1.3), but also by comparing their instances (approach 2 in Figure 1.3). The semantic integration methods using instance level information are termed *extensional* methods, in contrast to the *intensional* methods that concerned with concept-level (i.e. concept definitions) and/or schema-level (i.e. hierarchy in taxonomy) information.

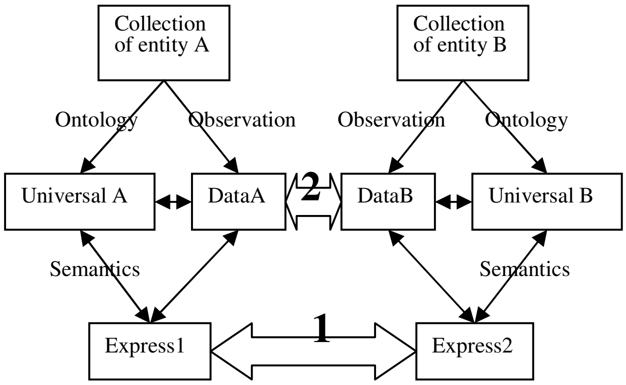


Figure 1.3 Semantic integration approaches: intensional approach (1) and extensional approach (2)

1.2 Semantic Integration of LULC Information

One of the most important issues of interoperability is needed in LULC data. LULC has environmental impacts on many different aspects throughout the globe; examples include the global carbon cycle, global climate, atmospheric composition, regional climates (through changing surface energy and water balance), the hydrologic cycle, anthropogenic nutrient inputs to the biosphere from fertilizers and atmospheric pollutants, water quality, coastal and freshwater ecosystems, biodiversity (through the loss, modification, and fragmentation of habitats), degradation of soil and water, overexploitation of native species, and local food supply (Foley et al. 2005). Evaluating the impacts of LULC requires the integration of multiple LULC data sources, as a single source cannot provide adequate accuracy or geographic coverage.

This research aims to address a fundamental geographical question: how it is possible to integrate and compare LULC information derived by different authorities using very different classificatory approaches. We start by reviewing current semantic integration methods in Geographic Information Science (GI Science).

1.2.1 Semantic Integration Methods in GI Science

Semantic integration is achieved through ontology alignment, which requires the measurement of similarities between concepts. Similarity theory was originally developed for psychological models to explain human-similarity judgment. Schwering (2008) summarized five main categories of semantic similarity measurements and illustrated the potential application in GI Science of each category. Based on different notions on the knowledge representation and similarity calculation, these categories are geometric, feature-based, network, alignment and transformational methods. A similar categorization can also be found in Goldstone and Son's work (2005). In addition to these methods, recent research incorporates the information theory in the similarity measurement.

Geometric model

The geometric model uses a distance in a multidimensional space to represent similarity (Rips et al. 1973, Gardenfors 2000). Each dimension of the space corresponds to a quantitative property of concept, with the property value being proportional to the dimensional coordinate. A concept is then projected to a vector. Based on this multidimensional vector space representation of semantics, the similarity is measured as the linear decay function of the distance between vectors.

When applying the geometric model, it is necessary to quantify properties of each concept, and an underlying assumption is all properties (dimensions in the vector space) are independent.

Feature-based model

The feature-based model is widely used (Kavouras et al. 2005, Rodriguez and Egenhofer 2004) since invented in the 1970s (Tversky 1977). Semantic features are defined as a concept's distinguishing classifiers, such as attributes, functions, and parts (Miller 1990). The "universe of discourse" (Feng and Flewelling 2004), which maintains a complete collection of features that can represent all concepts in current context, is built. Then a common model, a contrast model, or a ratio model (Tversky 1977) can be employed to compute similarity from shared and different features.

The feature-based model cannot handle disjoint similarity i.e. where there are no shared features. Also it is impossible to find a complete set of characteristic attributes describing a real world object. The major problem is to choose representative attributes (Kuhn 1995).

Network model

The network model is always an intuitive choice when the concepts to be compared form a taxonomy-like structure, in which nodes represent concepts, and edges represent semantic relations between concepts (Sunna and Cruz 2007). Semantic similarity is measured by the shortest path between nodes (concepts) (Collins and Quillian 1969). Traditional semantic relations attach much importance to hyponymy and hypernymy, but do not consider mereological (part-of) relations, which describe the relation between parts and whole. Mereological (or partial) relation is introduced to the semantic network in Guarino's work (1995).

The network model is determined by the predefined semantic network architecture. Resnik (1995) pointed out that the network model and the edge-counting method rely on the notion that "links in the taxonomy represent uniform distances", which is usually not true: "there is a wide variability in the 'distance' covered by a single taxonomic link, particularly when certain sub-taxonomies are much denser than others (Resnik 1995)".

Alignment model

Developed from feature-based model, the alignment model not only measures the feature-based similarity, but also considers the alignment of features and includes it into similarity measurement (Goldstone 1994). The alignment model is applied to spatial scene comparison, where objects are considered to be features and the spatial relation between objects contributes matching.

Information-theoretic model

Many recent approaches incorporate information in measuring semantic similarity. Lin (1998) proposed an information-theoretic definition of similarity, derived from a set of assumptions on similarity in the way the entropy/information is defined. Based on the notion that shared information corresponds to similarity, Lin formally defines the similarity (*sim*) between *A* and *B* as

$$sim(A,B) = \frac{2\log P(common(A,B))}{\log P(description(A,B))},$$

in which *common(A,B)* is the amount of information needed to describe the commonality of *A* and *B*, *description(A,B)* is the information needed to fully describe both *A* and *B*. Based on Lin's similarity in taxonomy, Resnik (1999) proposed a measure of semantic similarity using WordNet (Fellbaum 1998). Formally, Resnik defines the similarity (*sim*) between concept c_1 and c_2 as

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)] ,$$

where $S(c_1,c_2)$ is the set of concepts that subsume both c_1 and c_2, or in WordNet, set of hyponyms of both c_1 and c_2, and $p(c)$ is the probability of c's occurrence.

Extensional methods

The discussion above summarized widely used semantic integration methods adopting an intensional approach, as all their knowledge representations, including quantitative properties, position in taxonomy, features, and occurrence pattern all relates to a concept rather than instances.

There are also a few works using information specific to instances. For example, Ehrig et al. (2005) developed a comprehensive framework for measuring similarity within a single ontology. Their framework is defined in three layers on which the similarity between concepts can be measured using instance level, conceptual, and contextual information respectively. Then the similarity is aggregated as a weighed average of the individual similarities from each layer. Based on Ehrig et al.'s work, Albertoni and De Martino (2008) proposed an asymmetric similarity assessment among instances belonging to the same ontology.

1.2.2 Two Approaches to Semantically Integrating LULC Classification Systems

LULC data are described by LULC classification systems. The interoperability of LULC data depends on the semantic integration of LULC classification systems. A typical LULC classification system organizes LULC categories in a taxonomic structure, in which each category is defined by a name (label) and often followed by a textual description.

From the above review on semantic integration methods, we find not all integration methods are applicable to LULC classification systems, because external references or repositories commonly used in generic semantic methods are not available in the study of LULC. For example, a corpus for training purposes is essential to the information-theoretic model. But in the study of LULC, a general purpose corpus is not applicable, because LULC category descriptions, written and read by land use experts on the purpose of a scientific specification, are different from a general-purpose corpus in terms of vocabulary, word sense and frequency, and the length of paragraphs.

As for the network model, the soundness of semantic hierarchy (taxonomy) determines the performance of similarity measurement. While upper ontology (an ontology describing shared concepts across knowledge domains), such as Cyc (http://www.cyc.com/), DOLCE (http://www.loa-cnr.it/DOLCE.html), and WordNet, is too general to compare concepts from a narrow domain such as LULC, no domain ontology exists in LULC because "a lot of effort is required to construct ontology from scratch not just from a technical point of view but more importantly the process of knowledge extraction from domain experts and arriving at a consensus view" (Bhogal et al. 2007). The design and construction of domain ontology is labor intensive, time consuming and difficult (Kashyap, 2001).

Ruling out information-theoretic model and network model, many research works in semantic integration of LULC information are based on the use of features. In chapter 2, we will further discuss the applicability and drawbacks of these methods, and then propose a new information retrieval approach to semantically integrate LULC classification systems, which can overcome current methods' limitation in applicability and improve the performance.

Adopting Smith's definition of ontology (Smith 2004) (section 1.1.3), we find that LULC categories are universals. A LULC category, not only as a real world concept implies semantics, but also is populated by individual parcels of its kind, which are directly monitored by modern earth observing technologies, such as remote sensing. Hence, as explained in section 1.1.3 and illustrated in Figure 1.3, in addition to an intensional method that uses semantic expressions, an extensional method that uses the remote sensing information attached to parcels should be available.

However, existing methods of semantic integration of LULC information regrettably have not taken the important advantage of remote sensing. In chapter 3, we will present a remote sensing based approach to the semantic integration of LULC classification systems. We will use spectral and textural information derived from time series remotely sensed data to calculate the similarity between LULC parcels, and adopt an extensional approach to measure similarity between LULC categories. Then in chapter 4, we will test the generality of the remote sensing based method on more LULC information. In chapter 5, we will try to optimize the matching algorithm by refining the input and reduce procedure errors.

In chapter 6, we will present the improved matching result through integrating both intensional and extensional approaches. More importantly, we will prove, by combining the two approaches, we can have the ability to discover confounding conflicts and reconcile naming conflicts. Conclusions and future directions will be given in chapter 7.

Chapter 2: Employing Information Retrieval Methods to Improve Semantic Integration of LULC Data

2.1 Introduction

The interoperability of land use and land cover (LULC) data derived by different state authorities using different classificatory approaches is crucial to accurately capture regional land use dynamics, which has an impact on environment and socio-economics (Foley et al. 2005). Data interoperability requires different levels of integration (Sheth 1999). Among them, semantic integration is widely considered challenging (Rodriguez et al. 1999, Sboui et al. 2007), and will be the focus of this dissertation.

LULC data are described by LULC classification systems. A typical LULC classification system organizes LULC categories in a taxonomy structure, in which each category is defined by a name (label) and a textual description. The semantics of LULC data are expressed in textual definitions (names and descriptions). Therefore, semantically integrating different LULC data requires us to compare the definitions of LULC categories in different classification systems. The originality of the approach in this chapter is to apply methods in modern information retrieval to achieve this comparison. Thanks to this application, not only better integration results will be achieved, but also, unlike existing methods, our method does not rely on comparable characteristics, which are difficult to extract from textual descriptions.

But before we jump into the discussion on methodology, it is necessary to take a brief review on current semantic integration methods in Geographic Information Science (GI Science) and their applicability in LULC classification systems.

2.1.1 Semantic Integration in GI Science

Semantic integration aims to determine the relations between concepts, and this is based on the measurement of their semantic similarities (Euzenat and Shvaiko 2007). Similarity theory was originally developed for psychological models to explain human similarity judgment. Schwering (2008) summarized five categories of semantic similarity measurement and exemplify the potential application in GI Science of each category of measurement. Based on different notions on the knowledge representation and similarity, the categories are geometric, feature, network, alignment and transformational methods. In addition to these methods, recent research incorporates information in the similarity measurement. Table 2.1 listed the representative works for each type of methods, and a detailed specification of each model's knowledge representation and similarity calculation can be found in section 1.2.1.

Table 2.1 Methods to compare geospatial universals

Model	**Representative work**
Geometric model	Rips et al. 1973, Gardenfors 2000, Schwering & Raubal 2005-2
Feature-based model	Kavouras et al. 2005, Rodriguez & Egenhofer 2004, Ahlqvist 2005, 2008
Network model	Sunna & Cruz 2007
Alignment model	Goldstone 1994
Information-theoretic model	Lin 1998, Resnik 1999

2.1.2 Distinctiveness of LULC Category Descriptions

In section 1.2.2, we explained that not all generic semantic methods are applicable to the integration of the LULC classification systems, because external references or repositories commonly used in generic semantic methods (such as a corpus or domain ontology) are not available in this *ad hoc* semantic integration. Hence, many research works in semantic integration of LULC classification systems are based on the use of features.

FAO (Mücher et al., 1993) proposed a parametric land cover classification system, in which categories are defined by a combination of a set of independent diagnostic criteria or classifiers. Gregorio and Jansen (1998) claimed that any land cover identified anywhere in the world can be readily accommodated in this parametric classification. However, this claim is over-optimistic, since translating existing systems will be difficult itself, not to mention the errors introduced in the translation.

Different from FAO's "top-down" approach, many attempts of integrating existing LULC classification systems adopt a "bottom-up" approach. Based on Miller et al.'s work (1990) that categorized features into attributes, functions, and parts, Kavouras et al. (2005) extracted features from the textual description of a LULC category. Feng and Flewelling (2004) extended this model by assigning weight, which is calculated from classification taxonomy, to each feature. Rodriguez and colleagues (Rodriguez et al. 1999, Rodriguez & Egenhofer 2003, 2004) extended the model by introducing asymmetry in the similarity calculation. Adopting features defined by FAO (Mücher et al., 1993), Ahlqvist (2008) developed a fuzzy sets based approach to quantify semantic change between two categories in NLCD 1992 and 2001 systems.

Approaching similarity measurement differently, Sunna and Cruz (2007) focused on local structure (parents and siblings of a concept) of the taxonomy structure, and used it as the contextual information in semantic similarity measurement. However, this method is limited because it relied on initial similarity values, which must be calculated using other similarity measurements beforehand.

Kavouras and Kokla (2002) proposed a concept lattice based approach to formalize the comparison of land use categories using their subcategories and attributes. However, the requirement of a clear, non-overlapped and unambiguous identification of the attributes is difficult to meet.

In a nutshell, these semantic integration methods' using textual description has a major drawback: it is rather difficult to define comparable semantic representations, either the "features" in feature-based model, the "attributes" in concept lattice, or the "dimensions" in geometric model. Natural language processing (NLP) techniques are employed to automate this process, but satisfactory results are only obtained in narrowly restricted domains (Cunningham et al. 2002, Peng & McCallum 2006). Kuhn (1995) has pointed out the problem of choosing representative characteristics to describe a real world object. But in addition to Kuhn's concern, when integrating LULC classification systems, pre-selected representative characteristics may not even be explicitly defined in category descriptions, not to mention being extracted and used in comparison.

To tackle this problem, we borrowed the method of comparing documents in information retrieval, and applied it to the semantic integration of LULC classification system. Information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfied an information need from within large collections (usually stored on computers) (Manning et al. 2008). Since the emergence of World Wide Web, information retrieval has been developing fast in recent years, and always been providing the theoretical foundation to modern web search engines. Bag-of-words representation and cosine similarity is the basics of information retrieval and the heart of search engine. The success of today's search engines justified their capability of comparing documents, which suggests the rationale of applying them to the comparison of textual descriptions of LULC categories. In the rest of this chapter, section 2.2 will specify what the information retrieval method is and how it can be applied to our integration problem, and several variations of the algorithm are introduced along the way. Section 2.3 will present the result and summarize advantages and drawbacks of our approach.

2.2 Methodology

Adopting an information retrieval approach, category descriptions are first represented in bag-of-words model (2.2.1). Then a vector space model (VSM) is built upon term frequency and inverse document frequency in the collection of LULC category descriptions. A cosine similarity between category descriptions is then calculated (2.2.2). In section 2.2.3 and 2.2.4, two possible optimizations on emphasizing keywords (2.2.3) and incorporating semantic relatedness (2.2.4) are discussed.

2.2.1 Bag-of-Words Model

The bag-of-words model (Harris 1954) is a simplifying model representing documents. In this model, a text (i.e. a paragraph of LULC description) is represented as an unordered collection of words appeared in the text, disregarding grammar, structure, and the order of words (Manning et al. 2008). This simple representation is successful in information retrieval, because it is effective in indexing, invert indexing, and document comparison. However, the bag-of-words representation has no consideration of the structure, syntax and semantics.

Noise Word Filter

In the bag-of-words model, noise words, or “stop words”, are words that are common, short, and functioning words, such as the, is, at, which and on. These words are of little semantic importance, but may bias the similarity measurement. Noise words need to be filtered out. There are two ways to build the noise word list. Either it can be defined by experts (human input), or automatically generated by analyzing the frequency of words in training corpus. In this research, there is no applicable corpus for training purposes, and the noise word list is pre-defined manually.

Morphological Analysis

Morphological variance can also confuse the bag-of-words model. The popular morphological analysis in search engine is stemming, among which Porter stemmer (Porter 1980) is popular. Porter stemmer is a rule based stemmer, which apply a group of rules to an input word and transform it to a stem.

Despite its wide application in information retrieval, stemmer is not applicable in this research because of two reasons. First, stems may not be words. For example, the word ‘capability’ will be stemmed to ‘capabl’, which is not a word. In information retrieval, this is not a problem, because the system is based on string rather than semantics. As long as the stemmer conflates the morphological variances (e.g. capability and capabilities) to the same string (capabl), the information retrieval system can use this string to do the indexing, invert indexing, and execute queries. However, stems are not enough in this research because one of our optimization will consider the semantics by querying words in WordNet, in which entries need to be well-spelled words.

Second, stemmer has the error of omission and commission. The error of omission happens when the simple rule-based stemming conflates morphological variances of one word into different stems. For example, the word 'explain' and 'explanation' will be stemmed to 'explain' and 'explan' respectively. The error of commission happens when the stemming conflates morphological variances of different words into one same stem. For example, both 'university' and 'universe' will be stemmed to 'univers'. Omission and commission will bring irreversible error to the system.

The other type of morphological analyzer is lexicon-based, in which each word is queried in a huge predefined morphological variance list, and then its morphological root is returned. A lexicon-based morphological analyzer is accurate, but cumbersome and inefficient.

To combine the advantages of both approaches, this research adopts a hybrid morphological analyzer called "Morphy", which is used in the WordNet (Fellbaum 1998). WordNet is a widely used, broad-coverage semantic network for English, developed at the Cognitive Science Laboratory of the Princeton University. "Morphy" consists of an exception list and detachment rules. When an inflectional word comes, it will first be searched in the exception list, and if found, its base form will be returned. If there are no matches in the exception list, a set of detachment rules are employed to detect the inflectional ending and substitute it with the base form. The output from applying each rule is checked in WordNet, and will be returned as base form if its spelling is right. Some variances, such as ‘axes’, can have more than one base form (i.e. ‘axe’ and ‘axis’), the first base form is returned. Thus the rules are ordered by popularity.

In section 2.3, as an effort of optimization, we employed syntactic parsing, negation detection, and extraction strategy to locate noun phrases, and ultimately, the keywords. In this effort, the category description is no longer treated as a plain bag-of-words; instead, the structure and syntax are considered. Only after the keyword extraction, the bag-of-words representation is again adopted to represent the category description.

2.2.2 Weighting Words and Cosine Similarity

Having each category description represented as a bag of words, the frequency of each word is then calculated. Clearly, all words in the description are not equally important in comparing two categories: the word that is mentioned more often in a description is more important to describing that category, thus, if shared by two categories, a word with high occurrence brings more weight to the similarity measurement.

Towards this end, a weight, which reflects the number of occurrences of a word in a description, is assigned to each word in each description. Then a score between a word *t* and a description *d*, based on the weight of *t* in *d* can be simply defined as the number of occurrences of word *t* in document *d*. This weighting scheme is referred to as “term frequency” and is denoted $\mathrm{tf}_{t,d}$, with the subscripts denoting the word and the description in order (Manning et al. 2008).

In addition to term frequency, another consideration is that the discriminating power of each word varies depending on its occurrence in the whole collection of category descriptions: the more often a word is seen in the collection, the less discriminating power it has. A mechanism is needed to follow this trend. An immediate thought is to scale down the weights of words with high collection frequency by the total number of occurrences of a word in the collection. However, this collection-wide statistic is proven by many experiments to be less effective than document-level statistic, i.e. the number of descriptions containing a specific word. Towards this end, the inverse document frequency (idf) of a term *t* is defined as follows:

$$idf_t = \log \frac{N}{df_t},$$

in which N denotes the total number of descriptions in a collection, and df_t denotes description frequency, defined to be the number of descriptions in the collection that contains the word t. By its definition, the idf of a rare word is high, whereas the idf of a frequent word is low.

Combining the term frequency and inverse document frequency, the composite weighting scheme tf-idf assigns to word t in description d a weight given by

$$tfidf_{t,d} = tf_{t,d} \times idf_t.$$

Then each description builds a document vector from its bag of words, with one component corresponding to a word in the collection, together with the tf-idf weight for each component. If a word does not occur in the description, the weight for the corresponding component is 0. Based on document vector, the similarity between description d_k and description d_l is defined as the cosine of the angle between the two corresponding document vectors $\vec{d}_k$ and $\vec{d}_l$, that is,

$$sim(d_k, d_l) = \cos(\vec{d}_k, \vec{d}_l) = \frac{\vec{d}_k \cdot \vec{d}_l}{\sqrt{\vec{d}_k \cdot \vec{d}_k}\sqrt{\vec{d}_l \cdot \vec{d}_l}} .$$

In the equation, the dot product on document vector is defined as

$$\vec{d}_k \cdot \vec{d}_l = \frac{\sum_{i=1}^{n} w_{i,k} w_{i,l}}{\sqrt{\sum_{i=1}^{n} w_{i,k}^2}\sqrt{\sum_{i=1}^{n} w_{i,l}^2}} ,$$

in which $w_{i,k}$ is the tf-idf weight for word i in description k.

2.2.3 Optimization – Emphasizing Keywords

The classic bag-of-words model excludes stop words in the representation based on statistics or human input rather than semantics, and the weighting scheme is based on occurrences but semantics. In this section, we try to develop a word weighting scheme considering semantics.

Consider the following LULC category description: “Urban areas whose use does not require structures, or urban areas where non-conforming uses characterized by open land have become isolated.” Intuitively, despite the high idf value, the word “require” should not be considered as important as word “open” or “urban” in terms of representing the category. To select words carrying more representation power, namely keywords, we set a bunch of testing rules based on desired characteristics of a keyword.

Keywords must be positive

In the context of describing LULC categories, negated and exclusive words do not conform with the intention of the concept. For example, in the description of Beaches: “extensive shoreline areas of sand and gravel accumulation, with no vegetative cover or other land use”, the negated part “vegetative cover or other land use” falls out of the scope of the concept, and therefore should not be considered when extracting keywords. An example of exclusion can be found in the category description: “Included are golf courses, parks, recreation areas (except areas associated with schools or other institutions), cemeteries, and entrapped agricultural and undeveloped land within urban areas.” In this sentence, the phrase “areas associated with schools or other institutions” are excluded from the description of the category, and hence would be excluded from keywords selection.

Keywords are in noun phrases

In English, a noun phrase (NP) consists of the center noun(s) and the modifier(s), both of which are indispensable in expressing the complete meaning. For example, a NP ‘single family’ has the center word ‘family’ and its modifier ‘single’. It takes both ‘family’ and ‘single’ to complete the semantics.

In the context of describing LULC categories, the use of verbs and adverbs is mostly just for syntactic purpose. Other parts of speech, such as the pronouns and conjunctions, also contribute little to category semantics. The NPs carry most of the semantics of a category, and therefore should be used to extract keywords. For example, category “Phragmites Dominate Urban Area” has the description: “This category contains urban areas where the common reed, Phragmites australis dominates.” and the NPs in the description are “this category”, “urban areas”, and “the common reed, phragmites australis”, carrying the complete semantics. Words outside NPs, i.e. “contain” and “where”, are of little semantic importance.

NO “Noise words” in keywords

In the context of describing LULC categories, words like land, use, area, etc., although not in a general purpose stop word list, are used so pervasively that they drown out real informative words. Moreover, even if words like “land” are not found in a paragraph of description, still no extra information is gained. These words’ occurrence in this context is pervasive, predictable, and therefore of little semantic importance. Hence, these words should also be excluded from keyword selection.

Figure 2.1 shows the workflow and main steps of keyword extraction. The whole process, divided into 6 consequential modules, will be discussed in order.

Preprocessing

Preprocessing is the preliminary processing of textual description to prepare it as formal as possible in order to be accurately parsed by a syntactic parser. Steps in preprocessing include: formatting structural heterogeneity, consolidating the word and symbol use, and breaking the paragraph into sentences to facilitate parsing. To avoid introducing errors, it is important to confine preprocessing on format level. The syntax and semantics of the texts do not change in preprocessing.

In consolidating word and symbol use, compound words, semantically indivisibles, are broken into single words. For example, “military_installation” will be broken into "military" and "installation". But in terms of semantics, the whole (e.g. military installation) is more than the sum of its parts (military and installation). Hence, compounds must be restored in later stages.

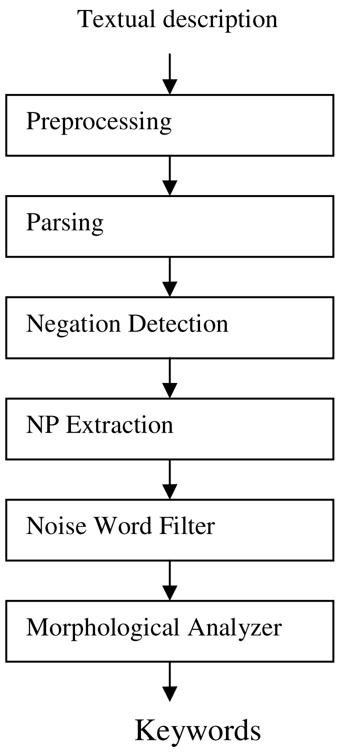


Figure 2.1 Flowchart of Keyword Extraction

Parsing

In natural language processing (NLP), parsing, or more formally, syntactic analysis, is the process of analyzing a sequence of words to determine their grammatical structure with respect to a given formal grammar. Given a sequence of words, a parser assigns each word a part of speech (POS) tag (i.e. noun, verb, adjective, pronoun, etc.), forms units like subject, verb, object, and determines the relations between these units according to some grammar formalism. In this research, the Berkeley Parser (Klein and Manning 2001) is adopted, because it is one of the most accurate and fastest parsers for a variety of languages, including English.

Negation Detection

One characteristic of keywords is the positivity. In the context of describing LULC categories, negated and exclusive phrases do not conform with the intention of the concept. Negation and exclusion indicator words (i.e. not, no, except, excluding, excluded, etc.) are located. Then the scope of negation or exclusion is decided, and content within the scope will be discarded from keyword extraction. Algorithm such as NegEx2 (Chapman et al. 2001) decides the negated scope without parsing, but it is less accurate. As we already have parsed sentences, finding the scope is straightforward and accurate – it is the immediate syntactic component containing a negation/exclusion indicator.

Noun Phrase Extraction

The noun phrases are marked during parsing, and only noun phrases falling out of the scope of negation and exclusion will be extracted.

For a nesting NP (a NP that contains one or more other NP(s)), we need to decide which NP within should be used to extract keywords. For example, NP "schools or other institutions" is a nesting NP, which contains NP "schools" and NP "other institutions". Selecting NPs to extract keywords is on the tradeoff between two processing gains. On one hand, the list of keywords should be kept short, otherwise using keyword is not different from the traditional bag-of-words model. To this end, we should use the NPs on the deepest nesting level to extract the keywords. On the other hand, the list of keywords should be complete. During the process of narrowing down to NPs on the deepest level, semantics loss due to the elimination of potential keywords is hard to avoid. Finding a one-size-fits-all strategy to extract minimum NPs without losing semantics is hard. Considering the semantics loss is irreversible, our approach is to keep all NPs (in the nesting), and to remove redundant keywords in a later processing.

In a parsed sentence, the scopes of NPs and the negated or exclusive parts will never overlap, because the English grammar, which is context-free, allows nesting but not overlapping. As the scope of a negation or exclusion word is within its direct component, NPs fall within this scope will, as a whole, be removed. A NP containing the negated or excluded scope will only have its negated or excluded part taken out, while the rest remains.

Compound Restoration

A NP consists of a central word and one or several modifiers. The structure of NP is analyzed to restore compound words. There are two types of structures, the flat structure and the nesting structure. A NP in flat structure has no NPs or other components nested, and the central word is the rightmost noun. The potential compound is generated by adding different combinations of modifiers in front of the central noun. Then the potential compound is searched in WordNet, and if found, the compound is restored. The searching starts from the potential compound with all modifiers, and if not found, it is shortened by discarding the farthest modifier and searched again until found. A nesting NP contains other NP(s) at some position, and can be treated recursively.

At this point, keywords are extracted from textual descriptions of LULC categories. A way to emphasize keywords in comparing descriptions is to assign heavier weights to keywords. In this research, we multiply the weights of keywords by 1.5 to emphasize their importance. In section 2.5, this keyword-enhanced model is compared with the model without keyword enhancement and the model using only keywords.

2.2.4 Optimization – Incorporating Semantics in Similarity Calculation

In the classic dot product of document vectors (section 2.2.2), an underlying assumption is that the components in the vector space are pair wise orthogonal. This means all words, corresponding to components, are independent, and the semantic relatedness (synonym, hyponym, or meronym) between any two words is negligible.

However, this assumption is over-simplifying the reality, as there are word mismatches between documents. For example, we may want to compare two LULC categories "retail" and "commercial", and assumedly both of the two contain only one word, i.e. "retail" and "commercial" respectively. Using classic dot product equation, the similarity between these two categories is 0, because they do not share words based on string match. However, term "retail" and "commercial" are not independent but related semantically. Therefore, incorporating semantic relatedness between the terms in the vector space model and/or similarity calculation becomes necessary (Chu et al. 2002).

Unsuitability of Expansion

In information retrieval, expansion is the process of expanding the terms in a document to match additional documents, and it is widely used to deal with word mismatch. In the previous example, if we expand the words retail and commercial by adding in their related terms e.g. shopping, the similarity between the two will be correctly increased.

Expansion can be achieved by human or by machine. Basic expansion involves techniques such as: adding in synonyms, hypernyms, and meronyms (Buscaldi 2005). More delicate expansion techniques have two approaches: global knowledge and local feedback. Global knowledge approach analyzes the corpus to discover word relationships, while local feedback approach analyzes documents retrieved by the initial query (Xu 1996). Bhogal et al. (2007) then separated global knowledge into two categories. One is corpus dependant knowledge (e.g. language model, like word-word co-occurrence, trained from corpus); the other is corpus independent knowledge (e.g. WordNet).

Local feedback and corpus dependant knowledge are not feasible in matching the LULC classification systems. But it seems corpus independent knowledge might be beneficial, because the category descriptions are relative short and may miss terms that will be string matched to related descriptions. However, a major hurdle in expansion is word sense disambiguation (WSD), because expansion should only performed on the intended meaning of keywords, otherwise expansion would decrease the precision.

WSD is still an open question in the NLP community; several efforts are contributing and effective in specific cases (Resnik 1999, Navigi and Velardi 2003). And due to the nature of the problem, all WSD methods employ a training process, meaning a corpus and/or a thesaurus are needed to provide knowledge about word senses based on usage and/or language model (context). Different methods may have different requirements on the corpus. While a substantial amount of untagged corpus would be adequate for some methods (Yarowsky 1995), some methods may need tagged and disambiguated corpus to achieve high accuracy (Mihalcea and Moldovan 2001). As discussed before, domain specific corpus is not available. But even it is available, Voorhees's experiments (1993) indicate that short statements (such as category descriptions) can be difficult to disambiguate because the "is-a" hierarchy is not sufficient to reliably select the correct sense of the noun.

With WSD inapplicable due to lack of corpus or low accuracy, expansion method is not a reliable method to handle word mismatch in this research. Instead, it is solved by incorporating semantic relation in similarity calculation.

There are two approaches to quantify the semantic relatedness between two words. The first one is based on semantic taxonomy, such as WordNet (Zhou and Wei 2008), and the second one is based on the co-occurrence of words, termed Latent Semantic Analysis (LSA) (Deerwester et al. 1990). In this research, we compare two approaches and try to choose the one with better performance on LULC category descriptions.

Quantifying Semantics Using WordNet

As introduced in section 2.1.1 network model, given a complete semantic network, the semantic similarity between two concepts depends on the shortest distance between their corresponding nodes in the network (Rodriguez et al. 1999). The shorter the distance, the more similar the two concepts are.

In WordNet, there are 19 kinds of relations for the nouns and 9 for the adjectives, including semantic and lexical relations. In a graph representation of WordNet (Fig 2.2), in which concepts (called “synsets” in WordNet) are the nodes and semantic relations are the edges. Quantitatively, the semantic relatedness is defined to be inverse proportional to the number of “hops” along the shortest path between the two concepts in WordNet, that is,

$$s(c_i, c_j) = \frac{1}{d(c_i, c_j)}.$$

In the equation, $d(c_i\ ,\ c_j)$ is the number of hops between concept c_i and c_j in WordNet.

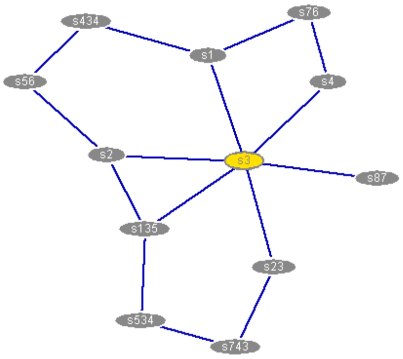


Figure 2.2 Subgraph of WordNet

The breadth-first search algorithm (BFS) is used to find the shortest path between two nodes. It is worth noticing that antonyms have opposite contribution to semantic similarity and therefore will terminate a search for semantic relatedness immediately.

Now we can calculate the semantic relatedness between concepts, but what we need in comparing two LULC descriptions is the semantic relatedness between words. For monosemous words, these two are equivalent as there is only one concept behind each word. For polysemous words, each of which represents multiple concepts, the semantic relatedness between words is set to be the closest semantic relatedness among all pair wise conceptual relationships (Resnik 1999). For example, consider how the similarity between word "field" and word "agriculture" would be decided. In all concepts behind "field", "a piece of land" has the closet relationship with "agriculture", and using this concept to calculate semantic relatedness is correct in this context. The equation computing similarity between term t_i and term t_j is then as follows,

$$s(t_i, t_j) = \max_{c_i \in C(t_i), c_j \in C(t_j)} [\frac{1}{d(c_i, c_j)}],$$

in which $C(t_i)$ is the set of all concepts correspond to term t_i.

Adopting the term similarity $s(t_i , t_j)$, the dot product is extended as follows,

$$d_k \circ d_l = \frac{\sum_{i,j} s(w_{i,k}, w_{j,l}) w_{i,k} w_{j,l}}{\sqrt{\sum_{i,j} s(w_{i,k}, w_{j,k}) w_{i,k} w_{j,k}} \sqrt{\sum_{i,j} s(w_{i,l}, w_{j,l}) w_{i,l} w_{j,l}}},$$

in which

$$s(w_{i,k}, w_{j,l}) = s(t_i, t_j) = \max_{c_i \in C(t_i), c_j \in C(t_j)} [\frac{1}{d(c_i, c_j)}].$$

Then if the term similarity is reduced to the Kronecker delta function,

$$s(w_{i,k}, w_{j,l}) = \delta(i, j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases},$$

the dot product is reduced to classic equation as in 2.2.2.

Quantifying Semantics Using Latent Semantic Analysis

Different from the WordNet approach, latent semantic analysis (LSA) is based on an assumption that words that are closely related in semantics tend to occur together in text, and therefore have similar occurrence distribute through the whole collection of documents. Now let us consider a term-document matrix, whose rows correspond to terms (words), columns correspond to documents (LULC descriptions), and each value corresponds to the occurrence of a term in a document. A technique in linear algebra called singular value decomposition (SVD) is employed to decompose the matrix and reduce the number of rows (words). SVD is not a total stranger to Geography community as it is the counterpart of Eigen decomposition for a non-square matrix; both decompositions are the foundation of principle component analysis (PCA), which has many applications in remotely sensed image processing.

The output of LSA is a low rank approximation of the original term-document matrix. Each description is represented by a low rank vector, based on which the cosine similarity between descriptions is calculated. Comparing to the original, the low rank approximation is much smaller and less noisy. More importantly, the semantic relatedness of words is considered, as the rank lowering is expected to merge the dimensions associated with words that have similar occurrence distribution. Returning to the example mentioned at the beginning of this section, a category named “commercial” and a category named “retail” will be related, if the word “commercial” and “retail” have similar occurrence patterns in the collection of LULC descriptions. An implementation of LSA in Python called “Gensim” is used in our method (Rehurek and Sojka 2010).

2.2.5 Summary on Methodology

In this section, we presented our innovative method to compare LULC descriptions, in which the bag-of-words model and cosine similarity are the basics, enhanced by two optimizations aiming the distinctiveness of LULC descriptions. In the first optimization, we developed the keyword enhancement strategy because a LULC description is usually much shorter than a regular document, and keywords have a more important role in deciding its concept’s intention. In the second optimization of incorporating semantics in similarity calculation, comparing lexical semantics in WordNet is the continuation of a previous work (Zhou and Wei 2008) done by a collaborator and the author, while the author first introduced latent semantic analysis to the semantic integration of LULC classification systems in this research. In the next section, we will present the results of applying this method (and optimizations) to the integration of LULC classification systems. A discussion based on our method’s performance follows.

2.3 Results and Discussion

2.3.1 Experimental Results

The experiment is to apply the method introduced in section 2.2 to the comparison of LULC classification systems used in the State of Maryland (MD LULC), Delaware (DE LULC), and New Jersey (NJ LULC) (Appendix I). The MD LULC data is prepared by Maryland Department of Planning (http://www.mdp.state.md.us/) based on aerial photo during year 2001 to 2002. The DE LULC data is downloaded from Geospatial One Stop data portal (http://gos2.geodata.gov/wps/portal/gos), prepared by the Office of State Planning Coordination (OSPC) of the Budget Development, Planning, and Administration Section of the Delaware Office of Management and Budget (http://stateplanning.delaware.gov/). The NJ LULC data set was prepared by Aerial Information Systems, Inc., Redlands, CA, under direction of the New Jersey Department of Environmental Protection (NJDEP), Bureau of Geographic Information System (BGIS) (http://www.state.nj.us/dep/gis/).

The MD LULC classification system has 22 categories. Each category is defined by a name and a paragraph of textual descriptions. In general, there are 20 to 30 words in the description. The NJ LULC classification system is defined in 79 categories, and each category contains 40 to 80 words. There are 38 categories in DE LULC classification system, and DE LULC is not elaborated by textual descriptions, only category names are given.

For every pair of categories from different classification systems of the three, the similarity values are calculated using the method including variations discussed in section 2.2. For a given category, its matching categories are the categories with similarity larger than a threshold. Obviously, the threshold controls the matching results. A lower threshold leads to more but less accurate matching categories, while higher threshold allows less but more accurate matching categories. Matching results achieved by algorithm are compared against human evaluation.

Human evaluation gives matching categories for each category, but not the similarity values. The matching categories are decided by graduate students of the Department of Geography, University of Maryland College Park. Human evaluators match LULC classification systems in a one-way fashion, that is, from the system with more categories to the system with fewer categories. For example, to match NJ LULC to MD LULC, for each category in NJ LULC classification system, the evaluators pick up one matching category or several matching categories from MD LULC classification system. If two classification systems (e.g. MD LULC and DE LULC) are approximately on the same level, two-way matching is enabled by switching the source and target. Therefore, four groups of evaluation are available to use NJ to MD (NJ2MD), NJ to DE (NJ2DE), MD to DE (MD2DE), and DE to MD (DE2MD). Evaluators are informed that ‘no matches’ is acceptable. The LULC map is not revealed to evaluators; therefore evaluators made decisions only based on category names and textual descriptions, rather than quantitative methods and direct observation on actual lands.

An automated semantic based data integration method applied to geospatial data portals (Zhou and Wei 2008) is used for comparison purpose. The semantic similarity measurement in that research uses a WordNet enriched feature-based model to compare "semantic factors" in two LULC categories. The semantic factors are keywords extracted by eliminating stop words and negated words from category descriptions.

The metrics to evaluate the algorithm performance is the widely used precision recall metrics. Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness. When using precision and recall in matching LULC classification systems, the set of possible matches for a given LULC category is divided into two sets, one of which is approved by human evaluation and considered "correct". Consider the multiple matches for a single category given by human evaluation are in an alternative relationship, recall is calculated as the number of categories which are given correct match(es) by algorithm divided by the number of all categories. Precision is then computed as the fraction of correct matches retrieved by algorithm among all retrieved matches.

As the threshold increases from 0 to 1 at the interval of 0.01, 100 pairs of recall and precision are calculated at each threshold, among which the general trend would be decreasing in recall and increasing in precision. Based on these 100 pairs of recall and precision, a precision-recall curve will be plotted using the recall as an independent variable on X axis, and the precision as a dependent variable on Y axis. Examining the entire curve is informative yet its saw-tooth shape is may blur the trend. To remove the jiggles, the classic simplification is the 11-point interpolated precision (Manning et. al. 2008), which measures the interpolated precision at the 11 recall levels of 0.0, 0.1, 0.2, . . . , 1.0. Then an average of this 11-point interpolated precision can be used to compare the overall performance of different methods. In Figure 2.3, we present the average of 11-point interpolated precision of each algorithm variation for matching DE LULC to MD LULC (a), MD LULC to DE LULC (b), NJ LULC to MD LULC (c), and NJ LULC to DE LULC (d).

Both the basic algorithm (the cosine similarity on bag-of-words representations) and its potential optimizations (keyword enhancement and semantics incorporation) are tested. In Figure 2.3, they are denoted by different labels. The first two letters in the label indicate what are in the "bag of words". "KW" stands for keywords only, "AW" for all words without keyword enhancement, "KE" keyword enhancement, and "NM" for category name only. As aforementioned, in KE, keywords are enhanced by multiplying the occurrence weight by 1.5. The second part of the label indicates the similarity measurement: "O" stands for orthogonal (standard) vector space model, "WN" stands for using WordNet to quantify lexical semantics, "LSA" stands for latent semantic analysis, and "FB" stands for the feature-based model as comparison (Zhou and Wei 2008). The number (1, 2, or 3) after label "WN" indicates the maximum number of "hops" that will be considered between nodes (words) in WordNet.

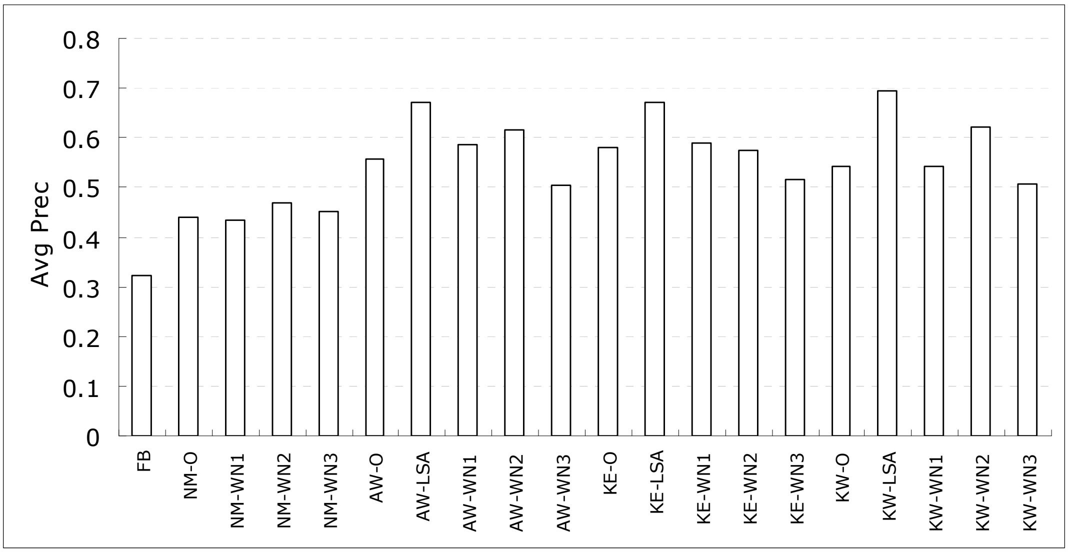


Figure 2.3-a Average precision when matching DE LULC to MD LULC

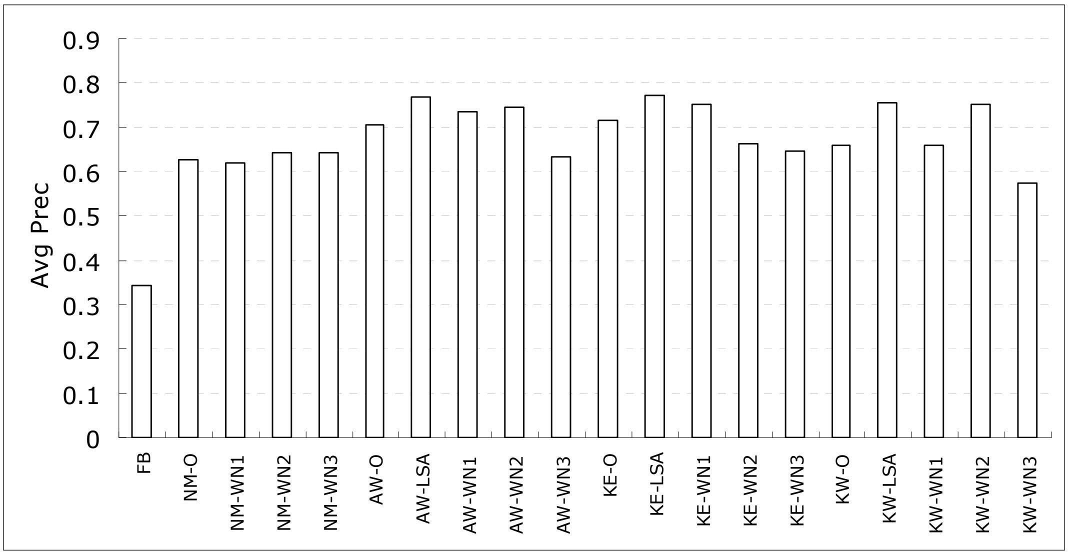


Figure 2.3-b Average precision when matching MD LULC to DE LULC

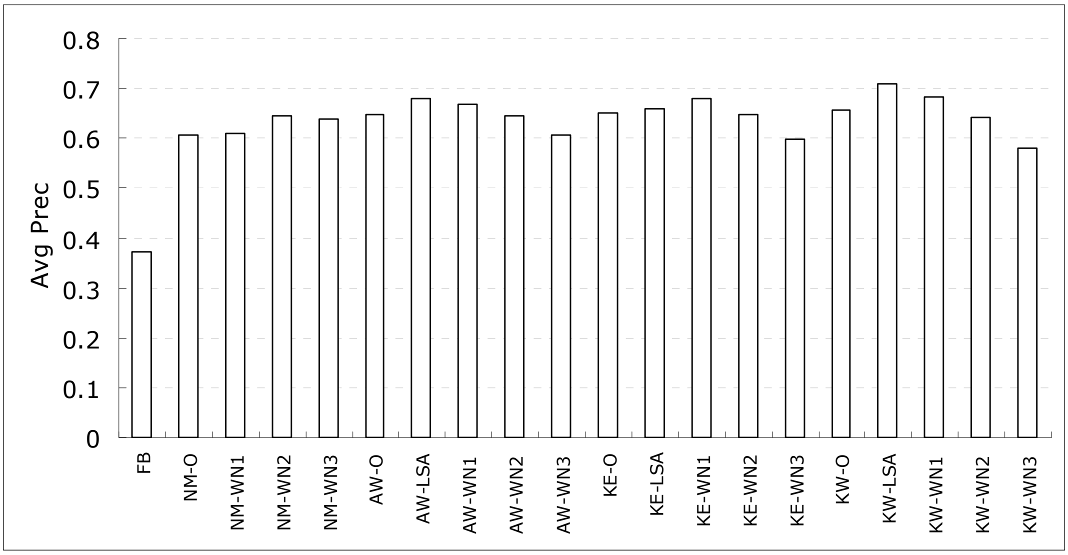


Figure 2.3-c Average precision when matching NJ LULC to MD LULC

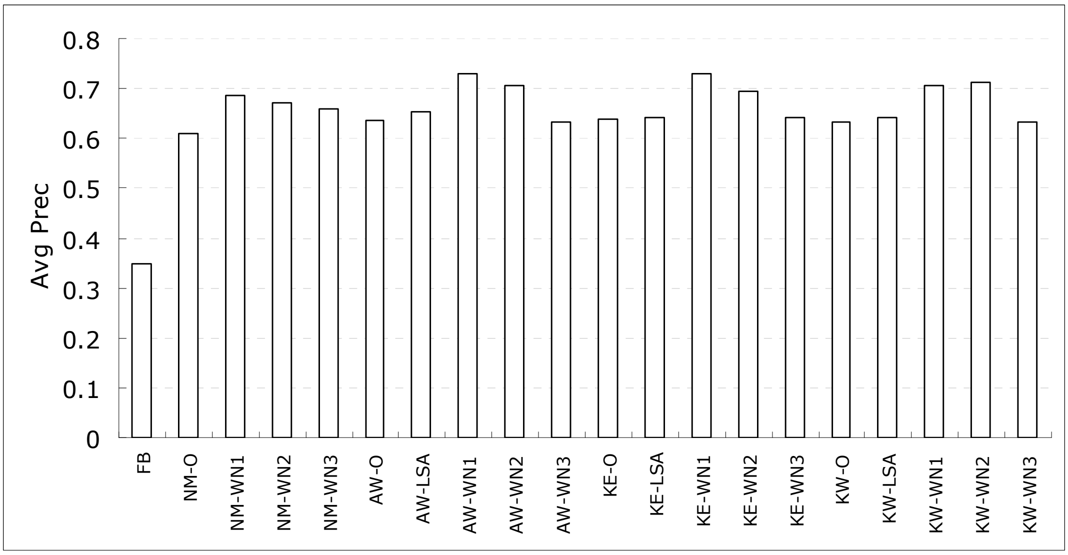


Figure 2.3-d Average precision when matching NJ LULC to DE LULC

2.3.2 Discussion

Over-performing feature-based model

In Figure 2.3, one of the most dominant findings is that given the same input, our algorithm has a great improvement over feature-based model (FB) in all four groups of experiments, no matter which variation of algorithm is in comparison. The performance score of feature-based model at best is at around 60% of our algorithm without optimizations. This result shows our algorithm can provide integration results much close to human evaluations than previous works.

Always Incorporating Descriptions

For each algorithm variation, using words in descriptions (AW-) leads to better performance than only using words in category names (NM-). This result supports our initial underlying assumption: incorporating textual descriptions in semantic integration is always favorable.

Keywords Enhancement

Removing the under-performers i.e. feature-based model and methods that only use words in names, we re-organize the order of methods in Figure 2.3 to better capture the comparison between different variations of our method (Figure 2.4).

When excluding consideration of semantics during similarity measurement, methods using keyword enhancement (KE-O) performs better than methods (AW-O) without keyword enhancement in all the four experiments, among which the largest improvement of 4% happens when matching DE LULC to MD LULC and small improvement from 1% to 2% is achieved in other experiments. We expect this result.

But using keywords alone (KW-O), the performance is uneven comparing to methods without keyword enhancement (AW-O): 3% and 7% drop when matching DE LULC to MD LULC and MD LULC to DE LULC, 2% increase when matching NJ LULC to MD LULC, and even performance when matching NJ LULC to DE LULC.

It seems keyword extraction improves the precision by eliminating words of less semantic importance, but it is inevitable for the process to drop some meaningful words along the way and reduce the conformance rate. As the language in LULC category description is supposed to be concise and complete, it is probable the information loss due to dropping words overshadows the gain of precision. Methods with keyword enhancement (KE-O), on the other hand, emphasized keywords without dropping words, and hence optimized the trade-off and gave better performance than AW-O. When using WordNet to capture semantic relation within one hop, methods with keyword enhancement (KE-WN1) also over-performed regular method (AW-WN1), while methods using only keywords have an uneven performance. But if we extend the semantic relation in WordNet deeper to 2 hops, using only keywords can be beneficial occasionally. To better understand this finding, we move ahead to the next discussion, in which we will examine the effectiveness of using WordNet and LSA to capture semantics in LULC descriptions.

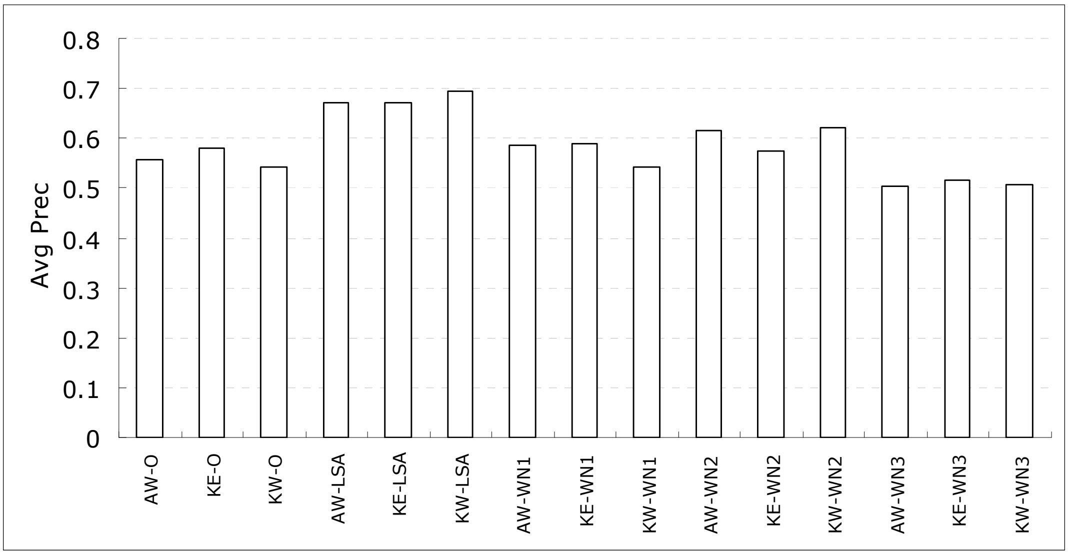


Figure 2.4-a Re-organized average precision when matching DE LULC to MD LULC

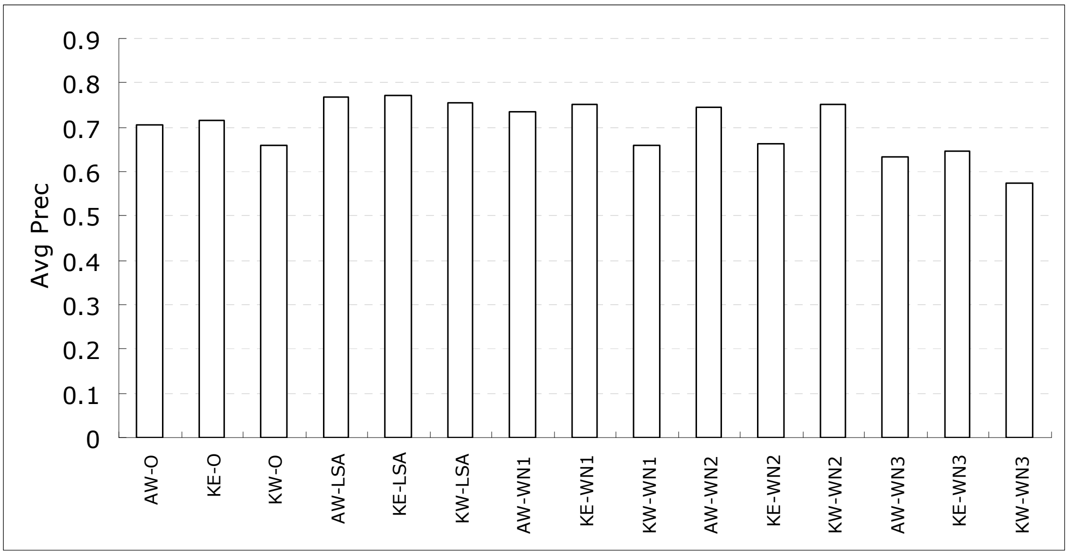


Figure 2.4-b Re-organized average precision when matching MD LULC to DE LULC

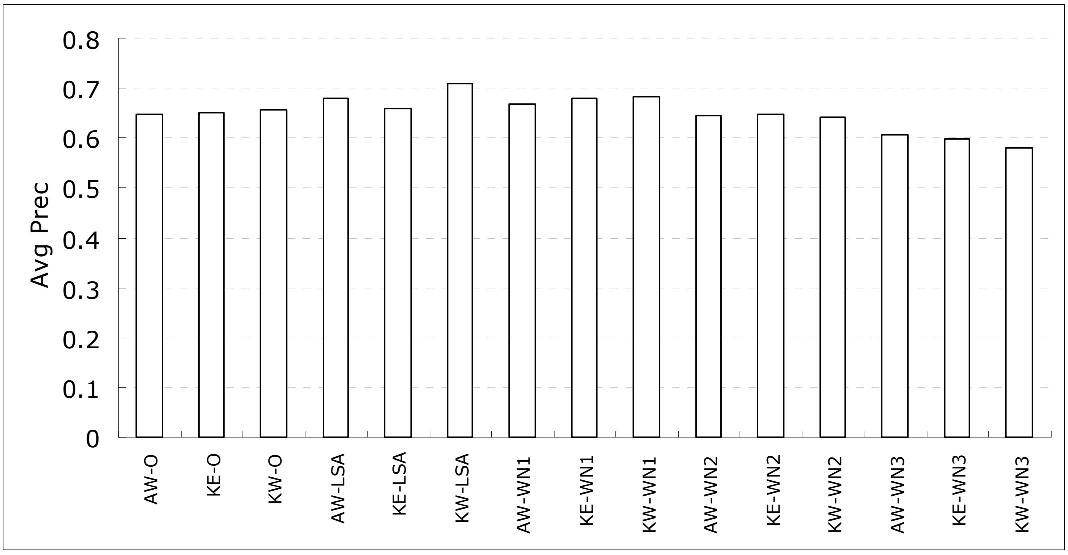


Figure 2.4-c Re-organized average precision when matching NJ LULC to MD LULC

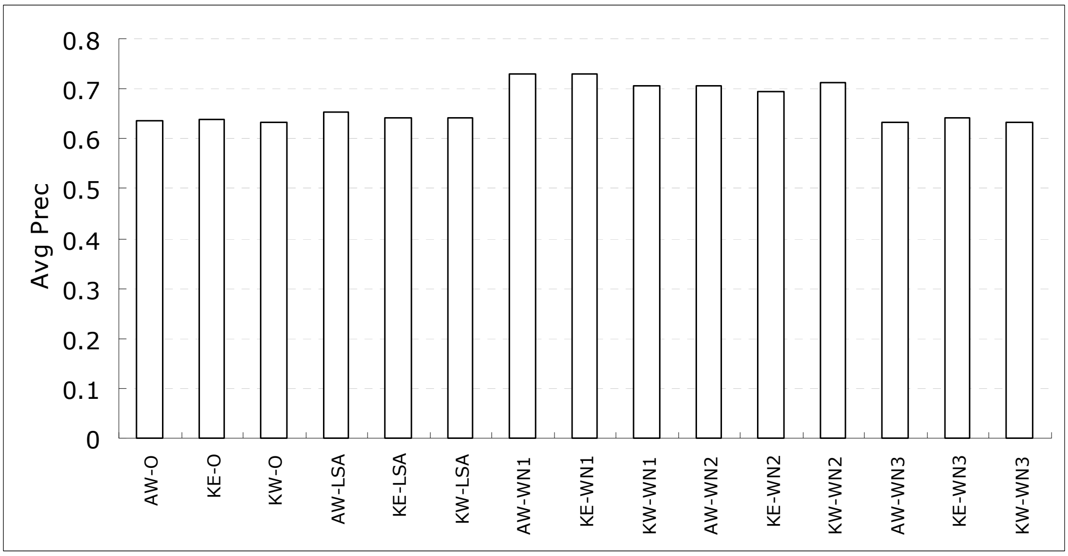


Figure 2.4-d Re-organized average precision when matching NJ LULC to DE LULC

LSA or WordNet

As discussed in 2.2.4, we employed two approaches, which are respectively based on LSA and WordNet, to incorporate semantics in comparing LULC descriptions. As results shown in Figure 2.4, either approach optimizes the methods and leads to better matching results than method using standard cosine similarity in all four experiments. Moreover, methods using LSA give the best performance in three experiments except matching NJ LULC to DE LULC, in which methods using WordNet prevailed.

For methods using WordNet, as the searching digs deeper in lexical relations, semantically remote words are more likely to be connected, which in turn links categories that are not considered to have a relation in LULC classification systems. For example, in WordNet, the word "service" and "home" are related within three hops in WordNet, which will lead to a false relation between Commercial and Residential categories. As the set of retrieved matches expands, more false matches are returned and decrease the precision. Using only keywords instead of keyword enhancement in comparison can reduce false relations when the search depth is 2, but when searching depth is 3 (-WN3), it was hard to defend the performance from many false relations.

Another issue when using WordNet to quantify semantic relatedness in LULC descriptions is that important topical relations in LULC descriptions may be omitted in WordNet. WordNet is a repository built upon "is-a" and "part-of" relations. Although there is a 'domain of synset – TOPIC' relation defined in WordNet, it is largely diluted in the whole vocabulary. Therefore, many topical relations that are important in measuring LULC concepts' semantic similarity are not acknowledged in WordNet. For example, in the context of LULC, closely related words – "agriculture" and "cropland", are not linked by WordNet, because neither "is-a" nor "part-of" can describe their topical relation.

These two issues restrict the WordNet's adaptability in comparing LULC descriptions. After all, WordNet is designed to relate concepts in general. Hypothetically, a well developed subset of WordNet containing words frequently used in LULC context and incorporating more topical relations could provide a better semantic reference, but building such an ad-hoc WordNet is not easier than manual semantic integration.

The model based on LSA, on the contrary, does not have these drawbacks because it is built upon the word frequency and co-occurrence in the LULC context. In general, it has superior performance. But when matching NJ LULC to DE LULC, although beating standard cosine similarity, LSA methods does not perform as well as methods using WordNet. After looking into the matching result, we find the reason rooting in the limitation of LSA's fundamental assumption. LSA assumes that semantic related words tend to occur together in text. This assumption is generally true but not in all circumstances, especially when two LULC categories are defined in different but semantically related words, which are rarely seen outside the two descriptions. For example, Saline Marshes in NJ LULC should be matched to Wetland DE, as marsh is a kind of wetland. Method using WordNet captured the semantic relation between marsh and wetland, while LSA missed it due to the lack of co-occurrence of term marsh and term wetland in the whole collection of LULC descriptions.

2.4 Conclusion on Lexical Semantic Integration

Our algorithm shows animprovement over previous feature-based method in all four of experiments: the 11-point interpolated precision of feature-based model at its best experiment reaches only 60% of our basic method's precision.

Based on all four experiments in this research, we recommend the methods using LSA, which over-performed other methods in three of four LULC classification matching experimetns. The benefit of keyword enhancement over LSA-based methods is not obvious (less than 3%), because denoising is already achieved in the low-ranking process of LSA. Considering LSA's mechanism, we believe the larger collection of LULC descriptions is in comparison, the better LSA will perform. But if LULC category descriptions are too brief and the whole collection of descriptions is too small, using co-occurrences may not be able to reconcile word mismatches (e.g. marsh and wetland). In this circumstance, the combination of keyword enhancement and using WordNet to quantify semantic relatedness should be working.

However, in the example mentioned before, lexical methods using LSA did not match Saline Marsh to Wetland due to word mismatch. Goh (1997) described this incompatibility of concepts because of word mismatch as one type of semantic heterogeneity, and termed it "naming conflict". Naming conflicts do not only happen in LSA based methods, but can also be found in WordNet based methods. For example, human evaluators consider Recreational of DE LULC and Open Urban

Land of MD LULC to be similar. However, all the lexical methods were lost in naming conflicts and none of them achieved this match.

More importantly, all integration methods using lexical semantics have a strong limitation. They are depending on how LULC categories are defined (in text) rather than how they are used in the field, but the usage of a LULC category does not necessarily follow its definition (Duckham & Worboys 2005). For example, the LULC category Wetland appeared in both MD LULC and DE LULC classification systems are matched together by human and algorithm. But by observing parcels labeled as Wetland in MD and DE, we find their main difference. Wetlands in MD always contain more water and less vegetation, and only resemble some of wetlands in DE, while other DE Wetland parcels are highly vegetated and resemble forests. Therefore Wetland MD is more like a subcategory of Wetland DE. Two concepts using same or similar labels are actually different, this underlying semantic heterogeneity is termed “confounding conflict” by Goh (1997). In the study of LULC, the confounding conflict may lead to severe cognitive problems when they are covered up by lexical semantic methods.

To overcome semantic heterogeneities that are not solvable using lexical semantic methods, we will further our research beyond lexical semantics into actual LULC status, and focus on using remotely sensed data to translate semantic heterogeneous LULC classification systems in the next chapter.

Chapter 3: Overcoming Semantic Heterogeneity Using Remote Sensing

3.1 Introduction

In chapter 2, we tried to compare and match LULC classification systems using lexical information contained in category name and descriptions. Our method is based on bag-of-words model and cosine similarity enhanced by LSA or WordNet, and the results showed great improvement comparing to previous feature-based model. However, we also found out that lexical semantic methods are having difficulties solving the semantic heterogeneities happened between different classificatory approaches.

In general sense, semantic heterogeneities are categorized into three types (Goh 1997). The *Naming conflict* means the naming schemes of the information differ. The *Confounding conflict* happens when information items (e.g. LULC labels) seem to have the same meaning, but differ in reality. The *Scaling and units conflict* happens when different reference systems are used to measure a value. The goal of semantic integration is to eliminate all three types of semantic heterogeneity. Comparing to the confounding conflict and naming conflict, the scaling and units conflict is straightforward and less difficult to solve. The focus of most research on semantic integration, including this one, is therefore on the first two types of semantic heterogeneities.

In LULC classification systems, the confounding conflict happens when labels and descriptions of two LULC categories seem to represent same or similar LULC status, but in reality different. For example, category wetland represents how complicated and semantically heterogeneous a LULC concept can be. Due to variations in vegetation species and coverage, water table height and period, parcels labeled as wetland in different areas can be way different in actual land cover status. It might be too easy for a lexical method to match Wetland MD to Wetland DE, without the consideration of the potential difference in their conceptual intensions. In Figure 3.1, although sharing the same label, wetland parcels in MD and DE have very different seasonal NDVI curves (calculated from time series Landsat ETM+ imagery from July 2001 to august 2002 path 14 row 33). To better understand the cause of this difference, a further discussion can be found in section 3.3.5. But intuitively, we can tell the two Wetland categories are different and inappropriate to be used interchangeably.

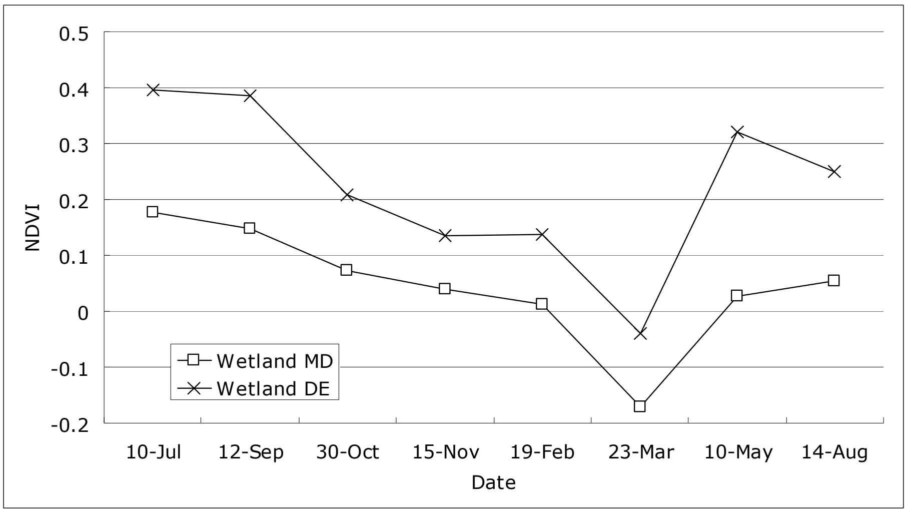


Figure 3.1 Average seasonal NDVI of Wetland parcels in MD and DE

On the other hand, the naming conflict happens when LULC categories under different labels represent same or similar LULC type from observation. For example, human evaluators consider Recreational of DE LULC and Open Urban Land of MD

LULC to be similar categories. However, different naming schemes make the lexical methods lost in naming conflicts and none of them achieved this match.

To discover confounding conflicts and reconcile naming conflicts, first, we need to know if two LULC categories are seemingly different or similar, then one step forward, we need to know if they are different or similar in reality. Our method in chapter 3 aimed to answer the first question and provided better solutions than previous methods. But answering the second question is beyond the capability of lexical methods and therefore calling for innovative approaches.

Fortunately, LULC status is observed by remote sensing, which provides a different way of understanding LULC categorization and how each category is used on the ground. It is logic to incorporate remote sensing into the integration of LULC classification systems. Before the discussion of how to incorporate remotely sensed data, we ought to briefly review the relation between remote sensing and LULC, through which we will find the reason and theoretical support of our method.

3.1.1 Remote sensing as a tool of understanding LULC

Remotely sensed data are increasingly used to describe LULC in the form of LULC mapping. As a means of observation on Earth, the remote sensing sensor measures the amount and spectral distribution of the solar energy reflected from the earth surface to infer the nature of the reflecting surface. A fundamental assumption here is that each type of earth surface (different types of vegetation, soil or sand, water, man-made surface, and so on) has an "individual and characteristic manner of interacting with incident radiation" that is described by the spectral response of that surface (Mather 2004). Generally, this assumption is reasonable and its validity is proven by decades of research in the remote sensing of environment. The spectral response curve of a piece of earth surface, which indicates its individuality, is termed the spectral signature, which is the foundation of the LULC mapping based on remote sensing. As remote sensing has such a great impact on LULC mapping, modern LULC classification systems are designed to use with remotely sensed data (Anderson 1976).

Now the remote sensing research community has become the provider of LULC information to a wider society. However, it is reported that the remote sensing community has concentrated too much on technical issues but not enough on semantic and ontological issues (Comber et al. 2004). Different disciplines or different agencies have different perceptions of land use categorization, which lead to semantically heterogeneous LULC classification systems. Interoperability of the LULC data will be impossible without the semantic integration of classification systems.

A typical LULC classification system organizes categories in a taxonomy structure, in which each category is defined by a name (i.e. label) and a paragraph of description explaining the name. The textual information extracted from names and descriptions is usually used in lexical methods (Kavouras et al 2005, Feng and Flewelling 2004, Rodriguez et al 1999, Rodriguez & Egenhofer 2003, 2004, Ahlqvist 2008, Kavouras and Kokla 2002) to integrate LULC classification systems. But all these methods have common problems: 1) it is difficult to quantify lexical semantics and convert them into comparable attributes; and 2) these methods are vulnerable to semantic heterogeneities. The first problem makes these methods difficult to use and even more difficult to automate. In chapter 3, we developed a new method using information retrieval techniques that aimed to solve this problem. But the second problem can be more severe because it may lead to deep conceptual ambiguities – the use of a LULC classification system does not reflect its definition (Duckham & Worboys 2005).

Recognizing this problem, Worboys and Duckham (2002) sought in spatiality for solutions. They believe a universal set of 'semantic atoms' exists among heterogeneous data sources, and semantic heterogeneity is entirely due to different groupings of the atoms. To integrate two classification systems of the same geographic area, Duckham and Worboys (2005) overlaid and intersected the different LULC maps, which splits land parcels into fragments. Each fragment is labeled by two categories from two classification systems. Based on this fragmentary land parcel map, Duckham and Worboys adopted a definition of semantic relations between two LULC categories based on how many fragments are shared by these two categories. Although acknowledging the importance of spatial attributes, Duckham and Worboys's method is geographically restricted: their method can only integrate LULC classifications mapping the same area.

Durbha and King (2005) introduce ontology to enable content-based image retrieval in remote sensing archives. First Durbha and King apply an unsupervised segmentation algorithm (Deng and Manjunath 2001) to extract homogeneous objects from remote sensing data archive. Then the objects are labeled by concepts in the ontology pre-developed from Anderson classification system (Anderson et al. 1976).

Based on the reasoning function supported by Web Ontology Language (OWL-DL) (Wache et al. 2001), the semantic-enabled content-based image retrieval is realized. Similar method is then applied to semantic reconciliation in data archives (Durbha et al. 2009). Worth noting is that Durbha and King's (2005) work, although in a different field, suggests a potential solution to the integration of LULC classification systems: reusing the remotely sensed data that have been widely used in LULC mapping.

3.1.2 Remote sensing to overcome semantic heterogeneity in LULC

Given its important role in LULC mapping, we believe using remotely sensed data to describe LULC categories can overcome semantic heterogeneities, because the difference in actual LULC status, either expected (between semantically different categories) or unexpected (between semantically same or similar categories), corresponds to different pixel values in the remotely sensed images due to spectral signature. Remotely sensed data have two distinctive advantages in understanding and comparing LULC categories: 1) an objective observation on the physical LULC status and 2) a quantitatively comparable measurement across LULC categories from different classification systems in different locations. Based on these two advantages, we now can answer the question if two LULC categories are different or similar in reality, and sweep the restriction of only comparing LULC classifications in same geographical areas incurred in Duckham and Worboys (2005)'s method.

In this chapter, we will present a remote sensing based approach to the semantic integration of LULC classification systems. Our method will use spectral and textural information derived from time series remotely sensed data to measure similarity between LULC categories from different classification systems, and then discovers confounding conflicts and reconciles naming conflicts. Adopting this approach can not only automate the integration process, but also provide more reliable integration results. In the rest of this chapter, we will detail our method in section 3.2, and present the matching result and compare it with human evaluations in section 3.3. Then a conclusive remark in section 3.4 will follow.

3.2 Methodology

In LULC mapping, the semantics of each category is employed to label the geospatial data, either on pixel basis or parcel basis. In other words, LULC mapping is the process of endowing the LULC semantics to the geospatial data (mostly remotely sensed) and producing LULC maps, in which each labeled parcel embodies the semantics of its LULC category. By this logic, overlaying LULC map on remotely sensed image and exploring the parcel level data patterns is a straightforward way to restore the link between semantics and remotely sensed data. Hence, our method compares the semantics of different LULC categories via the parcel level patterns in spectral and textural. The method is demonstrated through the experiment of matching LULC classification systems of Maryland (MD LULC) and Delaware (DE LULC).

3.2.1 Study area and LULC data

The study area, eastern Maryland and Delaware, is covered by a single Landsat scene (Path 14; Row 33) (Figure 3.2). As adjacent areas, eastern Maryland and Delaware have similar LULC types due to similar climate, topography, and hydrology, but these similar LULC types are defined in different classificatory approaches (Figure 3.3).

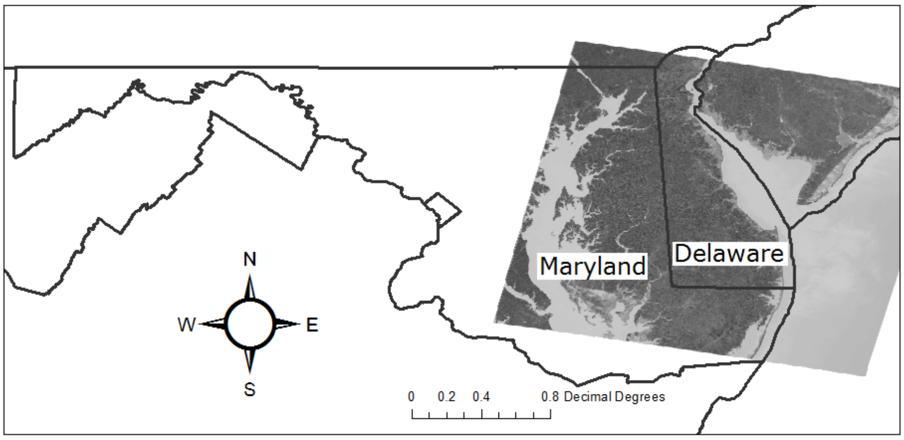
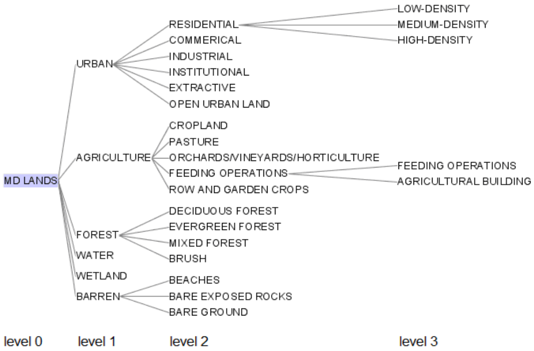


Figure 3.2 Study area: MD and DE covered by Landsat scene path 14; row 33



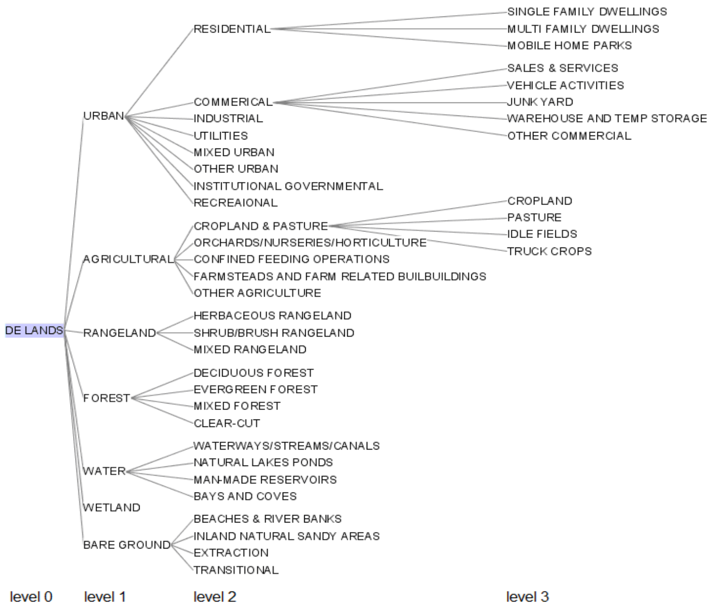


Figure 3.3 Taxonomy of MD and DE LULC Classification Systems

The Maryland LULC (MD LULC) data is prepared by Maryland Department of Planning (MDP: http://www.mdp.state.md.us/) and obtained from their website, and the Delaware LULC (DE LULC) data is prepared by the Office of State Planning Coordination (OSPC) of the Budget Development, Planning, and Administration Section of the Delaware Office of Management and Budget (http://stateplanning.delaware.gov/), and obtained from Geospatial One Stop data portal (http://gos2.geodata.gov/wps/portal/gos). Both LULC datasets are created from aerial photo during year 2001 to 2002. The detailed classification systems used in two LULC maps are shown in Appendix I. Both classification systems are defined in taxonomy (Figure 3.3), in which level 0 categories, denoted by MD LANDS and DE LANDS, representing the generic concept of land, contain all land parcels. Level 1 categories are categories directly subsumed by level 0 categories, and so on so forth. There are 4 different levels (0-3) of categories in both classification systems, among which level 1 and level 2 approximately correspond to same levels in the Anderson system (Anderson 1976). All cateogries in both classifications have parcels covered by Landsat scene path 14 row 33.

3.2.2 Assessment of the LULC classification systems

Before the matching process begins, a thorough assessment of the LULC classification systems is needed to discover the conceptual ambiguities and errors. Otherwise, these ambiguities and errors will never be found but inherited to the matching process and make the integration error-prone.

An expert in LULC may have different ways to decide how well a LULC classification system is defined, but there is a rudimentary form to create classifications termed "*facet analysis*" (Ranganathan 1967). It provides a collection of rigid rules, which all classification systems should follow. Facet analysis is invented by information scientist Ranganathan in the 1930's, and primarily used to create classifications for the document collections in technical, scientific, and social scientific fields. In this tradition, facets are, in Wynar and Taylor's words (1992), "clearly defined, mutually exclusive, and collectively exhaustive aspects, properties, or characteristics of a class or specific subject." In simple words, facets are the characteristics of division. Ranganathan's facet analysis consists of three planes, each of which has several canons, postulates, and principles. Cannons are the rules must be followed by all classification systems, while postulates and principles are strong recommendations.

Ranganathan’s original facet analysis is presented as a detailed series of 46 canons, 13 postulates and 22 principles, many of which are beyond the scope of this research. Spiteri (1998) proposed a simplified model for facet analysis by combining and simplifying of Ranganathan’s canons, postulates, and principles, and the principles defined by the Classification Research Group (CRG). In this study, the assessment of LULC classification systems is based on Spiteri’s simplified facet analysis.

PRINCIPLES FOR THE IDEA PLANE

Principles for Choice of Facets

a) Principle of Differentiation

This principle advises that when dividing an entity, facets should have the ability to distinguish clearly among its component parts. In the LULC mapping, it means that for a land parcel, it should be decisive in determining which one and only one category this parcel belongs to. The definition of MD LULC classification system follows this principle well at level 1 categories. But at level 2, conceptual ambiguity is observed. For example, the definition of Mixed Forest poses a question on the differentiation between the mixed forest, and the deciduous and/or evergreen forest. Its definition says neither deciduous nor evergreen species dominate, and mixed forest is a combination of both. This is a vague definition as it did not define “dominate”.

The DE LULC classification system has several mixed categories, such as mixed urban, mixed rangeland, and mixed forest. As in the Anderson’s LULC classification system (Anderson 1976), DE LULC uses a threshold of 1/3 intermixture area to separate mixed from "pure" parcels, although accurate delineation of intermixture area might still be unclear.

b) Principle of Relevance

This principle advises that the choice of facets should reflect the purpose of classification. Both MD and DE LULC classification systems are produced and used by state urban planning authorities. The principle of relevance is followed in both systems, and can possibly explain some of the unbalanced extension of the classification system. Both classification systems have the highest level 3 subcategories for the category of residential, but do not divide level 1 category of wetland, because further categorizing wetland, an extremely important and notoriously complicated concept in LULC, is beyond the focus of urban planning.

c) Principle of Ascertainability

This principle advises that facets should be definite and ascertained, which was explained by Spiteri, for example, the date of death is inappropriate to use as a facet to divide live people, because it is unknown. Similarly, in LULC mapping, categorizing lands should be based on current status rather than planned uses in future. MD and DE LULC classification systems followed this principle.

d) Principle of Permanence

This principle advises that facets should reflect the permanent qualities of the entity to be divided, which means, for example, color is not applicable to classifying the chameleons. However, this reasonable principle is not easy to follow in LULC classification systems, because the LULC status, which is the basis of all facets, can change especially for transitional land uses. The impact of LULC change is expected to see in the integration results.

e) Principle of Homogeneity and Principle of Mutual Exclusivity

These two principles advise that each facet used to divide entities should represent only one characteristic of division and mutually exclusive among each other. These two principles ensure on each specific facet each component part is homogeneous, but mutually exclusive among each other. Each item in the classification has its own unique place (Spiteri 1998).

These two principles are strong restrictions. Homogeneity and mutual exclusivity are both relative to the discriminating power. Using a higher discriminating power, homogeneous categories can be divided further more. On the contrary, a lower discriminating power blurs the boundary between originally mutual exclusive subcategories. If we reasonably assume the same discriminating power is used through out the entire LULC mapping process, some LULC categories in MD and DE classification systems are less homogeneous than others. To this end, Wetland, again, as the only level 1 "leaf" category (no subcategories) in both systems stands out. It is obviously under-defined and may cause semantic heterogeneity. Different subcategories of Commercial in DE classification system, on the contrary, are more questionable on their mutual exclusivity.

f) Principle of Fundamental Categories

Different from Ranganathan's PMEST formula (Personality, Matter, Energy, Space and Time), this principle advises that there exist no fundamental categories (of facets) to all subjects. Facets should be derived based upon the nature of the subject being classified (Spiteri 1998). Every LULC classification system by nature is a good practice of this principle.

Canon of Exhaustiveness

In Ranganathan's original facet analysis (1967), the canon of exhaustiveness states that all classes and sub-classes in a classification system should present all aspects of their parent universe. This canon is excluded in Spiteri's simplified facet analysis, because Spiteri (1998) argues that the exhaustiveness is rather "difficult to determine and maintain". Spiteri is right about the difficulty, and it is reasonable to remove a hard-to-follow principle from the must-follows. However, exhaustiveness should be recommended, because without exhaustive subcategories, it is inevitable to see some items of the parent category but belong to none of its subcategories, which is a logical flaw.

Some LULC classification systems, such as the Anderson system (Anderson 1976), escape this predicament by adding in a "catch-all" subcategory, such as Other urban or built-up land. Setting up a "catch all" is good when it comes to assigning every and each land parcel a subcategory to avoid logical flaw. But the semantic clarity of this "catch-all" category is very much a challenge. DE LULC classification system is a modified Anderson system, and there is a catch-all category Other Urban or Built-up for urban land uses in the system. But in addition to this catch-call category, DE LULC includes urban categories such as Recreational and Utility, which in Anderson system composes the main part of the Other Urban or Built-up. Now the category of Other Urban or Built-up in DE LULC has an unclear intension, and is difficult for human evaluators to find a match in MD LULC.

Principles for Citation Order of Facets and Foci

a) Principle of Relevant Succession

This principle advises that the order of facets should reflect their natural scopes in the classification system. In LULC classification systems, this principle is usually well followed: parent categories represent broader LULC concepts than subcategories. The MD and DE LULC classification systems follow this principle closely.

Principle of Consistent Succession

This principle advises that the order of facets should not be modified once it is established, unless there is a change in the purpose, subject, or scope of the classification (Spiteri 1998). This principle is followed by MD and DE LULC classification systems.

Spiteri's simplified facet analysis (Spiteri 1998) also has the principles for the verbal and notational planes, which are about naming and coding the classification systems. These principles are beyond the scope of this research.

In Spiteri's simplified facet analysis on MD and DE LULC classification systems, several conceptual ambiguities are discovered. Solving these ambiguities are out of the reach of any lexical semantic methods, and the impact of these ambiguities on integration is estimated in the following result and discussion section (section 3.3).

3.2.3 Remote sensing data selection

Geospatial data are the direct observation of geospatial entities, such as land parcels. Each type of geospatial dataset, raster or vector, provides a unique perspective of observation and information into the nature of the geographical phenomenon (Durbha et al. 2009). In theory, every type of available geospatial data could be beneficial in terms of bringing the unique information into the algorithm. But in practical terms, each type of remotely sensed data has its individual scope of application, which is determined by the spatial, temporal, spectral, and radiometric resolution. In this study, we select the time series Landsat 7 ETM+ imagery to be the remotely sensed data source (Table 3.1) based on the following considerations.

Spatial resolution

Landsat ETM+ has eight channels covering the visible, near- and mid-infrared, and the thermal infrared, including a panchromatic channel. The panchromatic channel has a spatial resolution of 15 m, the thermal infrared channel has a spatial resolution of 60 m, and the rest six spectral channels have a spatial resolution of 30 m. In this study area, most parcels have an area larger than 10000 square meters. They can be captured by Landsat ETM+. Another consideration is the swath width. A Landsat ETM+ scene has a swath width of 185 km, which can cover the whole study area. It is practically beneficial that single scene coverage can save lots of efforts to balance the impact of the solar and sensor variation if using different scenes.

Temporal resolution

Landsat 7 has a repeat coverage interval of 16 days. At this temporal resolution, a time series of images are available to capture the phonological phenomenon, which is very important to differentiate vegetated LULC types. Even when clouds substantially occupy in the images on some dates, it is not difficult to find alternatives from a 2-3 years of collection.

Spectral and radiometric resolution

Landsat 7 ETM+ is an earth observing instrument. It is a proper choice for this research because it is designed to discriminate different LULC types via its eight spectral bands covering a rich range from visible to short-wave infrared in the electromagnetic spectrum.

Data availability

The method discussed in this chapter is data-oriented. The data availability is an important concern when the method is to be applied to broader and different areas. Landsat has a global coverage at the same spatial and temporal resolution, which enables this integration method to be applied to almost everywhere in the World, even across different continents, where lexical semantic methods are largely disabled because the LULC classification systems can be defined in different languages.

Table 3.1 lists the Landsat data used in comparing MD and DE LULC classification systems, including eight spectral bands (including thermal low gain (band 61) and high gain (band 62)) and one panchromatic band on eight different dates, scattered through a year time (2001-2002). In addition to spectral measurement, textural measurement Grey Level Co-occurrence Matrix (GLCM) (Haralick et al.1973) is also included. The GLCM measurements used in this study are contrast, correlation, entropy, and mean, calculated from 2001-07-10 Landsat NDVI and panchromatic image, respectively.

Table 3.1 External geospatial data

Data (parcel level mean and standard deviation)
ETM+ Band 1-8 2001-07-10
ETM+ Band 1-8 2001-09-12
ETM+ Band 1-8 2001-10-30
ETM+ Band 1-8 2001-11-15
ETM+ Band 1-8 2002-02-19
ETM+ Band 1-8 2002-03-23
ETM+ Band 1-8 2002-05-10
ETM+ Band 1-8 2002-08-14
GLCM texture: Contrast, calculated from NDVI 2001-07-10
GLCM texture: Correlation, calculated from NDVI 2001-07-10
GLCM texture: Entropy, calculated from NDVI 2001-07-10
GLCM texture: Mean, calculated from NDVI 2001-07-10
GLCM texture: Contrast, calculated from Band 8 2001-07-10
GLCM texture: Correlation, calculated from Band 8 2001-07-10
GLCM texture: Entropy, calculated from Band 8 2001-07-10
GLCM texture: Mean, calculated from Band 8 2001-07-10

3.2.4 Preprocessing

Preprocessing involves the correction of deficiencies and the removal of flaws present in the data. It is carried out before the data are used for a particular purpose (Mather 2004). In this study, preprocessing includes geo-referencing (if necessary) and removal of pixels on the parcel edges.

The accurate registration is required to overlay LULC map on remotely sensed images. Geometric errors in registration lead to the displacement of pixels in land parcels when overlaying LULC map on images, and then the data pattern extracted from pixels in each parcel are subject to mistakes due to the pixel displacement. To avoid this problem, registration is carried out if LULC maps and images do not fit well.

The pixels on parcel boundaries need to be removed because these pixels tend to cross the boundary and become mixed of multiple LULC types. To remove these pixels, the LULC map is rasterized on the same spatial resolution of the remotely sensed data. Then in this rasterized LULC map, pixels on parcel boundaries are assigned a different value than the value of pixels that fall completely within parcels. This boundary raster is then overlaid on each remotely sensed image. And through a raster calculation, pixels on the boundaries of parcels are picked and removed from each remotely sensed image and textural image.

3.2.5 Parcel level statistics

In remote sensing, object-based methods (Walter 2004) consider groups of pixels that represent existing objects rather than single pixels as inseparable units in processing. The advantage of this object-based approach is obvious: real world LULC is not delineated into tiny squares but into parcels, that is, objects. Adopting the object-based approach, parcel level statistics rather than pixel values, will be used to compare different LULC categories.

In the integration of MD LULC and DE LULC, mean and standard deviation of the pixels within each land parcel are calculated from 9 Landsat bands on 8 different dates plus 2 textural bands. That is, for each parcel, there are 160 feature values in total. If we imagine a vector space with each dimension corresponds to a feature value, each parcel projects to a point in this 160-dimentional vector space, and a LULC category is a cluster of points, each point corresponding to a parcel in this category. Now the spectral, textural and temporal information are all included in this calculation.

3.2.6 Extensional approach to similarity

Based on parcel level statistics, the similarity between LULC categories is calculated via an extensional approach – estimating similarity of concepts by comparing their instances. From an ontological point of view, all "valid" scientific concepts must have instances; otherwise it is worthless in terms of scientific research, since scientific research is meant to find rules (Smith 2004). Instances, whose commonality is reflected in the concept, provide important information in semantic integration. The semantic integration methods using instance level information are termed *extensional* methods, in contrast to the *intensional* methods that concerned with only concept-level (i.e. concept definitions) and/or schema-level (i.e. hierarchy in taxonomy) information. In this study, each LULC category (also a concept) was instantiated by individual land parcels, therefore instance level (parcel level) information and an extensional approach is applicable to the similarity measurement of LULC categories.

Extensional similarity is measured in matching source LULC classification system to the target. This matching is one-way matching, but two-way matching can be achieved by simply switching the source and the target. However, it is common that LULC classification systems to be compared are not on the same level, which means two way matching is not always reasonable. For instance, subcategories such as reservoir, ocean, natural lake, waterway are easily matched to its parent category water. However, it is not reasonable to match water to any one of its subcategories. In this study, DE LULC and MD LULC are approximately on the same level, and two way matching is possible.

3.2.7 Use of SVM classifier

In order to match DE LULC to MD LULC, if the algorithm can assign a DE LULC category to a parcel in MD LULC map, which among all the DE LULC categories has the most similar statistics to the MD parcel, a parcel-level match from the MD parcel to a DE LULC category is established. As a MD LULC parcel belongs to a MD LULC category, when it is matched to a DE LULC category by algorithm, this parcel-level match contributes to the match between the two categories. Quantitatively, the number of parcel-level matches indicates the strength of the category match, that is, the similarity of the two categories. Now the similarity measurement hinges on finding the assignment that maximizes the similarity (or minimizes the distance) between every DE LULC parcel and MD LULC categories in a high dimensional vector space. The problem becomes a typical supervised classification problem: a classifier can be trained from categorical parcel level statistics in MD LULC, and applied to DE LULC parcels.

Among all the classifiers available, Support Vector Machine (SVM) classifier (Cortes and Vapnik 1995, Huang et al. 2002) is selected, because of its superiority in classifying high dimensional dataset. The SVM implementation provided by LibSVM package (Fan et al. 2005) embedded in Weka software (Hall et al. 2009) is used in the algorithm. Thanks to its unique max-margin mechanism, SVM classifier has an extraordinary tolerance of errors in the training data (Song 2010), but with a side effect of sensitivity to disproportional training data. This side effect and its impact on integration will be discussed in section 3.3.5.

3.2.8 Refining SVM inputs

Before parcel level statistics can be used in a SVM classifier, an examination of its separability among all LULC categories would be essential, required by the Principle of Differentiation. A 10-fold cross-validation is carried out using SVM classifier on all parcel level statistics of MD and DE respectively, and the percentage of correctly classified parcels is a little lower than 50% in both MD and DE cross-validations. As training the SVM classifier should use representative parcels, the correctly classified parcels in cross-validation follow the patterns of their categories closely, and become the candidate training data set.

The low percentage of correctly classified parcels is because of two possible reasons: data quality flaws and varied parcels of same LULC category. Either reason may cause outliers - parcels scattered from the center of its category in the spectral space, which will confuse the boundary between categories, and then lead to mismatches in the integration. By filtering out outlier (scattered) parcels, the rest will be more cohesive and representative for their category. It is worth noticing that out task here is not to find all representative parcels, but to guarantee the training set is representative. This means outliers must be removed from candidate training set (if there are any), but false removal of several representative parcels is not a main concern.

The method of finding scattered parcels is studied in multivariate outlier detection (Rousseeuw and Zomeren 1990), and an implementation in R written by Professor Rand Wilcox (https://r-forge.r-project.org/projects/wrs/) is available. Rousseeuw and Zomeren's method picks outliers based on the Minimum Volume Ellipsoid (MVE) estimator, which is superior to classic Mahalanobis distance in its robustness. Not like Mahalanobis distance, MVE estimator is not easily biased by a small cluster of outliers.

By removing outliers from the parcels that are correctly classified in cross-validation, we have the representative parcels for each category that are ready to be used in SVM classifier for training purposes. In a new round of cross-validation runs on this training set, the percentage of correctly classified rises to more than 95% in both MD and DE training datasets.

3.3 Results and discussion

The two way matching results are showed in Table 3.2 (a) DE LULC to MD LULC and (b) MD LULC to DE LULC. In both tables, categories are denoted by codes introduced in Appendix I. The columns from left to right mean 1) DE (a), MD (b) LULC categories, 2) match(es) in MD (a) or DE (b) LULC by human evaluators, 3) conforming matches, and 4) non-conforming matches. Here conforming matches are the algorithm matches that conform to human evaluations, while non-conforming matches are those that do not conform to human evaluations. The number in the parentheses after each algorithm result (in column 3 and 4) is the similarity of that match, calculated as the ratio of the number of parcels in this match to the number of all parcels in the source category. For example, the match Warehs to Indstrl (0.545) in

Table 3.1 (a), means 54.5% of all parcels of DE LULC category Warehouse (Warehs) are matched to MD LULC Industrial (Indstrl) category, and it suggests a similarity of 0.545 between the two categories.

3.3.1 Human evaluation

Human evaluation gives matching categories for each LULC category, rather than the similarity. The evaluation was done by graduate students of the Department of Geography, University of Maryland College Park. Five evaluators are asked to find a matching category or categories from MD classification system for each DE LULC category, and the other way around. Evaluators were informed that 'no match' is acceptable in their results. The LULC data, however, is not given to evaluators. Human evaluators make decisions only based on category name, description, and *a priori* knowledge. Among 5 interpretation results, for each category, if no less than 2 votes from 5 evaluators agree on one match, the match is considered to be human interpretation result.

Table 3.2 Results of Matching DE LULC to MD LULC (a), and MD LULC to DE LULC (b).

a

DE LULC	MATCHED MD LULC		
	HUMAN	ALGORITHM	
		CONFORMING	NON-CONFORMING

SinFam	LowRes,MedRes	LowRes(0.198)
		MedRes(0.183)
			Pasture(0.116) OpenUrb(0.101)

MultFam	MedRes,HighRes	HighRes(0.484)	Comm(0.315)

MblHm	HighRes	HighRes(0.157)	LowRes(0.129) Comm(0.106)
			OpenUrb(0.106) MedRes(0.219)

Retail	Comm	Comm(0.372)	HighRes(0.108) Indstrl(0.412)

VclAct	Indstrl,Comm	Indstrl(0.625)
		Comm(0.25)	

JunkYrd	Indstrl,Comm	Indstrl(0.3)
		Comm(0.3)
			HighRes(0.1) MedRes(0.133)

Warehs	Indstrl,Comm	Indstrl(0.545)
		Comm(0.272)	

OthrCom	Comm	Comm(0.315)	Brush(0.105) LowRes(0.105)
			Indstrl(0.315)

Indstrl	Indstrl	Indstrl(0.716)	Comm(0.221)
Utility	OpenUrb,Indstrl	Indstrl(0.111)	Pasture(0.154) Comm(0.259)
MixUrb	Comm,OpenUrb	Comm(0.372)	Indstrl(0.177) Inst(0.124)
OthrUrb	Comm		Pasture(0.187) Crop(0.119)

Inst	Inst	Inst(0.161)	Indstrl(0.318) Comm(0.258)
			HighRes(0.120)

Recreat	OpenUrb	OpenUrb(0.115)	Inst(0.107) Comm(0.155)
			AgrBldg(0.163) Indstrl(0.119)

Crop	Crop	Crop(0.502)	FeedOp(0.121) Pasture(0.151)
Pasture	Pasture	Pasture(0.314)	Crop(0.314) AgrBldg(0.169)
IdleFld	OpenUrb,Brush		LowRes(0.209) Pasture(0.321)

OrchHrt	OrchHrt		LowRes(0.111) Crop(0.222)
			Pasture(0.177) DeciF(0.177)

Feedlot	FeedOp	FeedOp(0.428)	Indstrl(0.381)

Frmstd	AgrBldg	AgrBldg(0.168)	Pasture(0.122) FeedOp(0.186)
			Indstrl(0.122)

DE LULC	**MATCHED MD LULC**		
	HUMAN	**ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
OthrAgr	Crop,AgrBldg	Crop(0.105)	Pasture(0.105) Indstrl(0.263) FeedOp(0.473)
HerbRng	Pasture,Brush	Pasture(0.281)	
ShrbRng	Brush,Pasture	Brush(0.218)	LowRes(0.231) DeciF(0.258)
MixRng	Brush,Pasture	Brush(0.144)	DeciF(0.152) LowRes(0.289) MedRes(0.115)
DecFrst	DeciF	DeciF(0.608)	Brush(0.102) LowRes(0.133)
EvrgrnF	EvrgrnF	EvrgrnF(0.729)	
MixFrst	MxFrst	MxFrst(0.242)	Brush(0.142) EvrgrnF(0.107) DeciF(0.288) LowRes(0.103)
ClrCut	BrGrnd,Brush	Brush(0.147)	FeedOp(0.107) Pasture(0.203) Crop(0.350)
Watrway	Water	Water(0.869)	
NtrlLk	Water	Water(0.739)	Wetland(0.173)
Rsrvr	Water	Water(0.789)	
BayCove	Water	Water(0.934)	
Wetland	Wetland	Wetland(0.123)	Brush(0.157) DeciF(0.317) MxFrst(0.159)
Beach	Beach	Beach(0.714)	Indstrl(0.285)
InldSnd	BrGrnd		Indstrl(0.6) Beach(0.28)
Extr	Extr		Indstrl(0.606) Comm(0.181)
Trans	Crop,Brush,BrGrnd		FeedOp(0.131) Indstrl(0.356)

b

MD LULC	**MATCHED DE LULC**		
	HUMAN	**ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
LowRes	SinFam	SinFam(0.382)	MixRng(0.150)
MedRes	MultFam,SinFam	SinFam(0.374)	MixRng(0.108)

HighRes	MultFam,MblHm	MultFam(0.313)
		MblHm(0.120)
			SinFam(0.177)

Comm	Retail,OthrCom, MixUrb,	MixUrb(0.157)
		Retail(0.174)	

Indstrl	Indstrl,JunkYrd, Warehs	Indstrl(0.276)	Retail(0.160)
Inst	Inst		SinFam(0.134) MixUrb(0.111)

MD LULC	**MATCHED DE LULC**		
	HUMAN	**ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
Extr	Extr	Extr(0.215)	SinFam(0.169)
OpenUrb	Recreat	Recreat(0.146)	SinFam(0.213)
Crop	Crop,OthrAgr	Crop(0.414)	SinFam(0.131)
Pasture	Pasture	Pasture(0.121)	SinFam(0.154) Crop(0.181)
OrchHrt	OrchHrt	OrchHrt(0.109)	Crop(0.140) Trans(0.171) Pasture(0.125) SinFam(0.140)
FeedOp	Feedlot	Feedlot(0.297)	Crop(0.133) Frmstd(0.126)
AgrBldg	Frmstd	Frmstd(0.103)	Crop(0.162) Feedlot(0.159) Pasture(0.115)
DeciF	DecFrst	DecFrst(0.204)	MixFrst(0.152) ShrbRng(0.162) MixRng(0.179)
EvrgrnF	EvrgrnF	EvrgrnF(0.423)	ShrbRng(0.113) MixFrst(0.136)
MxFrst	MixFrst	MixFrst(0.305)	ShrbRng(0.165) MixRng(0.127)
Brush	ShrbRng,MixRng, IdleFld,ClrCut, HerbRng,Trans	ShrbRng(0.202)	EvrgrnF(0.159)

Water	Watrway,Rsrvr, NtrlLk,BayCove	Rsrvr(0.330)
		BayCove(0.25)	

Wetland	Wetland		MixRng(0.228) Watrway(0.112) Rsrvr(0.110)
Beach	Beach	Beach(0.2)	Retail(0.2) BayCove(0.2) InldSnd(0.2) Indstrl(0.2)
BrGrnd	InldSnd,VclAct		Pasture(0.101) MixRng(0.101)

3.3.2 Evaluation Metrics

Algorithm matches with the similarity less than a threshold of 0.1 are considered to be noises and discarded. Here the threshold has an impact on the matching results. A higher threshold truncates more matches, most of which are non-conforming matches as they have lower similarities. A lower threshold, on the contrary, allows more matches in the table, and returns a more complete matching result. To compare the algorithm matches with human matches under a changing threshold, we introduce a modified precision-recall metric.

Precision-recall evaluation metric is widely used in information retrieval. It is a two-fold metric: the precision is a measure of exactness or fidelity, whereas the recall is a measure of completeness. In the matching of LULC classification systems, the precision follows its original definition as the fraction of conforming matches among all algorithm matches, but recall is a little different. By original definition, recall is the fraction of conforming matches among all human matches. However, the problem is multiple human matches for a category are in an alternative relationship, which means either one (or more than one) of the matches are correct, and the matches are not in an exclusive relationship. For example, according to human evaluation, the matches of DE LULC category MultFam should be MedRes or HighRes in MD LULC, which suggests DE LULC category Multiple Family Residential (MultFam) can either match to Medium Density Residential (MedRes) or to High Density Residential (HighRes). It does not specify Multiple Family Residential should match to one or both of the two. Instead, all three scenarios of matching (MedRes only, HighRes only, and both) are acceptable. Hence, the original definition of recall is not applicable as the number of human matches is not determinative. We modify the original definition of recall to be dividing the number of categories with conforming match(es) by the number of categories with human match(es), and call it conformance rate. Instead of the original definition of recall, the modified recall or the conformance rate metric is used in this study to measure how complete the algorithm compares to human evaluations.

A single measure named F measure will trade off precision versus conformance rate. It is defined as the harmonic mean of conformance rate and precision

$$F = \frac{2PC}{P + C}$$

where *P* is the precision and *C* is the conformance rate (Manning et al. 2008).

As the threshold increases from 0 to 1 at the interval of 0.01, 100 pairs of conformance rate and precision are calculated at each threshold, among which the general trend would be decreasing in conformance rate and increasing in precision, because increasing threshold leads to more exact but less complete integration results. Based on these 100 pairs of conformance rate and precision, a precision-conformance curve will be plotted using the conformance rate as an independent variable on X axis, and the precision as a dependent variable on Y axis. Examining the entire curve is informative yet complicated. The classic simplification is the 11-point interpolated average precision (Manning et. al. 2009), which measures the interpolated precision at the 11 recall levels of 0.0, 0.1, 0.2, . . . , 1.0.

3.3.3 Conformance of Human Evaluations

Before the evaluation of the algorithm, it is necessary to examine to what extent human evaluators agree with each other, because this will provide an upper limit of algorithm's performance.

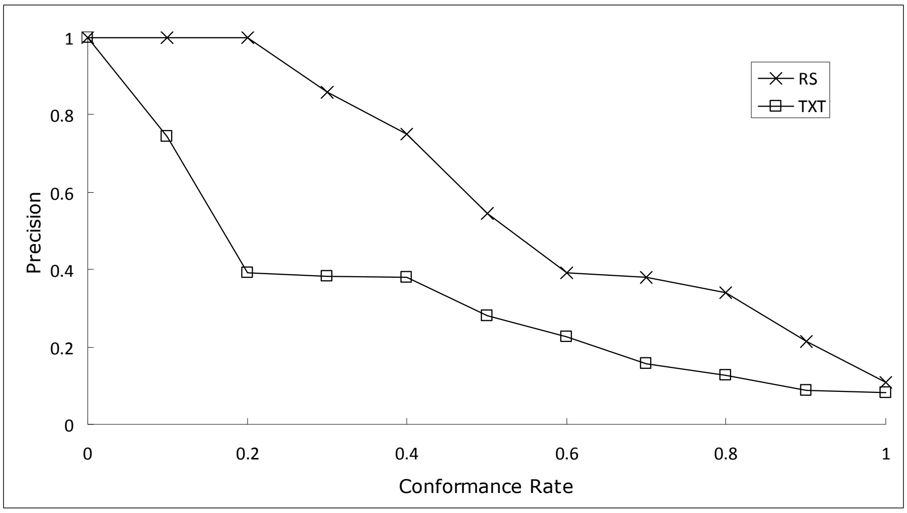
Matches reported by each evaluator are compared with the summary evaluation (matches upon which no less than two evaluators agree), and the F measure of this comparison is calculated. When matching MD LULC to DE LULC, the average conformance rate is 0.98, the average precision is 0.82, and the average F measure is 0.89. If we choose the least agreed evaluation, the conformance rate is 0.91, the precision is 0.66, and the F measure is 0.77. When matching DE LULC to MD LULC, the average conformance rate, precision, and F measure is 0.87, 0.94, and 0.90, respectively. The least agreed evaluation has a conformance rate of 0.72 and precision of 1.0, while its F measure is 0.84.

In the results, the average of conformance rate and precision is fairly high, which suggests evaluators can retrieve similar matches and agree with each other well in most of cases. But the relatively low F measure from the least agreed evaluation suggests matching LULC classifications can be challenging even for experts. This the need to improve the training interpreters.

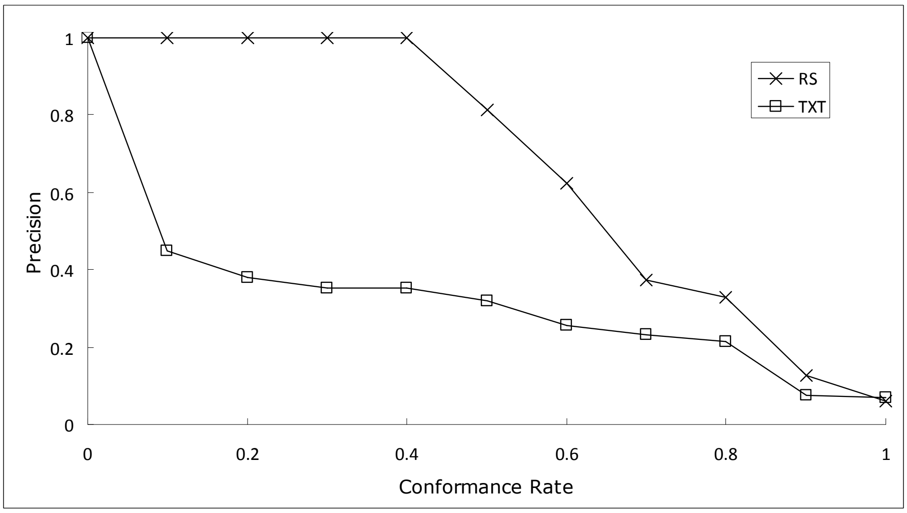
3.3.4 Performance Measurement

An automated lexical-semantics-based integration method applied to geospatial data portals (Zhou and Wei 2008) is applied to the MD and DE LULC data set, and the integration results are presented for comparison purpose. This work measures semantic similarity of two LULC categories by comparing “semantic factors” in the textual descriptions of categories in a feature-based model, where the “semantic factors” are meaningful keywords extracted from category descriptions excluding stop words and negated words.

In Figure 3.4, the precision-conformance curve is plotted using 11-point interpolated precisions. The method introduced in this chapter has a better precision at all 11 conformance rates in both DE to MD and MD to DE experiments. When matching DE LULC to MD LULC, feature based model has the average of the 11 precisions of 0.34 and F measure of 0.16. Using our remote sensing method, the average precision increased to 0.66 and the F measure is 0.50. When matching MD to DE, our remote sensing based method reaches an average precision of 0.61, and F measure of 0.48, while feature-based model has an average precision of 0.35 and F measure of 0.17. As remote sensing is the observation on physical LULC status and the human evaluation is only based on textual definitions, the conforming matches serve as the bridge between observation and semantics. The better performance over previous lexical semantic methods suggests this proposed approach’s feasibility in integrating LULC classification systems. However, comparing to F measure (about 0.90) of matching different human evaluations, remote sensing method is still limited. The next section will look into non-conforming matches and discuss the causes behind them.



a



b

Figure 3.4 Precision-conformance curve of Matching DE LULC to MD LULC (a), and MD LULC to DE LULC (b). RS stands for remote sensing, means the method introduced in this chapter. TXT stands for textual method, means the lexical semantic method in (Zhou and Wei 2008).

3.3.5 Discussion on non-Conforming matches

Inspired by Resnik's work on semantic similarity measurement in taxonomy (Resnik 1999), the degree of the difference between human and algorithm matches can be measured by the level of their minimal common upper category. The more semantically heterogeneous the non-conforming match is, the lower level it is measured. For example, DE LULC category Mobile Home Park is matched to MD LULC category High Density Residential by human and Medium Density Residential by algorithm. This difference is quantified by the level of the minimal common upper category of High Density Residential and Medium Density Residential, which is Residential on level 2.

Using this measurement, the average level of all non-conforming matches in the two way matching is close to 0.8, which means most of the non-conformance happens between level 2 categories. Among all non-conforming matches, the category of Wetland in MD LULC and DE LULC is the most semantically heterogeneous. Human evaluators easily achieved consent on the match of Wetland in MD LULC to Wetland in DE LULC based on their same names and similar descriptions, but many parcel level matches found by algorithm are at level-0 non-conformance: Wetland of DE LULC are similar to Deciduous Forest and Brush of MD LULC, while Wetland MD are matched to Mixed Rangeland and water body in DE. This heterogeneity is discussed in the following section addressing the reasons of non-conformance.

It is worth mentioning that despite the above quantitative measure of the semantic heterogeneity, the correspondence between categories in different areas will have differential impacts depending on the applications. For example, in a LULC classification used by urban planners (e.g. MD LULC), matching open urban land to pasture would be at highest level 0 heterogeneity, but in a regional runoff model, this mismatch is mostly acceptable. Therefore, the evaluation of semantic heterogeneity should also consider the application. But as we cannot predict the future application of integrated LULC classification, we will just stick to the original application of the source LULC classification, based on which aforementioned quantitative measure is developed.

In an extensional approach, the causes of mismatch or non-conformance are found in individual non-conforming parcels. Unfortunately, it is impractical to go through all non-conforming parcels, and find out the reason of its non-conformance. Instead, 10% of non-conforming parcels are selected on a stratified random sampling basis, and examined on original Landsat imagery as well as high resolution remote sensing imagery provided by ESRI World Imagery. The stratified sampling makes a random 10% sampling within the parcels of each non-conforming match, and it ensures every non-conforming match is studied. By examining the Landsat image and high resolution image on selected parcels, five main causes leading to non-conformance are separated. They are 1) conceptual ambiguity, 2) LULC data quality, 3) LULC change, 4) limitation of remote sensing, and 5) procedure error. Among these five causes, the conceptual ambiguity and the LULC data quality is controlled by the producer of the LULC map. Arguably, producer should update the LULC map according to the current LULC change, but it is not always feasible in reality because LULC change may still be happening at the time of LULC mapping. The limitation of remote sensing and the procedure error are on the user side. The limitation of remote sensing leads non-conforming matches because of insufficient discriminating power of the Landsat or remote sensing in general. The procedure error is errors brought by our matching algorithm.

Conceptual ambiguity

As pointed out in the assessment of MD and DE LULC classification systems (section 3.3.2), there are some conceptual ambiguities in the production of the LULC classification systems and maps, and these conceptual ambiguities lead to the latent semantic heterogeneities (both confounding conflicts and naming conflicts) that are embodied by the non-conforming matches. The only way to discover these non-conforming matches and underlying semantic heterogeneities would be using remote sensing. This is the goal of our research in this chapter, and makes our remote-sensing-based method indispensable.

In section 3.2.2, we found out that the use of mixed and "catch-all" categories is inevitable to reflect the complexity of LULC in reality and to keep the inner logic of classification systems. For example, Mixed Forest is used to categorize forest lands with intermixture of evergreen and deciduous forest species, and Other Urban and Built-up Land functions as the safe net to catch all urban land parcels belong to none of the pre-listed subcategories. However, allowing intermixture within a category is against the principle of homogeneity and will increase the possibility of non-conformance. As a result, for a mixed category, we are not surprised to see the individual components of the intermixture are matched together. For example, Mixed Forest DE is matched to Evergreen Forest MD and Deciduous Forest MD; Mixed Urban and Built-up Land DE is matched to Commercial MD, Industrial MD, and

Institutional MD. Although considered non-conforming, these matches are helpful to understand the constituents of the mixed and “catch-all” categories in comparison.

The conceptual ambiguity can happen in under-defined LULC categories, which will ultimately cause non-conforming matches. When matching DE LULC to MD LULC, the foremost noticeable under-defined category is the level 1 category Wetland. As a notoriously complicated LULC type, wetland parcels can be dominated by woody vegetation (forested wetland), or dominated by wetland herbaceous vegetation or non-vegetated at all. Obviously, they are very different LULC subtypes. DE LULC adopts the definition of Wetland in Anderson system (Anderson 1976), which defines wetlands as “the areas where the water table is at, near, or above the land surface for a significant part of most years.” It is also pointed out in Anderson system that “wetlands frequently are associated with topographic lows”. MD LULC defines Wetland briefly as “forested or non-forested wetlands, including tidal flats, tidal and non-tidal marshes, and upland swamps and wet areas.” From their definition, Wetlands in MD and DE both contain forested and non-forested wetlands and therefore should be considered similar. But are these two Wetlands also similar on the ground?

Algorithm matching result suggests different: Wetland of DE LULC are more similar to Deciduous/Mixed Forest and Brush of MD LULC, while Wetland MD are matched to Mixed Rangeland and water body in DE. It seems that DE Wetland is much woodier than MD Wetland. This finding is confirmed by average seasonable NDVI changes of Wetland parcels in MD and DE (Figure 3.1).

Now a closer examination on the actual LULC status of individual Wetland parcels is needed to provide insight into the semantic heterogeneity of Wetlands in DE LULC and MD LULC. Appendix II cell 1 presents a Wetland parcel in DE, which matched to Deciduous Forest MD. From the high resolution remotely sensed image, it is clear that the parcel is forested and in the middle of a larger Deciduous Forest parcel, where the vegetation cover shows no variation crossing the boundary. Actually, according to the definition of Forested Wetland in the Anderson system, the vegetation cover is not a decisive discriminating characteristic between forested wetland and forestland. Instead, wetlands are frequently associated with topographic lows and therefore have a "water table at, near, or above the land surface for a significant part of most years", which we wonder could be detected and mapped by the use of seasonal imagery. To this end, the temporal NDVI curves of the Wetland parcel in Appendix II cell 1 and its adjacent Deciduous Forest parcel should be different and separable. But as plotted in Figure 3.5, the two NDVI curves are too similar to separate. Furthermore, this is not an isolated case, 32% of all Wetland parcels in DE matched to Deciduous Forest MD, 16% to Mixed Forest MD, and another 16% to Brush MD. This non-conformance between definition and reality is either because not all topographically low forested lands are inundated with water and become wetlands, or the inundation below the canopy cannot be seen in remotely sensed images.

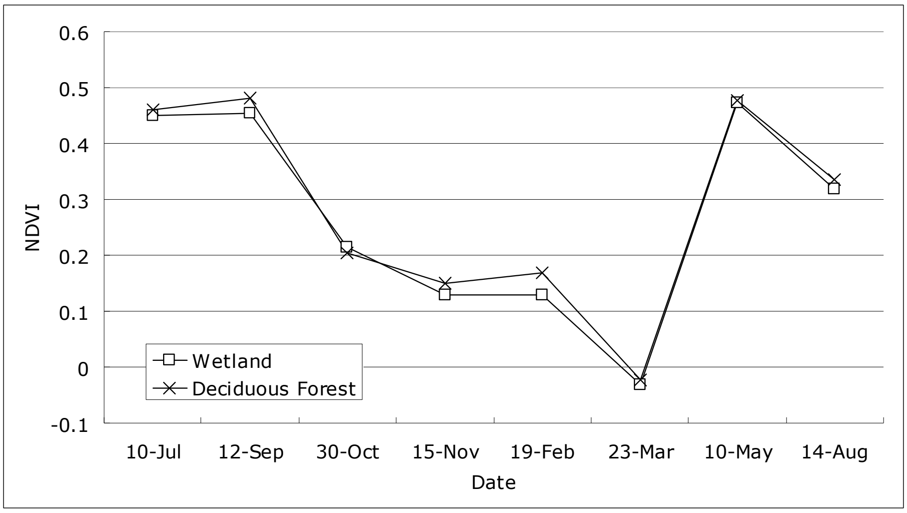


Figure 3.5 Seasonal NDVI of adjacent Wetland and Deciduous Forest parcels

Whichever the reason on the ground, it is certainly difficult to separate forested wetlands from general forestlands in remotely sensed images. Also, forested wetland parcels usually have features of the forestland, such as high percentage of canopy coverage, which make it legitimate to label them as forest. However, despite the practical difficulty and an easy “workaround”, 24.4% of the total parcels in DE LULC map are labeled as Wetland, much higher than 9.4% in MD LULC map. Why does the DE authority bother to label so many Forested Wetland parcels in their LULC map?

Their motivation is explained in the Anderson classification system used by the DE authority: “the wet condition is of much interest to land managers and planning groups and is so important as an environmental surrogate and control, such lands are classified as Forested Wetland.” In simple words, parcels featuring both Forest and Wetland are labeled as Wetland because Wetland is of more environmental importance. Here this strategy of labeling parcels featuring multiple LULC types, termed the *multiple labeling* strategy, will become a cause of non-conformance. In MD LULC, Wetlands are not empowered the same priority over Forest Lands as in DE LULC. This difference in multiple labeling strategy leads to very much different Wetland parcels in reality, and then the confounding conflict in the concept of Wetland between MD LULC and DE LULC. Again, this semantic heterogeneity cannot be discovered by lexical semantic methods.

Since most Wetland parcels in DE are too woody to match to MD Wetland, in the other way around it is difficult for Wetland MD to find a right match in DE LULC, because Wetland parcels in DE are treated as a whole to train the SVM classifier, which means only features of the majority (as forested wetland) count. Hence, Wetland MD is constrainedly matched to Mixed Rangeland (Appendix II cell 2) and Water (cell 3) in DE LULC based on the ratio of vegetation and water in the parcel. This result does not necessarily indicate a high similarity between Wetland MD or Mixed Rangeland or Water DE. Instead, it is only because, on the parcel level, the SVM classifier always gives out matching even the similarity is low.

This mechanism of the SVM classifier leads to another question, namely how to identify LULC types that simply do not match? Let us consider a LULC category in one area that has no match to any category in another. If it has heterogeneous parcels, although SVM has to match each of its parcel to a category, then parcels will be dispersed to multiple different categories, while each match has only a few parcels and therefore a low similarity, and hence can be easily filtered using a slightly higher threshold. But if the category is homogeneous, most its parcels will be forced to match to one or two categories. In this case, these inappropriate matches, with potentially decent similarity values, are highly likely to be noisy and decrease the conformance rate and precision. To fix this problem, we could use classifiers that actually give the similarity when matching a parcel to each target category, and stop matching if none of the similarity values is high enough. A potential candidate of this classifier is maximum likelihood classifier, recommended in Song (2010). Clearly more work is needed to establish robust procedures to identify non-conforming classes between regions.

Multiple labeling also explains several non-conforming matches involving both natural and urban LULC types. In some cases, human activities change the land cover so much that the features of the original land cover type are permanently removed (e.g. commercial or industrial). But in other cases, features of the original land cover may survive some extent of human employment. In these cases, although land parcel has the features of both urban and its original natural LULC; it is always labeled as the urban LULC category. For example, many Single Family Residential land parcels in MD developed on rangelands still keep a feature of rangelands. In the parcel shown in Appendix II 4, the development in the parcel happens along the parcel boundary, and most of the rangeland area remains intact. Algorithm matches such parcels to Mixed Rangeland in DE LULC judging its primary land cover status, which is considered to be a mismatch by human evaluation. Multiple labeling is also found in non-conforming matches such as Farmstead DE to Pasture MD (Appendix II 5), and Agricultural Building MD to Pasture DE (Appendix II 6), because agricultural structures are often built on and surrounded by agricultural lands, including croplands or pasture lands.

For similar LULC categories, deciding the semantic relationship (e.g. which one is a broader concept) between them can be difficult through the interpretation of textual descriptions, because it is hard to quantify semantics and convert them into comparable attributes (section 3.1). Approaching the problem differently, the extensional method employed in two way matching provides estimation on the semantic relationship via the set inclusion and intersection between some LULC categories. For example, Table 3.3, extracted from Table 3.2, presents the two-way matching for Deciduous Forest and Evergreen Forest. When matching is from DE LULC to MD LULC, Evergreen Forest DE matched to Evergreen Forest MD entirely (except less similar (<10%) matches). When matching is from MD LULC to DE LULC, only part of Evergreen Forest MD matched to Evergreen Forest DE, while the rest matched to Shrub Rangeland DE or Mixed Forest DE (Figure 3.6). Combining these two pieces of information, we can make estimation that Evergreen Forest DE is subsumed by Evergreen Forest MD – it has a narrower conceptual scope. The categories of Deciduous Forest are a little complicated, as neither the DE category nor the MD category is clearly subsumed by the other, but instead, they are overlapped (Figure 3.7). But still, as much higher fraction of Deciduous Forest DE parcels are matched to Deciduous Forest MD than the other way around, Deciduous Forest in MD LULC should have a broader scope.

Table 3.3 Results of Matching Deciduous Forest and Evergreen Forest

a

DE LULC	**MATCHED MD LULC**		
	HUMAN	**ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
DecFrst	DeciF	DeciF(0.608)	Brush(0.102) LowRes(0.133)
EvrgrnF	EvrgrnF	EvrgrnF(0.729)	

b

MD LULC	**MATCHED DE LULC**		
	HUMAN	**ALGORITHM**	
		CONFORMING	**NON-CONFORMING**

DeciF	DecFrst	DecFrst(0.204)	MixFrst(0.152)
			ShrbRng(0.162)
			MixRng(0.179)

EvrgrnF	EvrgrnF	EvrgrnF(0.423)	ShrbRng(0.113)
			MixFrst(0.136)

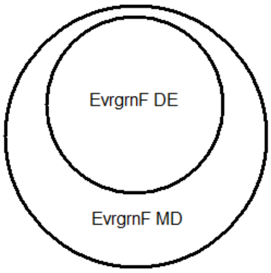


Figure 3.6 Set Inclusion of Evergreen Forest

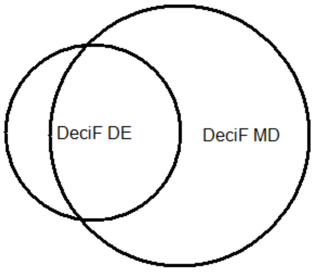


Figure 3.7 Set Intersection of Deciduous Forest

Data Quality

After separating conceptual ambiguities, the next cause of non-conforming matches relates to the data quality of LULC maps. Ideally, a concept is represented by the commonality of all its instances, and every instance exemplifies the *intension* of the concept. However, in a real world, errors are pervasive: they will happen on the conceptual level as conceptual ambiguities, and they will happen in instances as data errors and will undermine the representative power of instances as an important means to understand the concept. If we consider a LULC category as a concept, then a land parcel is an instance of this concept, and therefore each parcel is expected to comply with the categorical definition. However, as errors in the delineation and labeling of land parcels are inevitable, non-conforming matches resulted from these errors are also inevitable.

A common data quality issue reflected in LULC mapping roots in mistakes during the delineation process. Mistakes in locating the boundary of a parcel can change its LULC status and lead to non-conforming matches. For example, as shown in Appendix II cell 7, the water body that naturally located in the center of a wetland parcel is not excluded during delineation. These wetland parcels in MD then have a strong feature of water in their LULC status, and are matched to Waterway or Reservoir of DE LULC.

In some cases, when discerning the boundary of land parcels, producers of the LULC map did not exclude artificial structures from land parcels that by definition should be "structure-free" and caused mismatches. For example, a Pasture parcel in Maryland is shown in Appendix II cell 8. This parcel is labeled as Pasture, which is the primary LULC type of the parcel. However, several buildings are included in the parcel, which make it more similar to Single Family Dwellings DE. This parcel may also represent an omission of the multiple labeling strategy, under which it should be labeled by the use of the building rather than the pasture land occupying the main part of it. There is a procedure error rooted in the SVM classifier that should also be held accountable for this non-conforming match, and we will discuss it in the following paragraphs on procedure errors.

A common type of data errors in the parcel boundary delineation is caused by map displacement. Appendix II cell 9 and 10 shows a displaced Deciduous Forest parcel in MD matched to Mixed Rangeland DE, because part of the parcel is in the neighboring rangeland area. Eliminating mixed/transitional pixels on parcel boundaries always helps algorithm tolerate errors due to the displacement. But for a rather displaced small parcel, such as the one in Appendix II cell 10, the elimination does not work correctly. Pixels eliminated on the edge are not the actual transitional pixels, while the actual transitional pixels are kept and lead to non-conforming matches. This displacement is not due to mis-registration (geometric errors should be corrected in the preprocessing phase), but attributes to sporadic producer errors.

Based on scrutinizing sampling data (10% of all non-conforming parcels), the displacement leads to limited number (around 5%) of non-conforming parcels in random locations. For all parcels in the map, this percentage should be lower because data quality issues happen more frequently on non-conforming parcels than conforming parcels.

LULC Change

LULC change detection is one of the earliest remote sensing applications. LULC change within a land parcel leads to the difference between the actual LULC status and its labeled category. This difference can be detected by the algorithm, and leads to non-conforming matches. LULC change can happen to any LULC type, but some types, such as Transitional, Bare ground, or Clear-cut, are especially vulnerable. For example, in Appendix II cell 11, newly grown crops change the LULC status of a Clear-cut parcel in DE to cropland. The algorithm detected this change and matched it to Cropland MD. This parcel is an example of deforestation and agricultural expansion. On the contrary, vegetation growth on unmanaged lands leads to LULC change and non-conforming matches such as Bare Ground MD to Mixed Rangeland DE (Appendix II cell 12), Brush MD to Evergreen Forest DE (cell 13), Shrub Rangeland DE to Deciduous Forest MD (cell 14), and Mixed Rangeland DE to Deciduous Forest (cell 15).

Deforestation and urbanization, another type of widely observed LULC changes, are also found in our study area. In Appendix II cell 16, the canopy cover of the Mixed Forest parcel greatly decreases due to development, and finally the parcel is disqualified from forestland. This parcel is matched to Mixed Rangeland instead.

Limitation of Remote Sensing

A substantial amount (42% of all non-conforming parcels in matching DE LULC to MD LULC, 34% in matching MD LULC to DE LULC) of non-conforming parcels are due to the limitation of remote sensing's ability of deciding the relationship between observed land cover status and its actual land use. Land cover and land use are different concepts. Land cover emphasizes the physical cover of the land (Turner et al. 1995), while land use emphasizes the human employment of land resources (Vink 1975, FAO 1995). Remote sensing observes land cover, and human activities on lands are interpreted based on this observation.

In LULC mapping, land cover functions as a "surrogate" (Anderson 1976) standing between remote sensing and land use. However, this surrogating relationship is not always decisive for all land cover and land use types. Even largely reduced by employing time series imagery and textural data, the uncertainty persists in this surrogating relationship, and will lead to non-conforming matches in the matching result, especially those with high level non-conformance.

This limitation of remote sensing is especially common when separating urban land use categories. After all, remote sensing cannot tell the use of individual buildings and structures directly. For example, a DE parcel used for Retail, as shown in Appendix II cell 17, is matched to MD LULC category High Density Residential. In urban areas, remote sensing, even at the highest spatial resolution, is not capable of mapping land uses accurately.

It is possible this limitation can attribute to the insufficient spatial resolution of Landsat ETM+. This argument brings up several preventives that have been taken to alleviate the influence and make up the insufficiency. First, mixed pixels on parcel boundaries are removed. Second, image time series is incorporated as an important measurement of vegetation phenology. Third, textural information is also incorporated, which is especially helpful in separating LULC categories with similar color in remote sensing images but different patterns, such as agricultural lands (Appendix cell 18) and rangelands (cell 19).

It seems the limitation of spatial resolution can be easily fixed by incorporating higher resolution data. But in practice it is actually much more complicated, because high resolution remote sensing has issues in data coverage and availability, and sacrifices temporal and spectral abundance.

Another approach to break this limitation goes beyond remote sensing. For example, to separate agricultural lands, such as Orchards/Nurseries/Horticulture, Pasture, and Cropland, remote sensing alone is not enough. GIS data about soil types, topography, and local climatology need to be incorporated. By combining specific GIS data such as building information with remote sensing, Wu et al. (2009) successfully separate urban land uses in Austin, Texas. In addition to GIS data, lexical semantics also hold the key to solving this problem. For example, while it is very difficult to separate Feeding Lots from Farmsteads in remote sensing, the most naïve natural language processing (e.g. string comparison) is adequate to tell the difference and match them correctly. The following chapter has the discussion on combining lexical semantics with remote sensing.

Procedure Error

Procedure error is the algorithm error rooted in the matching process and irrelevant to conceptual ambiguity, data quality, LULC change or the limitation of remote sensing. As our method is completely automated, there is no room for uncertainties and human factors in the whole algorithm, and this makes the algorithm deterministic. Now it may seem a little paradoxical to discuss procedure error, because if there are procedure errors in the algorithm, why the errors are not fixed in the first place? The answer is procedure errors are the inevitable byproduct resulted from the core functions of the algorithm. There are two types of procedure errors isolated in the algorithm, the first one relates to the side effects of the SVM classifier and the second one concerns the removal of pixels on parcel boundaries.

In the matching results, an interesting finding is the categories of Low Density Residential MD and Single Family Dwellings DE seem to be two popular choices in matching, even for categories that seem semantically and spectrally different from the two. When matching MD LULC to DE LULC, 7 various categories, ranging from High Density Residential to Pasture, have more than 10% of all their parcels matched to Single Family Dwellings of DE LULC. Also when matching DE LULC to MD LULC, 6 various categories, such as Mobile Home Parks and Mixed Forest, matched to Low Density Residential MD with a similarity of more than 10%.

This interesting finding cannot be explained by causes discussed before. It concerns the mechanism of support vector machine (SVM) classifier (Cortes and Vapnik 1995) used in the algorithm (section 3.3.7). Despite the complicated mathematics behind SVM, its main idea is quite straightforward: finding a hyperplane to separate the instances into two classes, which can maximize the margin between the two separated classes. In his comprehensive study of various classifiers, Kuan (2010) pointed out that SVM classifier generally over-performs other classifiers in terms of accuracy and error tolerance. This is the reason why SVM is chosen to be the classifier in this study.

However, this effective design of SVM also leads to an unwanted side-effect. As Kuan (2010) pointed out, “when a class is given more training data, the hyperplanes around this class will be pushed outwards, eroding other classes (Kuan 2010).” This eroded class therefore attracts more matches. The eroded class, Single Family Residential, constitutes 12% of all parcels in DE LULC map, and Low Density Dwellings constitutes 19% of all parcels in MD. In addition to the quantitative advantage, these two categories also have high priority in multiple labeling, which means a variety of original land covers, such as Rangeland, Forest, or Wetland, will be observed in these categories. The variety of land cover status means a large deviation and therefore much erosion in the spectral space, and leads to many non-conforming matches to Low Density Dwellings MD and Single Family Residential DE. Kuan (2010) discussed the influence of class proportions in the training set on the classification result, but an almighty optimization strategy is yet to come.

The second type of procedure errors relates to the removal of pixels on parcel boundaries. As discussed before, this removal is necessary and beneficial to increasing the accuracy. However, for a parcel in a long and narrow shape (Appendix II cell 20), pixels to be removed on the parcel boundary compose a large portion of the whole parcel. Therefore, losing a substantial amount of its pixels, the Deciduous

Forest parcel (Appendix II cell 21) becomes unidentifiable and matched to Low Density Residential incorrectly. An easy fix to this error is to only consider parcels large enough and in relatively regular shapes. Further discussion of this remedy can be found in chapter 5.

Table 3.4 Percentage of non-conforming parcels attribute to different causes

	Conceptual Ambiguity	Data Quality	LULC Change	Limitation of RS	Procedure Error
DEtoMD	45.4%	0.1%	4.2%	41.8%	8.5%
MDtoDE	35.7%	15.1%	10.8%	34.3%	4.1%

Calculated from 10% sampling of all non-conforming parcels, the percentage of parcels attribute to each of the five causes are shown in Table 3.4. In the table, we can see Conceptual Ambiguity and Limitation of Remote Sensing are two main causes that are responsible for 87.2% of non-conforming parcels in the matching from DE LULC to MD LULC, and 70% from MD LULC to DE LULC. DE LULC map is better representing the actual LULC status of the time than MD LULC map, because only 0.1% non-conforming matches attribute to data quality issues of the DE LULC map and 4.2% non-conforming matches are caused by unrecorded LULC changes, while these two numbers are 15.1% and 10.8% for MD LULC. Procedure errors are responsible for only a few cases of non-conformance in this two-way matching.

The isolation of different causes of non-conformance is necessary. Conceptual ambiguity is on the conceptual level, where the latent semantic heterogeneity occurs and goes all the way down to non-conforming matches. Having conceptual ambiguities isolated, data quality is the issue that the semantics or the *intension* of concept is not represented by LULC data correctly, due to mistakes in producing maps and/or LULC change. These are the causes of non-conformance on the producer side. On the user side, non-conformance is caused by the limitation of remote sensing (or particularly Landsat ETM+) and procedure errors.

After different causes separated, non-conformance on the producer side or the user side should be treated differently. On one hand, finding the non-conformance on the producer side is the purpose of remote-sensing-based algorithm. Conceptual ambiguities reflect the difference between how a LULC concept is defined in text and how it is used on the ground, and provide a valuable insight needed in the semantic integration of LULC classification systems. Finding issues on data quality and LULC change are more helpful to the ultimate goal – geospatial data interoperability. On the other hand, non-conformance on user side needs to be eliminated. Further discussion on this is in the following chapter 5.

3.4 Conclusions

In this chapter, we presented the method of applying remote sensing to the integration of LULC classification systems. Remote sensing is a means of observation on actual LULC status, and it observes individual parcels. We therefore calculated parcel level statistics from spectral and textural data, and imported these statistics of parcels from different areas in a SVM classifier as training and testing respectively. Then an extensional similarity measurement is adopted to calculate category similarity from parcel level matches, and the matching categories are compared with human evaluations, which are based on names and descriptions.

The matching results showed this remote sensing based approach largely improved performance over the previous lexical semantic method (Zhou and Wei 2008): the average of 11-point precision has improved from 0.34 to 0.66 when matching DE LULC to MD LULC, and from 0.35 to 0.61 when matching MD LULC to DE LULC.

More importantly, remote sensing based method discovered and reduced semantic heterogeneities between LULC descriptions. Based on discovering confounding conflicts, the method estimated the semantic relation between LULC categories. Since the method compares LULC categories by their actual LULC status via remote sensing rather than potential confusing names, it naturally has the capability to reconcile naming conflicts. With semantic heterogeneities discovered and reduced, remote sensing here served as the translation between semantic heterogeneous LULC classification systems, and hence enabled LULC data interoperability, which is the foundation of regional LULC dynamics analysis.

From examining mismatched parcels on high resolution imagery, five causes leading to non-conformance between algorithm and human evaluation are separated. They are 1) conceptual ambiguity, 2) LULC data quality, 3) LULC change, 4) limitation of remote sensing, and 5) procedure error. Among these five causes, conceptual ambiguity, source of semantic heterogeneities, is responsible for the majority (45% and 36%) of all non-conforming parcels in the experiments of matching DE to MD and MD to DE.

Conceptual ambiguity, LULC data quality, and LULC change are controlled by the producers of the LULC map, and the limitation of remote sensing and the procedure error are on the user side. Producer side and user side non-conformance should be separated and treated differently. Discovering producer side non-conformance (mainly caused by semantic heterogeneities in LULC classifications and defects in LULC data) is one important purpose of remote sensing based matching algorithm. Reducing the impact of user side non-conformance helps producer side non-conformance stand out. The method of reducing user side non-conformance will be discussed in chapter 5.

In this chapter, our method is illustrated by matching MD LULC and DE LULC. In the next chapter, we will test the method's geographical generality by incorporating LULC classification system from New Jersey area.

Chapter 4: Generality of Integration Based on Remote Sensing

4.1 Introduction

In chapter 3, we proposed using remote sensing to overcome semantic heterogeneities between different LULC classification systems. The method is demonstrated by matching MD LULC and DE LULC, and the matching results showed our remote sensing based approach successfully matched semantic heterogeneous LULC classification systems used in MD and DE.

But in order for this remote sensing based matching to be meaningful in LULC study, one successful example of integration is not enough. Instead, we need to know if the matching method based on remote sensing can be applied to other classification systems and how well it will perform. Namely, does it have the geographical generality needed in LULC study?

In terms of methodology, there is not a single step in the algorithm that is location specific or related to geographical limitation. Instead, the geographical generality of a remote sensing based matching method will depend on the availability and applicability of remote sensing data. If remote sensing is able to translate between the LULC classification definitions, matching based on remote sensing can be applied to LULC information from even remote areas, where lexical semantic methods is powerless due to the linguistic barrier. Landsat data, as the data of choice of the method in chapter 3, has a global coverage, and it is designed to provide observation on Earth surface, especially LULC. Hence, the data availability and applicability would not limit the geographic generality of this method.

LULC depends on many factors, such as soil type, climate, topography, hydrology, et al., and it varies geographically. If the LULC classifications in two places are greatly heterogeneous due to one or several of these factors rather than mere naming conflicts, we could expect the matching result contain many confounding conflicts and hence become less conforming to human evaluations that are based on textual descriptions.

Based on the discussion above, we make a hypothesis that the matching method based on remote sensing has the geographical generality and is able to produce integration results conforming to human evaluation, if two areas have similar geographical factors. In this chapter, we will test this hypothesis by adding a new LULC classification system used by the State of New Jersey (NJ LULC) (Figure 4.1, Appendix I) to the testing datasets, and compare it with MD LULC and DE LULC. As adjacent areas, NJ LULC has similar soil type, climate, hydrology, and topography as MD LULC and DE LULC.

4.2 Study area and data

In Figure 4.2, we can see the southern New Jersey is covered by the Landsat scene path 14 row 33, the same scene used in matching DE LULC and MD LULC. Although the matching algorithm is not location specific, it does require the study area is covered by same or similar sensors. Choosing this adjacent New Jersey area enables the reuse of previous Landsat images and textural data (Table 4.1).

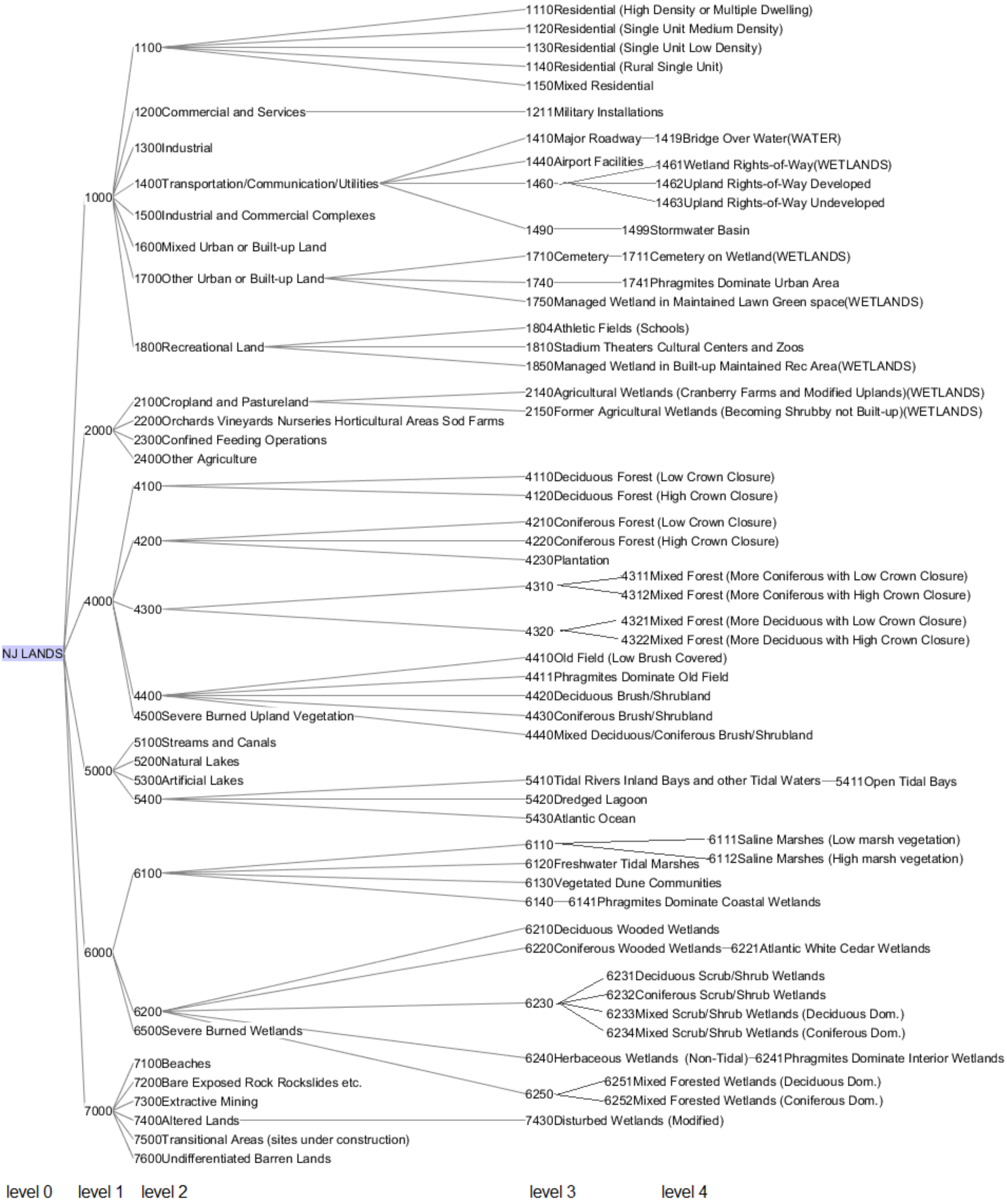

NJ LANDS
1000
1100
1110Residential (High Density or Multiple Dwelling)
1120Residential (Single Unit Medium Density)
1130Residential (Single Unit Low Density)
1140Residential (Rural Single Unit)
1150Mixed Residential
1200Commercial and Services
1211Military Installations
1300Industrial
1400Transportation/Communication/Utilities
1410Major Roadway
1419Bridge Over Water(WATER)
1440Airport Facilities
1460
1461Wetland Rights-of-Way(WETLANDS)
1462Upland Rights-of-Way Developed
1463Upland Rights-of-Way Undeveloped
1490
1499Stormwater Basin
1500Industrial and Commercial Complexes
1600Mixed Urban or Built-up Land
1700Other Urban or Built-up Land
1710Cemetery
1711Cemetery on Wetland(WETLANDS)
1740
1741Phragmites Dominate Urban Area
1750Managed Wetland in Maintained Lawn Green space(WETLANDS)
1800Recreational Land
1804Athletic Fields (Schools)
1810Stadium Theaters Cultural Centers and Zoos
1850Managed Wetland in Built-up Maintained Rec Area(WETLANDS)
2000
2100Cropland and Pastureland
2140Agricultural Wetlands (Cranberry Farms and Modified Uplands)(WETLANDS)
2150Former Agricultural Wetlands (Becoming Shrubby not Built-up)(WETLANDS)
2200Orchards Vineyards Nurseries Horticultural Areas Sod Farms
2300Confined Feeding Operations
2400Other Agriculture
4000
4100
4110Deciduous Forest (Low Crown Closure)
4120Deciduous Forest (High Crown Closure)
4200
4210Coniferous Forest (Low Crown Closure)
4220Coniferous Forest (High Crown Closure)
4230Plantation
4300
4310
4311Mixed Forest (More Coniferous with Low Crown Closure)
4312Mixed Forest (More Coniferous with High Crown Closure)
4320
4321Mixed Forest (More Deciduous with Low Crown Closure)
4322Mixed Forest (More Deciduous with High Crown Closure)
4400
4410Old Field (Low Brush Covered)
4411Phragmites Dominate Old Field
4420Deciduous Brush/Shrubland
4430Coniferous Brush/Shrubland
4440Mixed Deciduous/Coniferous Brush/Shrubland
4500Severe Burned Upland Vegetation
5000
5100Streams and Canals
5200Natural Lakes
5300Artificial Lakes
5400
5410Tidal Rivers Inland Bays and other Tidal Waters
5411Open Tidal Bays
5420Dredged Lagoon
5430Atlantic Ocean
6000
6100
6110
6111Saline Marshes (Low marsh vegetation)
6112Saline Marshes (High marsh vegetation)
6120Freshwater Tidal Marshes
6130Vegetated Dune Communities
6140
6141Phragmites Dominate Coastal Wetlands
6200
6210Deciduous Wooded Wetlands
6220Coniferous Wooded Wetlands
6221Atlantic White Cedar Wetlands
6230
6231Deciduous Scrub/Shrub Wetlands
6232Coniferous Scrub/Shrub Wetlands
6233Mixed Scrub/Shrub Wetlands (Deciduous Dom.)
6234Mixed Scrub/Shrub Wetlands (Coniferous Dom.)
6240Herbaceous Wetlands (Non-Tidal)
6241Phragmites Dominate Interior Wetlands
6250
6251Mixed Forested Wetlands (Deciduous Dom.)
6252Mixed Forested Wetlands (Coniferous Dom.)
6500Severe Burned Wetlands
7000
7100Beaches
7200Bare Exposed Rock Rockslides etc.
7300Extractive Mining
7400Altered Lands
7430Disturbed Wetlands (Modified)
7500Transitional Areas (sites under construction)
7600Undifferentiated Barren Lands
level 0
level 1
level 2
level 3
level 4

Figure 4.1 Taxonomy of NJ LULC classification system

The originator of NJ LULC is the New Jersey Department of Environmental Protection (NJDEP), Office of Information Resources Management (OIRM), Bureau of Geographic Information Systems (BGIS). Using a modified Anderson classification system, "the 2002 NJ LULC data was created by comparing the 1995/97 LULC map from NJDEP's geographical information systems database to 2002 color infrared (CIR) imagery and delineating areas of change (http://www.state.nj.us/dep/gis/digidownload/metadata/lulc02/w01lu02.htm)."

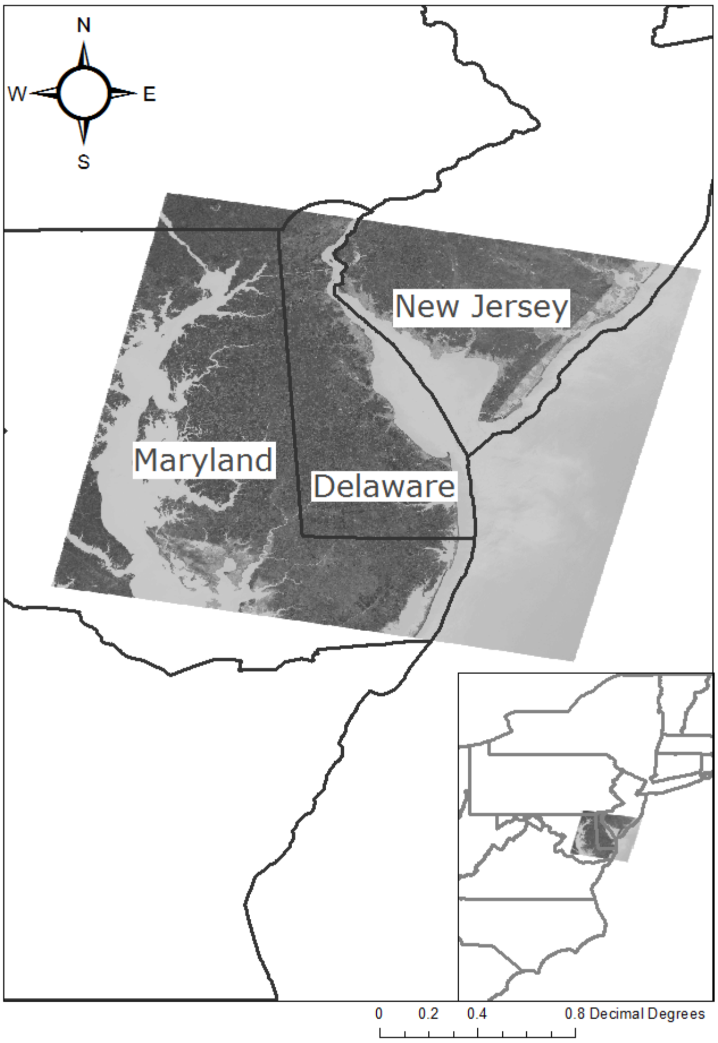


Figure 4.2 Study area: NJ covered by Landsat scene path 14; row 33

4.3 Assessment of the NJ LULC classification system

In the taxonomy presented in Figure 4.1, the categories with meaningful names are used in the LULC map, while those denoted only by numeric codes are not used in the LULC map but as space holders in the taxonomy. Obviously, these space holders represent meaningful LULC concepts. A complete description of all categories of NJ LULC, either used in LULC map or just as space holders in the taxonomy can be found on the official website of the State of New Jersey: http://www.state.nj.us/dep/gis/digidownload/metadata/lulc02/anderson2002.html.

An obvious difference between NJ LULC and DE or MD LULC is that NJ LULC is more finely divided than MD and DE LULC. About half of the NJ LULC categories actually used in map are on level 3; level 2 and level 4 categories compose the other half. But in MD LULC and DE LULC, most categories in use are on level 2, along with several level 3 categories. Wetland, as the level 1 category in both MD LULC and DE LULC classification systems, is further divided in NJ LULC.

As introduced in section 3.3.2, an assessment based on Spiteri's simplified model for facet analysis is performed to locate the potential conceptual ambiguities and errors in NJ LULC classification system. Same principles in section 3.3.2 are tested on the categories of NJ LULC.

PRINCIPLES FOR THE IDEA PLANE

Principles for Choice of Facets

a) Principle of Differentiation

This principle advises that categories in a classification system should be clearly separated, and therefore each land parcel belongs to a specific category with total certainty. NJ LULC categories (including mixed categories) follow this principle well. Theoretically, there will be no difficulties to assign the appropriate category of NJ LULC to each parcel, as long as the parcel is described thoroughly. However, since NJ LULC is not designed to use with specific remotely sensed data, a parcel, especially which belongs to high level categories, might not be fully described by Landsat. In these cases, conceptually separable does not mean separable at a specific mapping scale, and therefore the confusion between high level categories may be observed in the matching result.

b) Principle of Relevance

This principle advises that the choice of facets should reflect the purpose of the classification. NJ LULC data is originated by environmental protection authorities to provide information for regulators, planners, and others interested in LULC changes. Extending the Anderson system to level 4, NJ LULC classification system fulfills its purpose well.

c) Principle of Ascertainability

NJ LULC classification system is built upon definite and ascertained facets.

d) Principle of Permanence

It is nearly impossible to find permanent qualities in remote sensing that can be used to categorize the parcels that are undergoing LULC change. As usual, the impact of LULC change in NJ LULC is inevitable to see in the integration results.

e) Principle of Homogeneity and Principle of Mutual Exclusivity

Homogeneity and mutual exclusivity are both relative to the discriminating power. Under a consistent discriminating power, the principles of homogeneity and mutual exclusivity require each category to be homogeneous but mutually exclusive among each other. In NJ LULC, however, using a parent category and its direct subcategories to label the LULC map broke the conceptual homogeneity and mutual exclusivity. For example, category Other Urban or Built-up Land (1700), Cemetery (1710) and Cemetery on Wetland (1711) are all used in NJ LULC map. This confusing categorization broke the homogeneity of category Other Urban or Built-up Land and category Cemetery, and destroyed the mutual exclusivity between each two of the three. Several more similar confusing categorizations are observed in NJ LULC. A logical explanation to this embedded categorization is that a parcel, if assigned to any subcategory, will be excluded from the parent category. However, NJ LULC did not explicitly define this mutual exclusivity. In matching NJ LULC to MD and DE LULC, although logically flawed, this type of confusion will not cause serious non-conformances because it happens to high level categories while categories in MD and DE LULC are mainly on level 2. That is, if a parent category and its subcategories of NJ LULC all match to a same category in MD or DE LULC, their inner mutual exclusivity is not considered any more.

f) Principle of Fundamental Categories

As a modified Anderson classification system, NJ LULC classification system is built on the facets related to the nature of LULC analysis.

Canon of Exhaustiveness

To ensure the exhaustiveness, “catch-all” subcategories, such as Other Urban or Built-up Land, Other Agriculture, and Undifferentiated Barren Land are included in NJ LULC classification systems. Normally, the semantic purity of the “catch-all” categories is very much a challenge. In NJ LULC, however, the intension of these “catch-all” categories is clearly defined by the well-written category descriptions and the fully developed taxonomy.

Principles for Citation Order of Facets and Foci

a) Principle of Relevant Succession

The hierarchy of NJ LULC classification system reflects the natural scopes of each level of categories.

b) Principle of Consistent Succession

The order of facets used in NJ LULC classification system is not modified after established.

The result of the assessment shows, except some logic flaws resulted form overlapping super category and subcategories, NJ LULC in general is well categorized and documented. However, as aforementioned, NJ LULC classification system is not designed to use with specific remotely sensed data, and conceptually separable high level categories may not still be separable in Landsat images, and non-conformance due to this inseparability will show in matching results. Furthermore, it is important to point out that since the matching process involves at least two LULC classification systems, reliable integration is not only based on the soundness of classification on the source side (NJ LULC), but also on the target side (MD LULC and DE LULC). All confusions in MD and DE LULC definitions (see section 3.3.2) will also have an impact on the integration.

4.4 Matching Results

As before, the human evaluation gives matching categories for each category rather than the similarity value. Human evaluators match LULC classification systems in a one-way fashion: from the system with higher level categories to the system with lower level categories, in this case, from NJ LULC to MD or DE LULC.

The matching result is shown in Table 4.1 (a) NJ LULC to MD LULC and (b) NJ LULC to DE LULC. In both tables, categories are denoted by codes introduced in Appendix I. The columns from left to right mean 1) NJ LULC categories, 2) match(es) in MD (a) or DE (b) LULC by human evaluators, 3) conforming matches (algorithm and human match same), and 4) non-conforming matches (algorithm and human match different). The number in the parentheses after each algorithm result (in column 3 and 4) is the similarity of that match, calculated as the ratio of the number of parcels in this match to the number of all parcels in the source category.

Table 4.1 Results of Matching NJ LULC to MD LULC (a), and NJ LULC to DE LULC (b).

a

NJ LULC	**MATCHED MD LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
HighRes	HighRes	HighRes(0.511)	Indstrl(0.116) Comm(0.186)
MedRes	MedRes	MedRes(0.317)	Comm(0.139) HighRes(0.211)
LowRes	LowRes	LowRes(0.190)	Brush(0.143) MedRes(0.270)
RurlRes	LowRes	LowRes(0.192)	MedRes(0.157) Brush(0.163)
MixRes	LowRes MedRes HighRes		Wetland(0.5) OpenUrb(0.5)
Comm	Comm	Comm(0.361)	HighRes(0.296) Indstrl(0.229)
Milit	Inst		Indstrl(0.636) HighRes(0.181)
Indstrl	Indstrl	Indstrl(0.577)	HighRes(0.118) Comm(0.211)
Transp			Indstrl(0.275) Wetland(0.120) HighRes(0.147) Comm(0.178)
Road			Comm(0.193) HighRes(0.387) Wetland(0.129) MedRes(0.193)
Bridge			Water(1.0)
Airport			HighRes(0.189) Indstrl(0.405)
WtlndWa	Brush Wetland	Brush(0.370) Wetland(0.191)	DeciF(0.131) LowRes(0.101)
UpldWaD			OpenUrb(0.117) Inst(0.117) Comm(0.176) LowRes(0.117) HighRes(0.176)
UpldWa	Brush	Brush(0.310)	Pasture(0.132) LowRes(0.152)
StrmBas			Indstrl(0.181) HighRes(0.163) Comm(0.228)
ICCmplx	Indstrl Comm		HighRes(1.0)
MixUrb	ALL URBAN	Comm(0.313) HighRes(0.509)	
OthrUrb	BrGrnd OpenUrb		HighRes(0.166) Indstrl(0.133) Comm(0.207)
Cemet	OpenUrb	OpenUrb(0.103)	Pasture(0.151) AgrBldg(0.103)
WtCemet	OpenUrb Wetland		AgrBldg(0.5) Pasture(0.5)
Phrg			Wetland(0.5) HighRes(0.25) Comm(0.25)
MngWtld	Pasture Wetland	Pasture(0.222)	Comm(0.111) HighRes(0.126)
Recreat	OpenUrb		HighRes(0.109) Comm(0.157) Indstrl(0.220)
Athlet	Inst	Inst(0.107)	Comm(0.123) Indstrl(0.256) FeedOp(0.194)
Stadium	Inst		HighRes(0.263) Indstrl(0.526)
MngWtRe	OpenUrb Wetland	OpenUrb(0.148)	Pasture(0.172) AgrBldg(0.111) Inst(0.111)
CrpPstr	Crop Pasture	Pasture(0.221) Crop(0.169)	FeedOp(0.124)

NJ LULC	MATCHED MD LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
AgriWet	Crop Pasture Wetland	Crop(0.137) Pasture(0.210)	FeedOp(0.110) AgrBldg(0.146)
FmAgrWt	Wetland Brush	Brush(0.163)	Pasture(0.236) LowRes(0.221) DeciF(0.105)
OrchHrt	OrchHrt		AgrBldg(0.138) FeedOp(0.112) Indstrl(0.129) Crop(0.107) Pasture(0.113)
FeedOp	FeedOp		Comm(0.102) Indstrl(0.346)
OthrAgr	FeedOp AgrBldg GdnCrop	AgrBldg(0.124) FeedOp(0.105)	Pasture(0.142) Indstrl(0.186)
DeciF10	DeciF Brush	Brush(0.352)	MedRes(0.102) LowRes(0.121)
DeciF50	DeciF	DeciF(0.175)	Brush(0.383) LowRes(0.134)
ConiF10	EvrgrnF Brush	EvrgrnF(0.340) Brush(0.392)	
ConiF50	EvrgrnF	EvrgrnF(0.675)	Brush(0.126)
Plant	EvrgrnF	EvrgrnF(0.751)	
MxCnF10	MxFrst EvrgrnF Brush	Brush(0.419) EvrgrnF(0.276)	
MxCon50	MxFrst EvrgrnF	EvrgrnF(0.468)	Brush(0.326)
MxDec10	MxFrst DeciF Brush	Brush(0.483)	EvrgrnF(0.170)
MxDec50	MxFrst DeciF		EvrgrnF(0.206) Brush(0.514)
OldFld	Brush	Brush(0.129)	Comm(0.108) Pasture(0.166)
PhrgOld	Brush	Brush(0.103)	MedRes(0.103) Wetland(0.517)
DecBrsh	Brush	Brush(0.254)	LowRes(0.165) DeciF(0.123)
ConBrsh	Brush	Brush(0.273)	EvrgrnF(0.274)
MxBrush	Brush	Brush(0.342)	LowRes(0.157)
BrUplnd	BrGrnd		Brush(0.416) EvrgrnF(0.5)
Stream	Water	Water(0.125)	Brush(0.125) EvrgrnF(0.333) Wetland(0.375)
NatLake	Water	Water(0.289)	Wetland(0.5)
Rsrvr	Water	Water(0.594)	Wetland(0.152)
TdlRiv	Water Wetland	Water(0.728) Wetland(0.221)	
TdlBay	Water	Water(0.767)	Wetland(0.125)
Dredge	Water	Water(0.294)	HighRes(0.176) Wetland(0.441)
Ocean	Water	Water(0.619)	Wetland(0.142)
SlMrsh	Wetland	Wetland(0.638)	Water(0.247)
SlMrshV	Wetland	Wetland(0.603)	EvrgrnF(0.116)
FrMrsh	Wetland	Wetland(0.595)	Water(0.107)
VegDune	Wetland	Wetland(0.204)	Comm(0.108) Indstrl(0.397)
PhrgCWt	Wetland	Wetland(0.677)	
DecWdWt	Wetland DeciF	DeciF(0.149)	Brush(0.283) EvrgrnF(0.254)
ConWdWt	Wetland EvrgrnF	EvrgrnF(0.831)	
CedarWt	EvrgrnF Wetland	EvrgrnF(0.905)	
DecBrWt	Brush Wetland	Brush(0.25) Wetland(0.219)	EvrgrnF(0.225)

NJ LULC	MATCHED MD LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING

ConBrWt	Brush Wetland	Wetland(0.114)
		Brush(0.179)
			EvrgrnF(0.507)

MxBrWtD	DeciF MxFrst Wetland	Wetland(0.171)	EvrgrnF(0.336) Brush(0.254)

MxBrWtC	EvrgrnF MxFrst Wetland	Wetland(0.211)
		EvrgrnF(0.382)
			Brush(0.236)

HrbWtNT	Brush Wetland	Wetland(0.340)
		Brush(0.151)	

PhrgWet	Wetland	Wetland(0.462)	Brush(0.148) EvrgrnF(0.118)
MxFrWtD	DeciF MxFrst Wetland		EvrgrnF(0.519) Brush(0.249)
MxFrWtC	EvrgrnF MxFrst Wetland	EvrgrnF(0.701)	Brush(0.149)
BrndWet	BrGrnd Wetland	Wetland(0.5)	EvrgrnF(0.5)
Beach	Beach	Beach(0.305)	HighRes(0.186) Indstrl(0.238)
BrGrnd	BrGrnd		
Extr	Extr		Beach(0.142) Indstrl(0.630)
AltLnd			Indstrl(0.278) Comm(0.245)
DstrbWt	Wetland	Wetland(0.292)	Indstrl(0.102) Brush(0.116)
Transi	OpenUrb BrGrnd		Indstrl(0.384) Comm(0.103)
Barren	BrGrnd		Indstrl(0.705)

b

NJ LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING

HighRes	MultFam MblHm	MultFam(0.433)
		MblHm(0.140)
			Indstrl(0.125)

MedRes	SinFam MblHm	SinFam(0.227)
		MblHm(0.150)
			MultFam(0.183) MixRng(0.128)

LowRes	SinFam	SinFam(0.202)	MixRng(0.228)
RurlRes	SinFam	SinFam(0.152)	MixRng(0.189)
MixRes	SinFam MultFam MblHm	MblHm(0.5)	Watrway(0.5)
Comm	Retail VclAct Warehs OthrCom	Retail(0.159)	MultFam(0.328) Indstrl(0.107)
Milit	Inst		Retail(0.363) InldSnd(0.181) MultFam(0.181)
Indstrl	Indstrl	Indstrl(0.316)	Warehs(0.134) Retail(0.131)
Transp			Indstrl(0.124) MultFam(0.125) Retail(0.110)
Road			SinFam(0.150) MultFam(0.473)
Bridge			BayCove(1.0)
Airport			MultFam(0.175) Feedlot(0.108) Indstrl(0.202)
WtlndWa	Wetland	Wetland(0.179)	MixRng(0.153) ShrbRng(0.280) EvrgrnF(0.161)
UpldWaD			MixRng(0.117) Feedlot(0.117) SinFam(0.176) MblHm(0.117) MixUrb(0.176) Frmstd(0.176)

NJ LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING

UpldWa	ShrbRng MixRng	MixRng(0.165)
		ShrbRng(0.317)	

StrmBas			MultFam(0.157)
ICCmplx	INDUSTRIAL & COMMERCIAL		MultFam(1.0)
MixUrb	MixUrb		MultFam(0.490)
OthrUrb	OthrUrb		MultFam(0.138)
Cemet	OthrUrb	OthrUrb(0.124)	Frmstd(0.117) SinFam(0.137)
WtCemet	OthrUrb Wetland		OrchHrt(0.5) Frmstd(0.5)
Phrg			Utility(0.25) MultFam(0.5) MblHm(0.25)

MngWtld	Pasture HerbRng Wetland	HerbRng(0.111)
		Pasture(0.126)	

Recreat	Recreat	Recreat(0.103)	
Athlet	Inst		Feedlot(0.164) Recreat(0.107)
Stadium	Inst		Indstrl(0.263) InldSnd(0.105) MultFam(0.157)
MngWtRe	Wetland Recreat	Recreat(0.222)	Pasture(0.123)

CrpPstr	Crop Pasture	Pasture(0.104)
		Crop(0.156)	

AgriWet	Crop Pasture Wetland	Crop(0.191)
		Pasture(0.110)	

FmAgrWt	IdleFld Wetland HerbRng ShrbRng MixRng	IdleFld(0.2)
		MixRng(0.173)
		ShrbRng(0.268)	

OrchHrt	OrchHrt		Crop(0.107) Trans(0.113)
FeedOp	Feedlot	Feedlot(0.204)	Frmstd(0.183) Indstrl(0.102)
OthrAgr	OrchHrt Feedlot OthrAgr	Feedlot(0.122)	Frmstd(0.134)

DeciF10	DecFrst ShrbRng MixRng	ShrbRng(0.196)
		MixRng(0.192)
			Wetland(0.198)

DeciF50	DecFrst		Wetland(0.254) MixRng(0.189) ShrbRng(0.216)

ConiF10	EvrgrnF ShrbRng MixRng	EvrgrnF(0.361)
		ShrbRng(0.152)
			Wetland(0.189)

ConiF50	EvrgrnF	EvrgrnF(0.685)	
Plant	EvrgrnF	EvrgrnF(0.758)	

MxCnF10	MixFrst EvrgrnF ShrbRng MixRng	ShrbRng(0.161)
		EvrgrnF(0.233)
		MixRng(0.123)
			Wetland(0.283)

MxCon50	MixFrst EvrgrnF	EvrgrnF(0.387)	ShrbRng(0.109) Wetland(0.286)

MxDec10	MixFrst DecFrst ShrbRng MixRng	ShrbRng(0.183)
		MixRng(0.130)
			Wetland(0.358) EvrgrnF(0.117)

MxDec50	MixFrst DecFrst		EvrgrnF(0.132) ShrbRng(0.166) Wetland(0.432) MixRng(0.103)

OldFld	HerbRng MixRng ShrbRng ClrCut	MixRng(0.104)
		ClrCut(0.113)	

PhrgOld	HerbRng MixRng	MixRng(0.344)	ShrbRng(0.103)

DecBrsh	ShrbRng MixRng	MixRng(0.195)
		ShrbRng(0.231)
			Wetland(0.114)

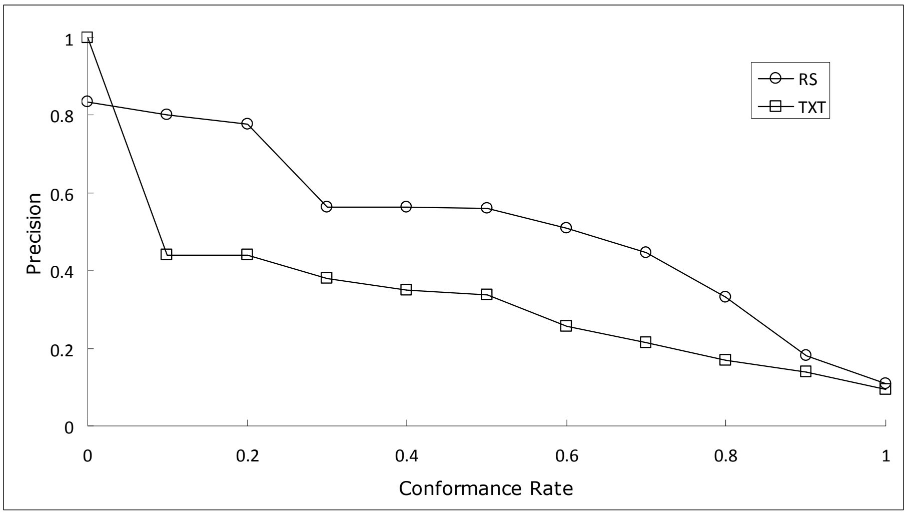
ConBrsh	ShrbRng MixRng	ShrbRng(0.141)
		MixRng(0.130)
			EvrgrnF(0.307)

NJ LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
MxBrush	HerbRng ShrbRng MixRng	ShrbRng(0.243) MixRng(0.180)	Wetland(0.157)
BrUplnd	ClrCut Trans		EvrgrnF(0.333) Wetland(0.416)
Stream	Watrway		EvrgrnF(0.25) Wetland(0.333) NtrlLk(0.291)
NatLake	NtrlLk	NtrlLk(0.421)	MixRng(0.105) Rsrvr(0.131)
Rsrvr	Rsrvr	Rsrvr(0.316)	BayCove(0.205) Watrway(0.155)
TdlRiv	BayCove Wetland	BayCove(0.533)	Rsrvr(0.133) Watrway(0.236)
TdlBay	BayCove	BayCove(0.678)	Watrway(0.178)
Dredge	Rsrvr Watrway	Watrway(0.705)	BayCove(0.117)
Ocean	BayCove	BayCove(0.761)	Indstrl(0.142)
SlMrsh	Wetland		EvrgrnF(0.139) Watrway(0.146) BayCove(0.538)
SlMrshV	Wetland		EvrgrnF(0.185) Watrway(0.190) MixRng(0.159)
FrMrsh	Wetland		Watrway(0.227) MixRng(0.161) BayCove(0.105)
VegDune	Wetland InldSnd	InldSnd(0.277)	Indstrl(0.132) MultFam(0.144)
PhrgCWt	Wetland		Watrway(0.219) EvrgrnF(0.189) MixRng(0.193)
DecWdWt	Wetland DecFrst	Wetland(0.426)	MixRng(0.141) EvrgrnF(0.119)
ConWdWt	Wetland EvrgrnF	EvrgrnF(0.783) Wetland(0.103)	
CedarWt	EvrgrnF Wetland	EvrgrnF(0.893)	
DecBrWt	ShrbRng Wetland MixRng	Wetland(0.215) ShrbRng(0.105) MixRng(0.171)	EvrgrnF(0.215)
ConBrWt	ShrbRng Wetland MixRng		EvrgrnF(0.534)
MxBrWtD	ShrbRng Wetland DecFrst MixRng	Wetland(0.223) ShrbRng(0.109) MixRng(0.143)	EvrgrnF(0.302)
MxBrWtC	ShrbRng Wetland EvrgrnF MixRng	Wetland(0.205) EvrgrnF(0.370) MixRng(0.176)	
HrbWtNT	Wetland HerbRng		ShrbRng(0.105) EvrgrnF(0.128) MixRng(0.145)
PhrgWet	HerbRng Wetland		Watrway(0.135) EvrgrnF(0.192) ShrbRng(0.130) MixRng(0.205)
MxFrWtD	Wetland DecFrst MixFrst	Wetland(0.469)	EvrgrnF(0.301)
MxFrWtC	Wetland EvrgrnF MixFrst	Wetland(0.305) EvrgrnF(0.539)	
BrndWet	HerbRng Wetland	Wetland(0.5)	NtrlLk(0.5)
Beach	Beach	Beach(0.149)	Watrway(0.126) Indstrl(0.246) BayCove(0.104) InldSnd(0.208)
BrGrnd			
Extr	Extr	Extr(0.275)	Indstrl(0.153) InldSnd(0.178)
AltLnd			
DstrbWt	Wetland		EvrgrnF(0.147)
Transi	Trans		Feedlot(0.121)

NJ LULC	**MATCHED DE LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
Barren	InldSnd	InldSnd(0.235)	Extr(0.220) Trans(0.132)

Matches in Table 4.1 with the similarity less than a threshold of 0.1 are considered to be noises and discarded. We use the precision-conformance metric, introduced in section 3.4.2, to evaluate the algorithm to human matches. The 11-point interpolated precision (Manning et al. 2009) is presented in Figure 4.3 (a) matching NJ LULC to MD LULC, and (b) matching NJ LULC to DE LULC. The algorithm is compared with an automated feature based semantic integration method applied to geospatial data portals (Zhou and Wei 2008). As shown in Figure 4.3, our integration method based on remote sensing has a much better overall performance than previous feature-based method in both NJ to MD and NJ to DE experiments. When matching NJ LULC to MD LULC, the conformance rate is 78%, and when matching NJ LULC to DE LULC, the conformance rate is 70% (both numbers are calculated when threshold is set to 0.1).

a



b

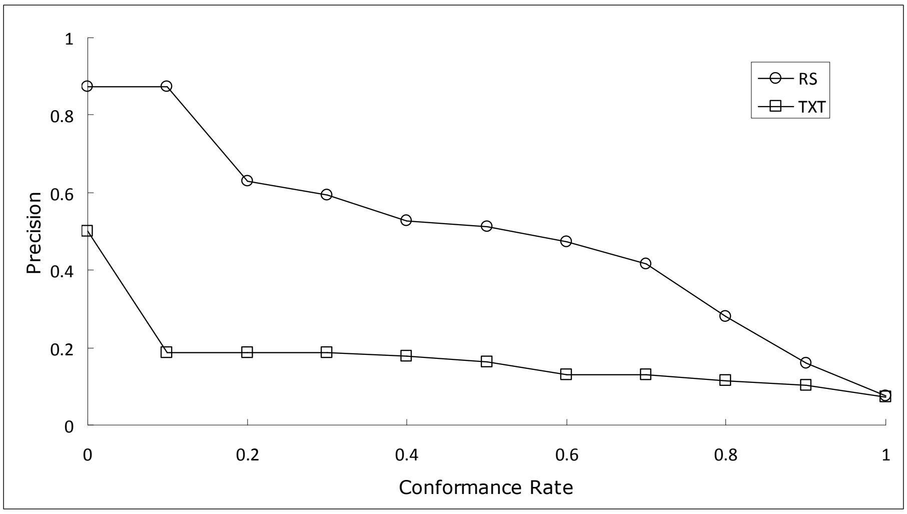


Figure 4.3 Precision-conformance curve of Matching NJ LULC to MD LULC (a), and NJ LULC to DE LULC (b). RS stands for our remote sensing based integration method. TXT stands for the lexical semantic method in (Zhou and Wei 2008).

4.5 Discussion on Non-conformance

The disparity of non-conforming matches is measured by the level of their minimal common upper category. The average disparity level of matching NJ to MD is about 0.50 and 0.41 for NJ to DE. A value near 0.5 means that about half of the non-conforming matches happen on level 2 and the other half on level 1, given there are very few level 3 non-conforming matches. If we look at the level 0 non-conforming matches, we can find several of them are caused by errors in parcel delineation. For example, Beach parcels in NJ are mistakenly matched to Bays and Coves of DE LULC (Appendix II 27) due to wrong parcel boundaries. But for most of the level 0 non-conforming matches, two main categories are observed. The first category includes mismatches between urban and non-urban categories, which are caused by several causes discussed in section 3.3.5. The second main category involves the conceptual ambiguities in the definition of Wetland. As before, a random 10% sampling for each non-conforming match is used to quantify the impact of each cause on integration.

Level 0 non-conforming matches

55% of level 0 non-conforming matches are categorized as the urban/non-urban mismatch, when matching NJ LULC to MD LULC. There are several reasons responsible for this type of mismatch, among which *multiple labeling* (section 3.4.3) is commonly observed. For example, Residential (Single Unit, Low Density) of NJ LULC is matched to Brush of MD LULC (Appendix II 22), because some residential areas are built on bushy areas that still preserve the features of brush land.

On the other hand, matching non-urban parcels to urban categories is more complicated. For example, some parcels of NJ LULC category Mixed Deciduous/Coniferous Brush/Shrubland are matched to Low-Density Residential of MD LULC. Multiple labeling alone cannot explain this type of match. Although Low-density Residential lands can be similar to Mixed Deciduous/Coniferous Brush/Shrubland, it is hard to explain why these Mixed Deciduous/Coniferous Brush/Shrubland parcels are not matched to conforming categories, such as Brush land. Behind this are several reasons. First, as aforementioned, a substantial amount of MD Low-density residential parcels are in brush or shrub land, and this explains the spectral similarity. Second, if there is a noticeable non-vegetated "hole" (e.g. bare soil) in a brush parcel, Landsat data are not able to accurately tell if it is a bare ground or a building top (section 3.4.3, Limitation of Remote Sensing). In these cases, Low-Density Residential is a more preferable category for these brush parcels. Third, a SVM classifier's side effect introduced in the procedure error part of section 3.4.3 also contributes to this non-conforming match. When training a SVM classifier, a category (such as Low-Density Residential) with large and heterogeneous training samples tends to push the hyperplane outward and therefore becomes a more preferable category to match. The combination of these three factors leads to the match from several non-urban types in NJ to urban categories of MD LULC.

Other causes of urban/non-urban mismatches include land use change and procedure error. Land use change may convert previous non-urban parcels to urban areas. For example, Transitional Areas (sites under construction) parcels matched to Industrial after the construction is complete (Appendix II 24). A kind of procedure errors that relates to the removal of pixels on parcel boundaries is also partially responsible for several mismatches involving LULC parcels that usually are long and narrow in shape. For example, some Dredged Lagoon parcels in NJ are matched to Low-Density Residential of MD LULC (Appendix II 25 and 26). This is a strange mismatch, because spectrally they should be way different. We look at the parcels of this mismatch and find out that Dredged Lagoon parcels are always too narrow to be effectively captured by Landsat ETM+, and mixed pixels are pervasive in the parcels. In this case, removing pixels on edges cannot "purify" pixels, but has uncertain effects on the matching. An easy fix of this problem is to discard small parcels from the matching process, and our next chapter will discuss the results of this experiment.

When matching NJ LULC to DE LULC, urban/non-urban non-conforming matches constitute a proportion of 27% of all level 0 non-conforming matches. The majority (62%) of level 0 non-conforming matches between NJ LULC and DE LULC are caused by the semantic heterogeneity between the definitions of the category and sub-categories of wetland. NJ LULC has a rather detailed definition of wetland (Figure 4.1, Appendix I). On the contrary, as pointed out in section 3.3.2, DE LULC has an ambiguous definition of Wetland, which induced many non-conforming matches when integrating MD LULC and DE LULC. The integration of MD and DE LULC suggests that the majority of Wetland parcels in DE LULC are highly vegetated as forestland, and therefore non-forested wetland parcels in MD are difficult to find a right match in DE LULC. Similarly, when matching NJ LULC to DE LULC, a substantial amount of forestland parcels are matched to Wetland DE LULC, while non-forested wetland parcels (e.g. different kinds of marsh lands) in NJ are matched to water bodies, because there are no similar categories for them in DE LULC. Figure 4.4 shows the seasonal NDVI of a parcel of Mixed Forest (More Deciduous with High Crown Closure) in NJ, along with the average seasonal NDVI of Deciduous Forest and Wetland in DE. This parcel is matched to Wetland DE by algorithm, and its NDVI curve explains this result. A further divided Wetland category in DE LULC should help increasing the matching accuracy, but as it is currently aggregated to be a level 1 category, all mismatches are considered level 0 non-conformance. On the other hand, ambiguously aggregated as well, Wetland in MD LULC happens to be more similar to the non-forested subcategories than forested subcategories of Wetland in NJ LULC. In Figure 4.5, the seasonal NDVI of a parcel of Deciduous Wooded Wetlands NJ is more similar to the NDVI of Evergreen Forest than Wetland in MD. It is interesting to observe in the figure that the NDVI curve of Deciduous Wooded Wetlands has a dip on September 12th, which is the sign of water inundation.

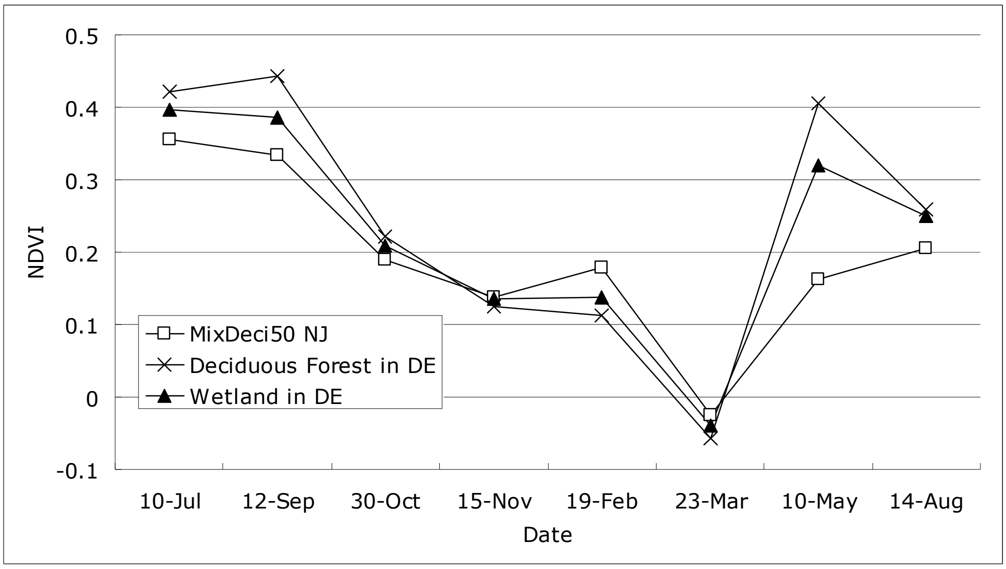


Figure 4.4 Seasonal NDVI curves of Mixed Forest (More Deciduous with High Crown Closure) NJ LULC, denoted as MixDeci50 NJ, and Deciduous Forest and Wetland of DE LULC.

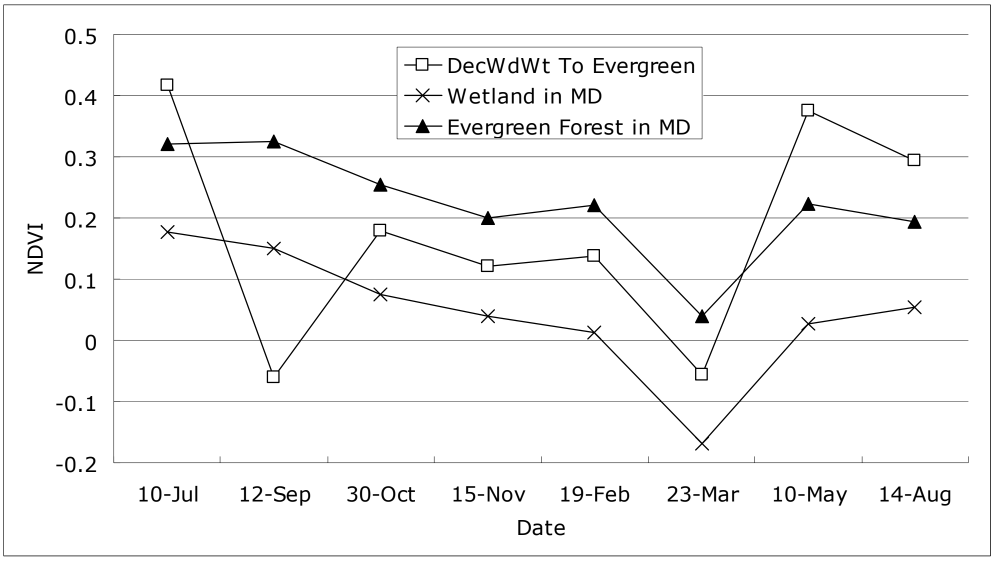


Figure 4.5 Seasonal NDVI curves of Deciduous Wooded Wetlands, denoted as DecWdWt, and Wetland and Evergreen Forest of MD LULC

Level 1 non-conforming matches

By definition, level 1 non-conforming matches are among level 2 subcategories under each one of the level 1 categories. They are more commonly caused by the limitation of remote sensing in LULC mapping, and anticipated when locating the possible semantic heterogeneities in NJ LULC (section 4.3). For example, algorithm matching Athletic Fields (Schools) NJ to Recreational DE is disapproved by human evaluators because Athletic Fields (Schools) NJ are always associated with schools, but Recreational DE by definition must not. Industrial parcels in NJ are matched by algorithm to Commercial MD and Retail Sales/Wholesale/Professional Services DE. Beaches NJ and Inland Natural Sandy Areas DE are matched by algorithm. Identifying and correcting these mismatches is beyond the capability of Landsat or remote sensing in general.

But for some level 2 categories that are separable using remote sensing, non-conforming matches actually provided insight into how the category is used in the field. For example, in NJ LULC, Deciduous Forest (High Crown Closure) contains deciduous stands with crown closures greater than 50%, and the majority of deciduous forests in New Jersey should be in this category, said in the category definition. Based on this definition, human evaluation undoubtedly matches this category to Deciduous Forest in MD and DE LULC. However, our algorithm matches 38.3% of its parcels to Brush MD, much higher than 17.5% to Deciduous Forest MD, and for DE LULC, our algorithm matches 40.5% to Rangeland DE and 25.4% to Wetland DE, while less than 10% (lower than predefined threshold) to Deciduous Forest DE. High resolution remotely sensed imagery shows that these Deciduous Forest (High Crown Closure) parcels do not have 50% canopy coverage as defined (Appendix II 28). Figure 4.6 shows the seasonal NDVI of a Deciduous Forest (High Crown Closure) parcel and explains why it is more similar to Brush rather than Deciduous Forest of MD LULC.

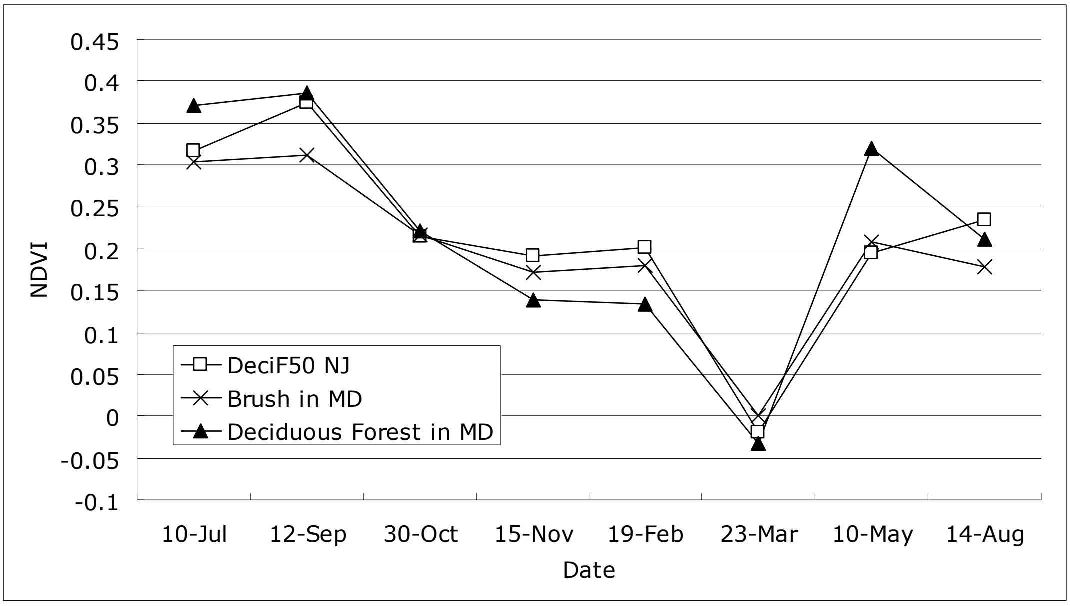


Figure 4.6 Seasonal NDVI curves of Deciduous Forest (High Crown Closure) of NJ LULC, denoted as DeciF50 NJ, and Brush and Deciduous Forest of MD LULC

4.6 Conclusion on Generality

Remote sensing is a widely used data source of LULC mapping because of spectral signature: the difference in actual LULC status corresponds to different pixel values in the remotely sensed images. Based on the spectral signature, we proposed using remote sensing to compare LULC categories via their actual LULC status observed by remote sensing sensors. Therefore, the generality of remote-sensing-based integration of LULC classification systems depends on two factors: the availability and applicability of remote sensing data and the comparability of LULC categories.

Experimentally, as showed in Figure 4.3, the matches found by the remote sensing based method are much more conforming to human evaluations than those of feature-based method in previous work (Zhou and Wei 2008) in all our experiments involving MD LULC, DE LULC, and NJ LULC. This result confirmed the generality of using remote sensing to integrate LULC classifications in neighboring regions.

However, if a method is data driven, it may be limited if data availability and quality cannot be assured. Our method depends on remotely sensed data to provide a consistent measurement on LULC status, which makes the requirement on data generality two-fold. Firstly, the LULC classifications in comparison must be in areas covered by same/similar type of remotely sensed data. This is one of the reasons why we use Landsat data: it has a global coverage. Secondly, remotely sensed data in use must be consistent regarding to the relation between land cover and reflectance values, which means if multiple scenes are involved, the effects of the atmosphere, sensor, and sun on land surface reflectance must be corrected. Just recently (June 2001), the Global Land Cover Facility (GLCF) at the University of Maryland College Park has launched the first global surface reflectance dataset based on the Landsat series of satellites (http://landcover.org/data/gls_SR/). These global reflectance data are important data sources to enable the remote sensing based algorithm to be applied globally. In future, it will be very interesting to incorporate these data in our method and use them to integrate LULC classifications on continental and global scales.

To deal with these larger areas, the remote-sensing-based method has the potential to serve as a translation between classification systems in distant areas. The use of different languages poses an additional problem. But this linguistic barrier could be considered as a special type of semantic heterogeneity, which lexical semantic methods (except machine translation), including our information retrieval method, may or may not be able to handle. Remote sensing, however, in this case, could provide valuable translation between different languages in a LULC study.

However, it is very possible that LULC status, either at a continental scale or for distant areas around the globe (potentially in different languages), will not be comparable due to multiple factors including environmental factors such as climate and hydrology, and topography and also due to different land use practices resulting in non-comparable land use types.. Consequently these LULC classifications could be too heterogeneous to be integrated directly. Therefore, applying our remote sensing method at such very broad scales would be difficult to accomplish. Instead, in the future we may need to look for an expansion strategy, such that more comparable classifications from neighboring regions should be first compared and integrated: the integration will then progressively incorporate more comparable classifications and gradually grow to a continental scope.

For distant areas between which there are no obvious gradations in land use types (e.g. between the US and China and Brazil), the application of our remote sensing method will be more challenging. In this case, the direct application of remote sensing method can still disclose semantic heterogeneities, mostly in form of compounding conflicts, but achieving a highly reliable matching is likely to be difficult. To attack this problem more experiments will be needed, and this is one of many reasons why the newly launched global surface reflectance data are essential.

Chapter 5: Using Large Parcels to Reduce User Side Non-conformance

5.1 Introduction

In chapter 3, we separated five main reasons leading to non-conformance between algorithm and human evaluation, and they are 1) conceptual ambiguity, 2) LULC data quality, 3) LULC change, 4) limitation of remote sensing, and 5) procedure error. Among these five causes, conceptual ambiguity, LULC data quality, and LULC change are controlled by the producers of the LULC map, and the limitation of remote sensing and the procedure error are on the user side. The limitation of remote sensing leads to non-conforming matches because of insufficient discriminating power of Landsat or remote sensing in general. The procedure error is brought by the matching algorithm itself.

Producer side and user side non-conformance should be separated and treated differently. Discovering producer side non-conformance (mainly caused by semantic heterogeneities in LULC classifications and defects in LULC maps) is one of the purposes of remote sensing based matching algorithm. Reducing the impact of user side non-conformance helps producer side non-conformance stand out.

On the user side, a main kind of procedure error originates from the excessive removal of pixels on parcel boundaries. If a parcel is small and/or in a long and narrow shape, pixels to be removed on the boundary compose a large proportion of all pixels, and without pixels on boundary the parcel then becomes unidentifiable in the matching algorithm because of lack of usable pixels.

A solution to this problem is to refine the parcel level statistics to only consider large parcels and employ the matching algorithm on this refined statistics. For large parcels, the removal of pixels on boundary is no more than eliminating potential mixed pixels of different LULC types, and will benefit the accuracy. The other important reason to use large parcels is because they are much easier to be accurately described by Landsat, which in turn will enhance the discriminating power of Landsat data and benefit the accuracy as well.

In this chapter we will implement the similarity measurement using statistics calculated from large parcels in section 5.2, and in section 5.3 we will compare the integration results with the integration we got in chapter 3 and 4. Then in section 5.4, we will present a method to determine semantic relation between LULC categories based on two way matching. The conclusion of this chapter is given in section 5.5.

5.2 Methodology

In this research, we define large parcels to be the parcels that are large enough to contain 50 or more Landsat ETM+ pixels. In implementation, we adopt the same methodology as in chapter 3 but only include large parcels in calculating parcel level statistics. Then from the representative and accurately separable parcels (section 3.2.8) in the source LULC map, we select large parcels to train a SVM classifier, which is later employed to classify all large parcels in the target LULC map to obtain parcel level matches. The similarity between LULC categories is calculated via an extensional approach – estimating similarity of concepts by counting their parcel level matches. Different from previous calculation in chapter 3, this time only parcel level matches between large parcels are considered.

From the discussion above, we know using large parcels in the matching algorithm can successfully reduce the chance of procedure errors. Now the non-conforming matches are more likely caused by producer side reasons, such as conceptual ambiguities or the complicated relation between labeled land use and its actual land cover. Moreover, the spatial resolution of Landsat ETM+ images and its derived textural data is sufficient for large parcels to be accurately captured, which means the data representation of LULC categories and therefore the matching results will now reflect the actual land cover status more closely, and this will have a contribution in reducing the uncertainty between remote sensing and actual land cover status.

However, a better data representation of land cover does not simplify the relationship between land cover and land use, which is occupied by conceptual ambiguities, such as the multiple labeling. As Comber et al. (2005) pointed out, LULC classification systems are “not determined by the reflectance properties of land cover and their inferred relationship with biology alone; rather their specification combines policy objectives at regional, national or international levels with the individual and institutional objectives of those charged with creating the derived land cover map to inform policy.” Comber et al. (2005) then concluded that political processes have an influence on LULC classification systems as profound as do technical aspects, but its influence has never been disclosed to the data users.

We therefore should understand the goal of reducing user side non-conformance (via better data representation) is not to only achieve a matching result that is more conforming to human evaluations, but instead via emphasizing the producer side non-conformance, we can also disclose the semantic heterogeneities happened between LULC classification systems.

5.3 Comparison of Results

The algorithm matching results using large parcels are shown in Table 5.1 a (DE LULC to MD LULC), b (MD LULC to DE LULC), c (NJ LULC to MD LULC) and d (NJ LULC to DE LULC). In four tables, categories are denoted by codes introduced in Appendix I. The columns from left to right mean 1) Source LULC categories, 2) match(es) in target LULC by human evaluators, 3) conforming matches (algorithm and human match same), 4) non-conforming matches (algorithm and human match different), 5) conforming matches using large parcels, and 6) non-conforming matches using large parcels. The number in the parentheses after each algorithm result (in columns 3 to 6) is the similarity of that match, calculated as the ratio of the number of parcels in this match to the number of all considered parcels in the source category.

Table 5.1 Results of Matching Using Large Parcels

a

DE LULC	MATCHED MD LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING

SinFam	LowRes
	MedRes
		LowRes(0.198)
		MedRes(0.183)
			Pasture(0.116)
			OpenUrb(0.101)
				MedRes(0.492)
				LowRes(0.232)	

MultFam	MedRes
	HighRes
		HighRes(0.484)	Comm(0.315)	MedRes(0.238)
				HighRes(0.476)
					Comm(0.190)

MblHm	HighRes	HighRes(0.157)	LowRes(0.129)
			Comm(0.106)
			OpenUrb(0.106)
			MedRes(0.219)
				HighRes(0.1)	MedRes(0.625)
					LowRes(0.1)

Retail	Comm	Comm(0.372)	HighRes(0.108)
			Indstrl(0.412)
				Comm(0.464)	Indstrl(0.5)

Indstrl	Indstrl	Indstrl(0.716)	Comm(0.221)	Indstrl(0.718)	Comm(0.218)

MixUrb	LowRes
	MedRes
	HighRes
	Comm
	Indstrl
	Inst
		Indstrl(0.177)
		Inst(0.124)
		Comm(0.372)
				Inst(0.115)
				MedRes(0.115)
				Indstrl(0.192)
				Comm(0.5)	

OthrUrb	OpenUrb		Pasture(0.187)
			Crop(0.119)
					Crop(0.434)
					Extr(0.130)

Inst	Inst	Inst(0.161)	Indstrl(0.318)
			Comm(0.258)
			HighRes(0.120)
				Inst(0.166)	Comm(0.166)
					Extr(0.2)
					Indstrl(0.333)

Recreat	OpenUrb	OpenUrb(0.115)	Inst(0.107)
			Comm(0.155)
			AgrBldg(0.163)
			Indstrl(0.119)
				OpenUrb(0.386)	Crop(0.159)
					LowRes(0.113)
					Inst(0.113)

Crop	Crop	Crop(0.502)	FeedOp(0.121)
			Pasture(0.151)
				Crop(0.862)	

Pasture	Pasture	Pasture(0.314)	Crop(0.314)
			AgrBldg(0.169)
				Pasture(0.3)	Crop(0.5)
					Indstrl(0.1)
					FeedOp(0.1)

OrchHrt			LowRes(0.111)
			Crop(0.222)
			Pasture(0.177)
			DeciF(0.177)
					LowRes(0.111)
					GdnCrop(0.111)
					DeciF(0.222)
					Crop(0.555)

OthrAgr	Crop
	AgrBldg
		Crop(0.105)	Pasture(0.105)
			Indstrl(0.263)
			FeedOp(0.473)
				AgrBldg(1.0)	

ShrbRng	Brush
	Pasture
		Brush(0.218)	LowRes(0.231)
			DeciF(0.258)
				Brush(0.467)	EvrgrnF(0.145)
					DeciF(0.225)

MixRng	Brush
	Pasture
		Brush(0.144)	DeciF(0.152)
			LowRes(0.289)
			MedRes(0.115)
				Brush(0.25)	DeciF(0.5)
					LowRes(0.25)

DecFrst	DeciF	DeciF(0.608)	Brush(0.102)
			LowRes(0.133)
				DeciF(0.980)	

EvrgrnF	EvrgrnF	EvrgrnF(0.729)		EvrgrnF(0.945)	

MixFrst	MxFrst	MxFrst(0.242)	Brush(0.142)
			EvrgrnF(0.107)
			DeciF(0.288)
			LowRes(0.103)
				MxFrst(0.551)	DeciF(0.332)

ClrCut	Brush	Brush(0.147)	FeedOp(0.107)	Brush(0.259)	Crop(0.740)

DE LULC	MATCHED MD LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING

			Pasture(0.203)
			Crop(0.350)		

Rsrvr	Water	Water(0.789)		Water(0.823)	Wetland(0.117)
BayCove	Water	Water(0.934)		Water(1.0)	

Wetland	Wetland	Wetland(0.123)	Brush(0.157)
			DeciF(0.317)
			MxFrst(0.159)
				Wetland(0.238)	DeciF(0.309)
					EvrgrnF(0.103)
					MxFrst(0.298)

Beach	Beach	Beach(0.714)	Indstrl(0.285)	Beach(1.0)	

InldSnd			Indstrl(0.6)
			Beach(0.28)
					Indstrl(0.4)
					Beach(0.6)

Extr	Extr		Indstrl(0.606)
			Comm(0.181)
					Beach(0.166)
					Wetland(0.166)
					Indstrl(0.583)

Trans	Crop
	Brush
			FeedOp(0.131)
			Indstrl(0.356)
				Crop(0.190)	GdnCrop(0.190)
					FeedOp(0.142)
					Indstrl(0.380)

b

MD LULC	MATCHED DE LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING
LowRes	SinFam	SinFam(0.382)	MixRng(0.150)	SinFam(0.663)	MixRng(0.149)

MedRes	MultFam
	SinFam
		SinFam(0.374)	MixRng(0.108)	SinFam(0.496)	MixUrb(0.100)
					MblHm(0.158)

HighRes	MultFam
	MblHm
		MblHm(0.120)
		MultFam(0.313)
			SinFam(0.177)	MultFam(0.25)
				MblHm(0.2)
					SinFam(0.25)
					Indstrl(0.1)
					Retail(0.15)

Comm	Retail
	MixUrb
		MixUrb(0.157)
		Retail(0.174)
				MixUrb(0.104)
				Retail(0.388)
					Indstrl(0.179)

Indstrl	Indstrl	Indstrl(0.276)	Retail(0.160)	Indstrl(0.392)	Inst(0.142)
					Retail(0.107)
					Extr(0.178)

Inst	Inst		SinFam(0.134)
			MixUrb(0.111)
				Inst(0.312)	SinFam(0.208)
					MixUrb(0.125)

Extr	Extr	Extr(0.215)	SinFam(0.169)	Extr(0.375)	SinFam(0.187)
					Trans(0.125)

OpenUrb	Recreat	Recreat(0.146)	SinFam(0.213)	Recreat(0.333)	Pasture(0.111)
					SinFam(0.4)

Crop	Crop OthrAgr	Crop(0.414)	SinFam(0.131)	Crop(0.682)	

Pasture	Pasture	Pasture(0.121)	SinFam(0.154)
			Crop(0.181)
				Pasture(0.18)	SinFam(0.13)
					Crop(0.45)

FeedOp			Crop(0.133)
			Frmstd(0.126)
			Feedlot(0.297)
					Pasture(0.142)
					MixUrb(0.142)
					Extr(0.142)
					Recreat(0.142)
					Trans(0.285)
					Crop(0.142)

AgrBldg			Crop(0.162)
			Frmstd(0.103)
			Feedlot(0.159)
					Feedlot(0.166)
					Pasture(0.416)

MD LULC	**MATCHED DE LULC**				
	By HUMAN	**By ALGORITHM**		**USING LARGE PARCEL**	
		CONFORMING	**NON-CONFORMING**	**CONFORMING**	**NON-CONFORMING**
			Pasture(0.115)		

GdnCrop			Trans(0.129)
			Pasture(0.129)
			Crop(0.225)
					Trans(0.25)
					Crop(0.333)

DeciF	DecFrst	DecFrst(0.204)	MixFrst(0.152)
			ShrbRng(0.162)
			MixRng(0.179)
				DecFrst(0.354)	Wetland(0.116)
					MixFrst(0.230)
					ShrbRng(0.147)

EvrgrnF	EvrgrnF	EvrgrnF(0.423)	ShrbRng(0.113)
			MixFrst(0.136)
				EvrgrnF(0.566)	MixFrst(0.215)

MxFrst	MixFrst	MixFrst(0.305)	ShrbRng(0.165)
			MixRng(0.127)
				MixFrst(0.505)	ShrbRng(0.139)
					Wetland(0.101)

Brush	ShrbRng
	MixRng
	ClrCut
	Trans
		ShrbRng(0.202)	EvrgrnF(0.159)	ShrbRng(0.264)
				ClrCut(0.160)
					EvrgrnF(0.217)

Water	Rsrvr
	BayCove
		Rsrvr(0.330)
		BayCove(0.25)
				Rsrvr(0.156)
				BayCove(0.75)	

Wetland	Wetland		MixRng(0.228)
			Watrway(0.112)
			Rsrvr(0.110)
					MixRng(0.173)
					EvrgrnF(0.212)
					BayCove(0.177)
					SinFam(0.154)

Beach	Beach	Beach(0.2)	Retail(0.2)
			BayCove(0.2)
			InldSnd(0.2)
			Indstrl(0.2)
				Beach(0.5)	InldSnd(0.5)

c

NJ LULC	**MATCHED MD LULC**				
	By HUMAN	**By ALGORITHM**		**USING LARGE PARCEL**	
		CONFORMING	**NON-CONFORMING**	**CONFORMING**	**NON-CONFORMING**

HighRes	HighRes	HighRes(0.511)	Indstrl(0.116)
			Comm(0.186)
				HighRes(0.838)	

MedRes	MedRes	MedRes(0.317)	Comm(0.139)
			HighRes(0.211)
				MedRes(0.845)	

LowRes	LowRes	LowRes(0.190)	Brush(0.143)
			MedRes(0.270)
				LowRes(0.444)	Brush(0.111)
					MedRes(0.444)

RurlRes	LowRes	LowRes(0.192)	MedRes(0.157)
			Brush(0.163)
				LowRes(1.0)	

Comm	Comm	Comm(0.361)	HighRes(0.296)
			Indstrl(0.229)
					HighRes(0.5)
					Extr(0.25)
					Indstrl(0.25)

Indstrl	Indstrl	Indstrl(0.577)	HighRes(0.118)
			Comm(0.211)
				Indstrl(0.5)	HighRes(0.5)

Road			Comm(0.193)
			HighRes(0.387)
			Wetland(0.129)
			MedRes(0.193)
					Brush(0.5)
					Comm(0.5)

Airport			HighRes(0.189)
			Indstrl(0.405)
					Indstrl(1.0)

UpldWa	Brush	Brush(0.310)	Pasture(0.132)
			LowRes(0.152)
					MedRes(0.333)
					DeciF(0.333)
					LowRes(0.333)

NJ LULC	MATCHED MD LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING

OthrUrb	OpenUrb		HighRes(0.166)
			Indstrl(0.133)
			Comm(0.207)
					FeedOp(0.285)
					LowRes(0.142)
					Crop(0.571)

Cemet	OpenUrb	OpenUrb(0.103)	Pasture(0.151)
			AgrBldg(0.103)
					Crop(1.0)

Recreat	OpenUrb		HighRes(0.109)
			Comm(0.157)
			Indstrl(0.220)
				OpenUrb(0.363)	Inst(0.136)
					AgrBldg(0.181)

Athlet	Inst	Inst(0.107)	Comm(0.123)
			Indstrl(0.256)
			FeedOp(0.194)
				Inst(0.333)	Crop(0.333)
					AgrBldg(0.333)

Stadium	Inst		HighRes(0.263)
			Indstrl(0.526)
					Indstrl(1.0)

CrpPstr	Crop
	Pasture
		Pasture(0.221)
		Crop(0.169)
			FeedOp(0.124)	Crop(0.577)	Indstrl(0.141)

AgriWet	Crop
	Pasture
	Wetland
		Crop(0.137)
		Pasture(0.210)
			FeedOp(0.110)
			AgrBldg(0.146)
				Crop(0.437)	Indstrl(0.187)
					FeedOp(0.125)
					AgrBldg(0.25)

OrchHrt			AgrBldg(0.138)
			FeedOp(0.112)
			Indstrl(0.129)
			Crop(0.107)
			Pasture(0.113)
					AgrBldg(0.190)
					Crop(0.428)

OthrAgr	FeedOp
	AgrBldg
	GdnCrop
		AgrBldg(0.124)
		FeedOp(0.105)
			Pasture(0.142)
			Indstrl(0.186)
				AgrBldg(0.5)
				Crop(0.5)	

DeciF10	DeciF
	Brush
		Brush(0.352)	MedRes(0.102)
			LowRes(0.121)
				Brush(0.133)
				DeciF(0.266)
					MxFrst(0.6)

DeciF50	DeciF	DeciF(0.175)	Brush(0.383)
			LowRes(0.134)
				DeciF(0.579)	Brush(0.305)

ConiF10	EvrgrnF
	Brush
		EvrgrnF(0.340)
		Brush(0.392)
				EvrgrnF(0.428)
				Brush(0.571)	

ConiF50	EvrgrnF	EvrgrnF(0.675)	Brush(0.126)	EvrgrnF(0.95)	
Plant	EvrgrnF	EvrgrnF(0.751)		EvrgrnF(1.0)	

MxCnF10	MxFrst
	EvrgrnF
	Brush
		Brush(0.419)
		EvrgrnF(0.276)
				MxFrst(0.25)
				Brush(0.25)
				EvrgrnF(0.5)	

MxCon50	MxFrst EvrgrnF	EvrgrnF(0.468)	Brush(0.326)	EvrgrnF(0.628)	Brush(0.318)

MxDec10	MxFrst
	DeciF
	Brush
		Brush(0.483)	EvrgrnF(0.170)	MxFrst(0.222)
				DeciF(0.111)
				Brush(0.444)
					EvrgrnF(0.222)

MxDec50	MxFrst
	DeciF
			EvrgrnF(0.206)
			Brush(0.514)
				MxFrst(0.106)	EvrgrnF(0.147)
					Brush(0.692)

OldFld	Brush	Brush(0.129)	Comm(0.108)
			Pasture(0.166)
				Brush(0.333)	MedRes(0.333)
					Crop(0.333)

DecBrsh	Brush	Brush(0.254)	LowRes(0.165)
			DeciF(0.123)
				Brush(0.8)	DeciF(0.2)

ConBrsh	Brush	Brush(0.273)	EvrgrnF(0.274)		EvrgrnF(1.0)
MxBrush	Brush	Brush(0.342)	LowRes(0.157)	Brush(0.705)	EvrgrnF(0.176)
Rsrvr	Water	Water(0.594)	Wetland(0.152)	Water(0.972)	

TdlRiv	Water
	Wetland
		Water(0.728)
		Wetland(0.221)
				Water(0.954)	

TdlBay	Water	Water(0.767)	Wetland(0.125)	Water(1.0)	
Ocean	Water	Water(0.619)	Wetland(0.142)	Water(1.0)	

NJ LULC	MATCHED MD LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING
SlMrsh	Wetland	Wetland(0.638)	Water(0.247)	Wetland(0.911)	
SlMrshV	Wetland	Wetland(0.603)	EvrgrnF(0.116)	Wetland(0.8)	EvrgrnF(0.2)
FrMrsh	Wetland	Wetland(0.595)	Water(0.107)	Wetland(1.0)	

VegDune	Wetland	Wetland(0.204)	Comm(0.108)
			Indstrl(0.397)
					Beach(0.2)
					Indstrl(0.6)
					Comm(0.2)

PhrgCWt	Wetland	Wetland(0.677)		Wetland(0.794)	EvrgrnF(0.117)

DecWdWt	Wetland
	DeciF
		DeciF(0.149)	Brush(0.283)
			EvrgrnF(0.254)
				DeciF(0.246)	MxFrst(0.188)
					EvrgrnF(0.492)

ConWdWt	Wetland EvrgrnF	EvrgrnF(0.831)		EvrgrnF(0.833)	
CedarWt	EvrgrnF Wetland	EvrgrnF(0.905)		EvrgrnF(1.0)	

DecBrWt	Brush
	Wetland
		Brush(0.25)
		Wetland(0.219)
			EvrgrnF(0.225)	Brush(0.333)	EvrgrnF(0.666)

ConBrWt	Brush
	Wetland
		Wetland(0.114)
		Brush(0.179)
			EvrgrnF(0.507)		EvrgrnF(1.0)

MxBrWtC	EvrgrnF
	MxFrst
	Wetland
		Wetland(0.211)
		EvrgrnF(0.382)
			Brush(0.236)	MxFrst(0.666)	Brush(0.333)

HrbWtNT	Brush
	Wetland
		Wetland(0.340)
		Brush(0.151)
				Wetland(0.4)
				Brush(0.2)
					Extr(0.2)
					DeciF(0.2)

PhrgWet	Wetland	Wetland(0.462)	Brush(0.148)
			EvrgrnF(0.118)
					Brush(0.333)
					DeciF(0.666)

MxFrWtD	DeciF
	MxFrst
	Wetland
			EvrgrnF(0.519)
			Brush(0.249)
				MxFrst(0.172)	EvrgrnF(0.586)
					Brush(0.206)

MxFrWtC	EvrgrnF MxFrst Wetland	EvrgrnF(0.701)	Brush(0.149)	EvrgrnF(0.830)	

Beach	Beach	Beach(0.305)	HighRes(0.186)
			Indstrl(0.238)
				Beach(0.5)	Water(0.125)
					Indstrl(0.25)
					Extr(0.125)

Extr	Extr		Beach(0.142)
			Indstrl(0.630)
					Beach(0.7)
					Indstrl(0.3)

AltLnd			Indstrl(0.278)
			Comm(0.245)
					Indstrl(0.5)
					FeedOp(0.166)
					Crop(0.166)
					GdnCrop(0.166)

AltLnd1					
AltLnd2					

DstrbWt	Wetland	Wetland(0.292)	Indstrl(0.102)
			Brush(0.116)
				Wetland(1.0)	

d

NJ LULC	MATCHED DE LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING

HighRes	MultFam
	MblHm
		MultFam(0.433)
		MblHm(0.140)
			Indstrl(0.125)	MultFam(0.354)
				MblHm(0.387)
					Indstrl(0.193)

NJ LULC	MATCHED DE LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING

MedRes	SinFam
	MblHm
		SinFam(0.227)
		MblHm(0.150)
			MultFam(0.183)
			MixRng(0.128)
				SinFam(0.714)
				MblHm(0.142)
					MultFam(0.130)

LowRes	SinFam	SinFam(0.202)	MixRng(0.228)	SinFam(1.0)	
RurlRes	SinFam	SinFam(0.152)	MixRng(0.189)	SinFam(1.0)	

Comm	Retail	Retail(0.159)	MultFam(0.328)
			Indstrl(0.107)
				Retail(0.5)	Indstrl(0.5)

Indstrl	Indstrl	Indstrl(0.316)	Warehs(0.134)
			Retail(0.131)
				Indstrl(1.0)	

Road			SinFam(0.150)
			MultFam(0.473)
					MixUrb(0.5)
					EvrgrnF(0.5)

Airport			MultFam(0.175)
			Feedlot(0.108)
			Indstrl(0.202)
					Indstrl(1.0)

UpldWa	ShrbRng
	MixRng
		MixRng(0.165)
		ShrbRng(0.317)
					SinFam(1.0)

OthrUrb	OthrUrb		MultFam(0.138)	OthrUrb(0.142)	OthrAgr(0.142)
					SinFam(0.285)
					ClrCut(0.142)
					Crop(0.285)

Cemet	OthrUrb	OthrUrb(0.124)	Frmstd(0.117)
			SinFam(0.137)
					Crop(1.0)

Recreat	Recreat	Recreat(0.103)		Recreat(0.5)	SinFam(0.227)

Athlet	Inst		Feedlot(0.164)
			Recreat(0.107)
				Inst(0.333)	Crop(0.333)
					SinFam(0.333)

Stadium	Inst		Indstrl(0.263)
			InldSnd(0.105)
			MultFam(0.157)
				Inst(1.0)	

CrpPstr	Crop
	Pasture
		Pasture(0.104)
		Crop(0.156)
				Crop(0.656)	Trans(0.171)

AgriWet	Crop
	Pasture
	Wetland
		Crop(0.191)
		Pasture(0.110)
				Crop(0.812)	Trans(0.125)

OrchHrt	OrchHrt		Crop(0.107)
			Trans(0.113)
					Crop(0.476)
					Trans(0.253)

OthrAgr	OrchHrt
	OthrAgr
			Feedlot(0.122)
			Frmstd(0.134)
				OthrAgr(0.5)	Trans(0.5)

DeciF10	DecFrst
	ShrbRng
	MixRng
		ShrbRng(0.196)
		MixRng(0.192)
			Wetland(0.198)	ShrbRng(0.133)	Wetland(0.533)
					MixFrst(0.266)

DeciF50	DecFrst		Wetland(0.254)
			MixRng(0.189)
			ShrbRng(0.216)
				DecFrst(0.194)	MixFrst(0.182)
					Wetland(0.503)

ConiF10	EvrgrnF
	ShrbRng
	MixRng
		EvrgrnF(0.361)
		ShrbRng(0.152)
			Wetland(0.189)	ShrbRng(0.285)
				EvrgrnF(0.428)
					Wetland(0.285)

ConiF50	EvrgrnF	EvrgrnF(0.685)		EvrgrnF(0.98)	
Plant	OrchHrt EvrgrnF	EvrgrnF(0.758)		EvrgrnF(1.0)	

MxCnF10	MixFrst
	EvrgrnF
	ShrbRng
	MixRng
		ShrbRng(0.161)
		EvrgrnF(0.233)
		MixRng(0.123)
			Wetland(0.283)	EvrgrnF(0.25)	Wetland(0.75)

MxCon50	MixFrst
	EvrgrnF
		EvrgrnF(0.387)	ShrbRng(0.109)
			Wetland(0.286)
				EvrgrnF(0.787)	Wetland(0.150)

MxDec10	MixFrst
	DecFrst
		ShrbRng(0.183)
		MixRng(0.130)
			Wetland(0.358)
			EvrgrnF(0.117)
				ShrbRng(0.222)
				MixFrst(0.111)
					Wetland(0.555)
					EvrgrnF(0.111)

NJ LULC	MATCHED DE LULC				
	By HUMAN	By ALGORITHM		USING LARGE PARCEL	
		CONFORMING	NON-CONFORMING	CONFORMING	NON-CONFORMING
	ShrbRng MixRng				

MxDec50	MixFrst
	DecFrst
			EvrgrnF(0.132)
			ShrbRng(0.166)
			Wetland(0.432)
			MixRng(0.103)
					Wetland(0.633)
					EvrgrnF(0.100)
					ShrbRng(0.230)

OldFld	MixRng
	ClrCut
		MixRng(0.104)
		ClrCut(0.113)
				ClrCut(0.333)	Crop(0.666)

DecBrsh	ShrbRng
	MixRng
		MixRng(0.195)
		ShrbRng(0.231)
			Wetland(0.114)	ShrbRng(0.2)	Wetland(0.8)

ConBrsh	ShrbRng
	MixRng
		ShrbRng(0.141)
		MixRng(0.130)
			EvrgrnF(0.307)		EvrgrnF(1.0)

MxBrush	ShrbRng
	MixRng
		ShrbRng(0.243)
		MixRng(0.180)
			Wetland(0.157)	ShrbRng(0.117)	EvrgrnF(0.176)
					Wetland(0.588)

Rsrvr	Rsrvr	Rsrvr(0.316)	BayCove(0.205)
			Watrway(0.155)
					BayCove(1.0)

TdlRiv	BayCove
	Wetland
		BayCove(0.533)	Rsrvr(0.133)
			Watrway(0.236)
				BayCove(0.863)	Rsrvr(0.121)

TdlBay	BayCove	BayCove(0.678)	Watrway(0.178)	BayCove(1.0)	
Ocean	BayCove	BayCove(0.761)	Indstrl(0.142)	BayCove(1.0)	

SlMrsh	Wetland		EvrgrnF(0.139)
			Watrway(0.146)
			BayCove(0.538)
					EvrgrnF(0.723)
					BayCove(0.171)

SlMrshV	Wetland		EvrgrnF(0.185)
			Watrway(0.190)
			MixRng(0.159)
					MblHm(0.2)
					EvrgrnF(0.8)

FrMrsh	Wetland		Watrway(0.227)
			MixRng(0.161)
			BayCove(0.105)
					SinFam(0.285)
					EvrgrnF(0.571)
					MixRng(0.142)

VegDune	Wetland
	InldSnd
		InldSnd(0.277)	Indstrl(0.132)
			MultFam(0.144)
				InldSnd(0.4)	MixUrb(0.2)
					Indstrl(0.4)

PhrgCWt	Wetland		Watrway(0.219)
			EvrgrnF(0.189)
			MixRng(0.193)
					SinFam(0.117)
					ShrbRng(0.205)
					EvrgrnF(0.5)

DecWdWt	Wetland
	DecFrst
		Wetland(0.426)	MixRng(0.141)
			EvrgrnF(0.119)
				Wetland(0.768)	EvrgrnF(0.173)

ConWdWt	Wetland
	EvrgrnF
		EvrgrnF(0.783)
		Wetland(0.103)
				Wetland(0.119)
				EvrgrnF(0.880)	

CedarWt	EvrgrnF Wetland	EvrgrnF(0.893)		EvrgrnF(1.0)	

DecBrWt	ShrbRng
	Wetland
	MixRng
		Wetland(0.215)
		ShrbRng(0.105)
		MixRng(0.171)
			EvrgrnF(0.215)	Wetland(0.333)	EvrgrnF(0.666)

ConBrWt	ShrbRng Wetland MixRng		EvrgrnF(0.534)	Wetland(0.4)	EvrgrnF(0.6)

MxBrWtC	ShrbRng
	Wetland
	EvrgrnF
	MixRng
		Wetland(0.205)
		EvrgrnF(0.370)
		MixRng(0.176)
				Wetland(0.333)
				EvrgrnF(0.666)	

HrbWtNT	Wetland		ShrbRng(0.105)
			EvrgrnF(0.128)
			MixRng(0.145)
					BayCove(0.2)
					SinFam(0.2)
					EvrgrnF(0.2)
					OrchHrt(0.2)
					OthrUrb(0.2)

PhrgWet	Wetland		Watrway(0.135)
			EvrgrnF(0.192)
				Wetland(0.333)	ShrbRng(0.666)

NJ LULC	**MATCHED DE LULC**				
	By HUMAN	**By ALGORITHM**		**USING LARGE PARCEL**	
		CONFORMING	**NON-CONFORMING**	**CONFORMING**	**NON-CONFORMING**

			ShrbRng(0.130)
			MixRng(0.205)		

MxFrWtD	Wetland
	DecFrst
	MixFrst
		Wetland(0.469)	EvrgrnF(0.301)	MixFrst(0.103)
				Wetland(0.620)
					EvrgrnF(0.206)

MxFrWtC	Wetland
	EvrgrnF
	MixFrst
		Wetland(0.305)
		EvrgrnF(0.539)
				EvrgrnF(0.615)
				Wetland(0.369)	

Beach	Beach	Beach(0.149)	Watrway(0.126)
			Indstrl(0.246)
			BayCove(0.104)
			InldSnd(0.208)
				Beach(0.125)	OthrAgr(0.125)
					InldSnd(0.5)
					BayCove(0.125)
					Indstrl(0.125)

Extr	Extr	Extr(0.275)	Indstrl(0.153)
			InldSnd(0.178)
				Extr(0.1)	Beach(0.3)
					InldSnd(0.6)

AltLnd					Indstrl(0.333)
					MixUrb(0.166)
					MblHm(0.166)
					Inst(0.166)
					Crop(0.166)

AltLnd1					
AltLnd2					
DstrbWt	Wetland		EvrgrnF(0.147)		SinFam(1.0)

Table 5.2 Conformance Rate and Precision Comparison

	All Parcels			Large Parcels		
	Conformance	Precision	F	Conformance	Precision	F
DE to MD	0.84	0.35	0.49	0.76	0.42	0.54
MD to DE	0.82	0.34	0.48	0.77	0.31	0.44
NJ to MD	0.78	0.38	0.51	0.56	0.46	0.50
NJ to DE	0.70	0.42	0.53	0.56	0.42	0.48

The comparison of the conformance rate and precision between experiments using large parcels only and all parcels is given in Table 5.2. Using large parcels, the matching algorithm has a decreased conformance rate but an increased precision (at a threshold of 0.1) for all categories in all four experiments. However, in each MD LULC, DE LULC, and NJ LULC dataset there are several categories that do not have any parcels containing 50 or more Landsat ETM+ pixels (listed in Table 5.3). In other word, these categories contain no parcels qualified to participate the matching process, and will be considered "null". For these "null" categories, the extensional matching algorithm could never find a match for them, and these "no matches" will have a negative influence on the conformance rate. A more fair comparison between matching results using all parcels or large parcels only should therefore only include those "non-null" categories (Table 5.4). In this new comparison, using large parcels increased conformance rate in three of four experiments, and the precisions either increase as well or basically remain intact. F measure values, as trade-off between conformance rate and precision, have increased in all four experiments. It is inevitable that the matching result is incomplete when categories that have no large representative parcels are excluded, but as many small parcels are in urban areas, all major forest and wetland categories, which are of special interest in environment analysis and resource management, contain qualified large parcels and were included in the matching process.

Table 5.3 Categories Containing No Large Parcels

MD LULC	Orchards/vineyards/horticulture Bare ground
DE LULC	Vehicle Related Activities Junk/Salvage Yards Warehouses and Temporary Storage Other Commercial Utilities Idle Fields Confined Feeding Operations/Feedlots/Holding Farmsteads and Farm Related Buildings Herbaceous Rangeland Waterways/Streams/Canals Natural Lakes and Ponds
NJ LULC	Mixed Residential Military Installations Transportation/Communication/Utilities Bridge Over Water(WATER) Wetland Rights-of-Way(WETLANDS) Upland Rights-of-Way, Developed Stormwater Basin Industrial and Commercial Complexes Mixed Urban or Built-up Land Cemetery on Wetland(WETLANDS) Phragmites Dominate Urban Area Managed Wetland in Maintained Lawn Green space(WETLANDS) Managed Wetland in Built-up Maintained Rec Area(WETLANDS) Former Agricultural Wetlands (Becoming Shrubby not Built-up)(WETLANDS) Confined Feeding Operations Phragmites Dominate Old Field Severe Burned Upland Vegetation Streams and Canals Natural Lakes Dredged Lagoon Mixed Scrub/Shrub Wetlands (Deciduous Dom.) Severe Burned Wetlands Bare Exposed Rock, Rockslides, etc. Transitional Areas (sites under construction) Undifferentiated Barren Lands

Table 5.4 Conformance Rate and Precision Comparison (Non-null Categories)

	All Parcels			Large Parcels		
	Conformance	Precision	F	Conformance	Precision	F
DE to MD	0.88	0.33	0.48	0.92	0.46	0.61
MD to DE	0.88	0.41	0.56	0.94	0.40	0.56
NJ to MD	0.88	0.43	0.57	0.80	0.47	0.59
NJ to DE	0.71	0.42	0.52	0.77	0.42	0.54

5.4 Interpreting the Results – Determining Semantic Relation between LULC Categories

Our experiment results confirmed that using large parcels in the algorithm can reduce the user side causes of non-conformance, and fewer causes lead to less variety in non-conforming matches. For example, Cropland DE is one of the main agricultural categories. It has 1996 parcels, and 617 large parcels contain 50 or more pixels. Using all parcels in the matching algorithm, 50.2% Cropland parcels in DE are matched to Cropland MD, 12.1% to Feeding Operation MD, and 15.1% to Pasture MD. But when using only large parcels, the matching result converges: 86.2% Cropland parcels in DE are matched to Cropland MD, while no non-conforming matches are reaching a threshold of 0.1.

This convergence in matching candidates makes vague category integration clear. The analysis to extract semantic relations between LULC categories, based on the inclusion and intersection of sets of matching parcels (Figure 3.6 and 3.7), can now be generalized to all forest and wetland categories. For example, in the discussion in section 3.3.5, we concluded that Evergreen Forest DE is subsumed by Evergreen Forest MD. Now using large parcels in the matching algorithm, the subsuming semantic relation between Deciduous Forest categories in DE and MD became clear as well. Deciduous Forest MD as a concept and a category is broader than Deciduous Forest DE, because of the variety in its land cover status (partially similar to Mixed Forest DE, and Shrub/Brush Rangeland DE). MD LULC and DE LULC have similar concepts of Mixed Forest, but Mixed Forest DE might be more vegetated than Mixed Forest MD, because some of the Mixed Forest parcels in MD matched to Shrub/Brush Rangeland DE. As for the rangeland categories, Brush MD is similar to Shrub/Brush Rangeland more than Mixed Rangeland in DE LULC. The semantic heterogeneity of Wetland between MD LULC and DE LULC is already discussed in chapter 4, and an analysis based on similarities in Table 5.1 leads to a similar conclusion: Wetland DE contains much more vegetated parcels than Wetland MD does.

In NJ LULC, forest parcels are divided according to the combination of crown closure and dominant species into nine categories, which are Deciduous Forest (Low Crown Closure), Deciduous Forest (High Crown Closure), Coniferous Forest (Low Crown Closure), Coniferous Forest (High Crown Closure), Plantation, Mixed Forest (More Coniferous with Low Crown Closure), Mixed Forest (More Coniferous with High Crown Closure), Mixed Forest (More Deciduous with Low Crown Closure), and Mixed Forest (More Deciduous with High Crown Closure). In Table 5.1, we can see some parcels in the low crown closure forest categories are matched to Brush MD and Shrub/Brush Rangeland DE. But in general, the subsuming relations stand between subcategories with high and low crown closure and their parent categories' (Deciduous and Coniferous Forest) counterparts in MD LULC, except Mixed Forest (More Deciduous with Low Crown Closure) and Mixed Forest (More Deciduous with High Crown Closure), which appear more like brush rather than forest in MD. On the other hand, forest LULC matching between NJ LULC and DE LULC is complicated by Wetland DE. As found in chapter 3, many Wetland parcels in DE are substantially forested and cannot be separated from forestlands by remote sensing. As a result, Wetland DE attracted many non-conforming matches from forestland categories in NJ.

The subsuming semantic relation also stands between Brush MD and NJ brush subcategories, except that most Coniferous Brush/Shrubland NJ parcels seem more similar to Evergreen Forest in both MD and DE. As before, Wetland DE biased the matching result of several brush categories (i.e. Deciduous Brush/Shrubland and Mixed Deciduous/Coniferous Brush/Shrubland) from NJ LULC to DE LULC.

Judging from the matches of NJ wetland subcategories, we can determine that Coastal Wetland of NJ LULC (this concept is defined in classification but not used in map) is approximately equivalent to Wetland MD, as its subsumed subcategories, i.e. Saline Marshes (Low marsh vegetation), Saline Marshes (High marsh vegetation), Freshwater Tidal Marshes, and Phragmites Dominate Coastal Wetlands, all display very high similarities to Wetland MD. The only exception is Vegetated Dune Communities, which has a special sandy cover. As for the Interior Wetland of NJ LULC (also defined in classification but not used in map) and its subcategories (Deciduous Wooded Wetlands, Coniferous Wooded Wetlands, et al.), Evergreen Forest and Brush MD are more similar to them in terms of their common high vegetation coverage. When matching NJ LULC to DE LULC, however, Interior Wetland subcategories are naturally matched to Wetland DE, while Coastal Wetland categories are difficult to find an appropriate match. In the result, Evergreen Forest, instead of Wetland, is selected as the match for most Coastal Wetland parcels in NJ. However, it is worth noticing that this matching does not necessarily indicate an actual similarity in land cover status between Coastal Wetland NJ and Evergreen Forest DE, but more a result of the mechanism of SVM classifier, which on the parcel level constrainedly match a parcel to one and only one LULC category. For many Coastal Wetland parcels in NJ, Evergreen Forest DE is merely a reluctant choice, because SVM classifier constrainedly match parcel to a category and Evergreen Forest DE is the least different.

5.5 Conclusions

In this chapter, in order to reduce user side non-conformance caused by limitation of remote sensing and procedure errors, we refine the input data of the integration method to only include parcels that are large enough to contain 50 or more Landsat ETM+ pixels.

Although several urban LULC categories are disqualified from matching process because they do not have enough large parcels, all major forest and wetland categories, which are of special interest in environment analysis and resource management, contain qualified large parcels and remain in the matching process. For most of the classification systems, using large parcels increased the conformance rate and/or precision. The largestimprovement in conformance rate (6%) happens in the experiments of matching NJ LULC to DE LULC, and the largestimproviment in precision (12.5%) happens when matching DE LULC to MD LULC. This performance improvement attributes to reducing user side non-conformance, and consequently the semantic heterogeneity is further exposed.

As we can see in original matching results, the semantic relation between LULC categories is complicated by semantic heterogeneities. In original matching results, rather than simple subsuming, related LULC categories tend to overlap and intertwine together because of user side non-conformance, and this makes a straightforward hierarchical semantic integration difficult to extract. By using only large parcels, we effectively refined parcel level statistics and made a more reliable data representation of LULC categories, based on which the extraction of underlining semantic relations is achieved. Although semantic relations cannot be built between every two urban categories because of lack of large parcels in urban areas, the semantic integration of major forest and wetland categories were achieved. This integration is of great importance in the study of LULC, because it enables the indispensible data interoperability that supports the regional environment analysis and resource management.

Chapter 6: Integrating Lexical Semantic and Remote Sensing Results

6.1 Necessity of Integrating Results

In chapter 1, we introduced the view of ontology that it is built on universals (Smith 2004), and pointed out LULC categories are universals. A LULC category, on one hand implies semantics as a real world concept, on the other hand is populated by individual parcels of its kind, which are directly monitored by modern earth observing technologies, such as remote sensing. Hence, we adapt figure 1.3 to the integration of LULC classifications as figure 6.1, in which an intensional method uses descriptions of LULC categories and an extensional method uses the remote sensing information attached to parcels. These two methods are implemented in chapter 2 and chapter 3 respectively.

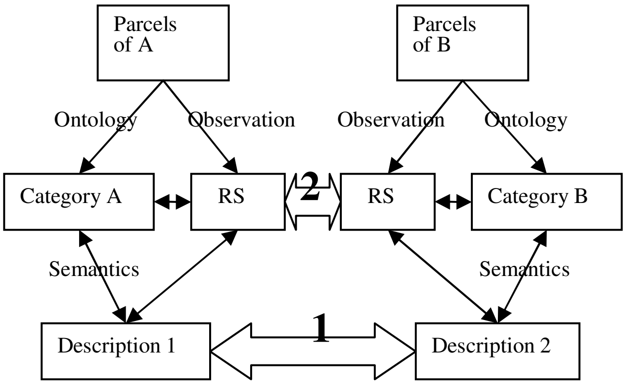


Figure 6.1 Semantic integration approaches in LULC classifications: intensional approach (1) and extensional approach (2)

Textual description and remote sensing are two different angles to understanding a LULC category, and as discussed in chapter 2 and 3, methods using either one of the two data sources are capable of matching LULC classifications at a satisfactory accuracy. But from the matching results, we found either method has limitations that can be compensated by the other. Specifically, lexical semantic method has difficulties with semantic heterogeneities, including naming conflicts and confounding conflicts, where remote sensing can serve as an independent source to discriminate.

On the other hand, having an independent source is also important to the remote sensing approach. Remote sensing observes ground land cover, which is the surrogate of actual land use. Uncertainty in this surrogacy leads to the limitation of remote sensing based integration. Differentiating some LULC categories (e.g. Farmstead and Feeding Lot) is beyond the capability of remote sensing, while the most naïve natural language processing (e.g. string comparison) technique is adequate to tell the difference and match them correctly.

Recognizing the necessity of combining the two information sources, in this chapter, we will discuss different approaches to integrate the two information sources to improving the matching of LULC classification systems.

6.2 A Simple But Effective Approach – Weighted Sum

As explained in section 2.1.1, semantic integration aims to determine the relations between concepts, and this is based on the measurement of their semantic similarities (Euzenat and Shvaiko 2007). The integration methods introduced in chapter 2 and 3, although using different data sources and methodology, have their outputs in the same form of similarity values. Hence, integration of the two methods can be achieved by employing compound similarity.

Compound similarity is concerned with the aggregation of different similarities (Euzenat and Shvaiko 2007). In chapter 5 of this book, Euzenat and Shvaiko summarized several different strategies to aggregate dissimilarities or distances. It is worth noticing these strategies can also be used to aggregate similarities. One of the most common families of distances is the Minkowski distance (Kruskal 1964). Comparing to other distances, Minkowski distance is well suited to independent dimensions and tend to balance the values between dimensions (Euzenat and Shvaiko 2007). The definition of Minkowski distance is as follows

$$\forall x, x' \in o, Minkowski(x, x') = \sqrt[p]{\sum_{i=1}^{n} \delta(x_i, x_i')^p} \ ,$$

in which o is a set of objects which can be analyzed in n dimensions, and $\delta(x_i,x_i')$ is the distance between two objects along the dimension i. Minkowski distance is a generalization of the widely used Euclidean distance (when p=2) and Manhattan distance (when p=1).

In some circumstances, several dimensions are more important than others. Their importance can be reflected in higher weights assigned to corresponding dimensions. By assigning weights to each dimension in Manhattan distance, we get weighted sum, defined as follows

$$\forall x, x' \in o, weighted\_sum(x, x') = \sum_{i=1}^{n} w_i \times \delta(x_i, x_i'),$$

in which $\delta(x_i,x_i')$ is the distance between two objects along the dimension i, and w_i is the weight of that dimension.

6.2.1 Aggregated Matching Results

As an instance of the Minkowski distances, it is important to reiterate that the weighted sum is applicable only when 1) the objects to be aggregated are in exactly the same unit and 2) the dimensions are independent.

In this research, we want to aggregate similarity values between LULC categories measured by two different methods. These two methods use different information sources (lexical semantics and remote sensing), and therefore their resultant similarity values are independent. The requirement on independence is hereby fulfilled. As for the first requirement, similarity values are unitless, but to ensure the comparability, the similarity from lexical semantics is normalized using the same strategy as in the remote sensing approach: the similarity values of matching one source category to different target categories are normalized, so that they sum up to 1.

Under this normalization strategy in remote sensing approach, the similarity of comparing A to B is not necessarily equal to the similarity of comparing B to A. This non-commutativity may seem paradoxical at first sight, but it reflects the complexity in real world LULC classifications. Previous analysis in section 3.3.5 shows that non-commutative similarity values between two LULC categories are the result of the difference in categories' conceptual scopes or levels in semantic hierarchy. For example, when matching DE LULC to MD LULC, a majority of Evergreen Forest parcels in DE matched to Evergreen Forest MD, but when matching MD LULC to DE LULC, only half of the Evergreen Forest parcels in MD matched to Evergreen Forest DE, while the rest matched to Shrub Rangeland DE or Mixed Forest DE.

Accordingly, the similarity value of comparing Evergreen Forest DE to Evergreen Forest MD, which is decided by the ratio of parcels in this match to all parcels for category Evergreen Forest DE, is 0.73, and the similarity is 0.42 when comparing Evergreen Forest MD to Evergreen Forest DE. Combining these two pieces of information, we can make estimation that Evergreen Forest DE is subsumed by Evergreen Forest MD. Actually, only based on non-commutative similarity calculations, we can discover the hierarchical semantic relationship between LULC categories.

As introduced in chapter 2, there are several algorithm variations of the lexical approach. We selected the method using Latent Semantic Analysis with keyword enhanced (KE-LSA), because it has a consistent good performance in all 4 experiments. Actually, our experiments showed the choice of algorithm does not impact the results very much (less than 5%), because all variations of lexical methods generate highly similar results, but greatly different from the results of remote sensing based methods.

Initially, we assign a same weight of 0.5 to the similarities calculated using lexical semantics (hereafter noted as SIM_{Sem}) and remote sensing (hereafter noted as SIM_{RS}). The matching result is presented in Table 6.1 a (DE LULC to MD LULC), b (MD LULC to DE LULC), c (NJ LULC to MD LULC) and d (NJ LULC to DE LULC). In four tables, categories are denoted by codes introduced in Appendix I. The columns from left to right mean 1) Source LULC categories, 2) match(es) in target LULC by human evaluators, 3) conforming matches (algorithm and human match same), and 4) non-conforming matches (algorithm and human match different). The number in the parentheses after each algorithm result (in columns 3 and 4) is the similarity of that match, calculated from weighted sum of SIM_{RS} and SIM_{Sem}. As before, matches with a compound similarity less than 0.1 are considered less important and discarded from the table.

Table 6.1 Matching Results Using Weighted Sum

a

DE LULC	MATCHED MD LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING

SinFam	LowRes
	MedRes
		LowRes(0.330)
		MedRes(0.281)	

MultFam	MedRes
	HighRes
		MedRes(0.181)
		HighRes(0.278)
			Comm(0.165) LowRes(0.280)

MblHm	HighRes	HighRes(0.321)	OpenUrb(0.131) Indstrl(0.163) MedRes(0.109)
Retail	Comm	Comm(0.671)	Indstrl(0.206)

VclAct	Indstrl
	Comm
		Comm(0.134)
		Indstrl(0.347)
			Extr(0.372)

JunkYrd	Indstrl
	Comm
		Indstrl(0.330)
		Comm(0.267)
			MedRes(0.125)

Warehs	Indstrl
	Comm
		Comm(0.173)
		Indstrl(0.541)
			Inst(0.124)

OthrCom	Comm	Comm(0.419)	AgrBldg(0.102) OrchHrt(0.110) Indstrl(0.157)
Indstrl	Indstrl	Indstrl(0.821)	Comm(0.136)
Utility	OpenUrb Indstrl	Indstrl(0.222)	Pasture(0.122) Comm(0.129)

MixUrb	All Urban but OpenUrb	Inst(0.141)
		Comm(0.194)
			MxFrst(0.139) OpenUrb(0.173)

OthrUrb	BrGrnd OpenUrb	OpenUrb(0.262)	Inst(0.147) AgrBldg(0.186)
Inst	Inst	Inst(0.528)	Indstrl(0.162) Comm(0.129)
Recreat	OpenUrb		Comm(0.128) AgrBldg(0.128)
Crop	Crop	Crop(0.713)	
Pasture	Pasture	Pasture(0.584)	Crop(0.157)
IdleFld	OpenUrb Brush	Brush(0.400)	LowRes(0.104) Crop(0.112) Pasture(0.160)
OrchHrt	OrchHrt	OrchHrt(0.482)	Crop(0.111)
Feedlot	FeedOp	FeedOp(0.625)	Indstrl(0.190)
Frmstd	AgrBldg	AgrBldg(0.473)	Inst(0.100)
OthrAgr	Crop AgrBldg		Indstrl(0.131) FeedOp(0.239) Brush(0.445)
HerbRng	Pasture Brush	Pasture(0.140)	
ShrbRng	Brush Pasture	Brush(0.477)	LowRes(0.115) DeciF(0.129)
MixRng	Brush Pasture		LowRes(0.144) MxFrst(0.436)
DecFrst	DeciF	DeciF(0.507)	MxFrst(0.211)
EvrgrnF	EvrgrnF	EvrgrnF(0.540)	MxFrst(0.237) DeciF(0.114)
MixFrst	MxFrst	MxFrst(0.406)	EvrgrnF(0.147) DeciF(0.250)
ClrCut	BrGrnd Brush	Brush(0.506)	Pasture(0.101) Crop(0.197)
Watrway	Water	Water(0.795)	

DE LULC	MATCHED MD LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
NtrlLk	Water	Water(0.466)	EvrgrnF(0.166) AgrBldg(0.172)
Rsrvr	Water	Water(0.809)	
BayCove	Water	Water(0.888)	
Wetland	Wetland	Wetland(0.546)	DeciF(0.162)
Beach	Beach	Beach(0.694)	Water(0.142) Indstrl(0.144)
InldSnd	BrGrnd		Crop(0.108) Indstrl(0.3) Beach(0.275)
Extr	Extr		Indstrl(0.365) Water(0.154) Crop(0.120)
Trans	Crop Brush BrGrnd		FeedOp(0.121) GdnCrop(0.116) OpenUrb(0.102) Indstrl(0.187)

b

MD LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
LowRes	SinFam	SinFam(0.398)	MultFam(0.188)

MedRes	MultFam
	SinFam
		SinFam(0.432)
		MultFam(0.189)	

HighRes	MultFam
	MblHm
		MblHm(0.312)
		MultFam(0.230)
			SinFam(0.234)

Comm	Retail
	OthrCom
	MixUrb
		OthrCom(0.136)
		Retail(0.234)	

Indstrl	Indstrl
	JunkYrd
	Warehs
		Warehs(0.165)
		JunkYrd(0.122)
		Indstrl(0.302)	

Inst	Inst	Inst(0.287)	MixUrb(0.109)
Extr	Extr	Extr(0.107)	VclAct(0.304) Feedlot(0.134)
OpenUrb	Recreat		SinFam(0.106) MixUrb(0.235) OthrUrb(0.251)
Crop	Crop OthrAgr	Crop(0.418)	TruckCrp(0.178)
Pasture	Pasture	Pasture(0.538)	
OrchHrt	OrchHrt	OrchHrt(0.383)	
FeedOp	Feedlot	Feedlot(0.610)	
AgrBldg	Frmstd	Frmstd(0.251)	
GdnCrop	TruckCrp	TruckCrp(0.434)	Crop(0.112)
DeciF	DecFrst	DecFrst(0.312)	MixFrst(0.182) EvrgrnF(0.150)
EvrgrnF	EvrgrnF	EvrgrnF(0.421)	MixFrst(0.165)
MxFrst	MixFrst	MixFrst(0.285)	MixRng(0.150) DecFrst(0.120) EvrgrnF(0.149)

Brush	ShrbRng
	HerbRng
	MixRng
	IdleFld
	ClrCut
		ClrCut(0.209)
		IdleFld(0.102)
		ShrbRng(0.193)	

Water	Watrway
	Rsrvr NtrlLk
		Watrway(0.125)
		Rsrvr(0.283)	

MD LULC	**MATCHED DE LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
	BayCove	BayCove(0.221)	
Wetland	Wetland	Wetland(0.418)	MixRng(0.121)
Beach	Beach	Beach(0.408)	Retail(0.1) BayCove(0.1) InldSnd(0.126) Indstrl(0.1)
BrGrnd	InldSnd VclAct		Extr(0.284)

c

NJ LULC	**MATCHED MD LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
HighRes	HighRes	HighRes(0.386)	Comm(0.100) MedRes(0.132) LowRes(0.115)
MedRes	MedRes	MedRes(0.269)	LowRes(0.128) HighRes(0.193)
LowRes	LowRes	LowRes(0.189)	HighRes(0.128) MedRes(0.231)
RurlRes	LowRes	LowRes(0.174)	MedRes(0.159) Brush(0.110)
MixRes	LowRes MedRes HighRes		OpenUrb(0.303) Wetland(0.260)
Comm	Comm	Comm(0.291)	HighRes(0.164) Indstrl(0.137)
Milit	Inst		Indstrl(0.332)
Indstrl	Indstrl	Indstrl(0.385)	Comm(0.134)
Transp			Indstrl(0.153) Comm(0.116)
Road			Comm(0.126) HighRes(0.209) MedRes(0.110)
Bridge			Water(0.719)
Airport			AgrBldg(0.116) Indstrl(0.244)

WtlndWa	Brush
	Wetland
		Brush(0.230)
		Wetland(0.161)	

UpldWaD			OpenUrb(0.146) Comm(0.127)
UpldWa	Brush	Brush(0.203)	OpenUrb(0.103)
StrmBas			Indstrl(0.112) HighRes(0.103) Comm(0.176)
ICCmplx	Indstrl Comm		HighRes(0.510)

MixUrb	LowRes
	MedRes
	HighRes
	Comm
	Indstrl Inst
		Comm(0.192)
		HighRes(0.266)
			OpenUrb(0.122)

OthrUrb	BrGrnd OpenUrb	OpenUrb(0.122)	HighRes(0.100) Comm(0.151)
Cemet	OpenUrb	OpenUrb(0.168)	
WtCemet	OpenUrb Wetland	OpenUrb(0.104)	AgrBldg(0.272) Pasture(0.259)
Phrg			OpenUrb(0.144) Wetland(0.262) HighRes(0.132) Comm(0.174)

NJ LULC	MATCHED MD LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
MngWtld	Pasture Wetland	Pasture(0.117)	OpenUrb(0.101)
Recreat	OpenUrb	OpenUrb(0.122)	Inst(0.104) Comm(0.124) Indstrl(0.123)
Athlet	Inst	Inst(0.158)	Indstrl(0.186)
Stadium	Inst	Inst(0.116)	HighRes(0.138) Indstrl(0.282)
MngWtRe	OpenUrb Wetland	OpenUrb(0.132)	

CrpPstr	Crop Pasture	Pasture(0.149)
		Crop(0.224)	

AgriWet	Crop Pasture Wetland	Pasture(0.112)	AgrBldg(0.154)
FmAgrWt	Wetland Brush	Brush(0.123)	LowRes(0.120) Pasture(0.118)
OrchHrt	OrchHrt	OrchHrt(0.151)	AgrBldg(0.109)
FeedOp	FeedOp	FeedOp(0.135)	Indstrl(0.192)
OthrAgr	FeedOp AgrBldg GdnCrop	AgrBldg(0.148)	Brush(0.107) Indstrl(0.109)

DeciF10	DeciF Brush	DeciF(0.134)
		Brush(0.205)
			MxFrst(0.137)

DeciF50	DeciF	DeciF(0.256)	MxFrst(0.208) Brush(0.196)

ConiF10	EvrgrnF
	Brush
		EvrgrnF(0.262)
		Brush(0.242)
			DeciF(0.134)

ConiF50	EvrgrnF	EvrgrnF(0.425)	Brush(0.107) DeciF(0.126)
Plant	OrchHrt EvrgrnF	EvrgrnF(0.436)	Brush(0.180)

MxCnF10	MxFrst
	EvrgrnF
	Brush
		MxFrst(0.208)
		EvrgrnF(0.200)
		Brush(0.256)
			DeciF(0.134)

MxCon50	MxFrst
	EvrgrnF
		EvrgrnF(0.291)
		MxFrst(0.193)
			DeciF(0.115) Brush(0.207)

MxDec10	MxFrst
	DeciF Brush
		DeciF(0.172)
		Brush(0.275)
		MxFrst(0.214)
			EvrgrnF(0.137)

MxDec50	MxFrst
	DeciF
		MxFrst(0.207)
		DeciF(0.165)
			EvrgrnF(0.152) Brush(0.289)

OldFld	Brush	Brush(0.182)	
PhrgOld	Brush	Brush(0.127)	Wetland(0.268)
DecBrsh	Brush	Brush(0.214)	DeciF(0.126)
ConBrsh	Brush	Brush(0.196)	EvrgrnF(0.175)
MxBrush	Brush	Brush(0.235)	DeciF(0.142) MxFrst(0.154)
BrUplnd	BrGrnd		Brush(0.283) EvrgrnF(0.279)
Stream	Water	Water(0.155)	EvrgrnF(0.215) Wetland(0.202) Brush(0.124)
NatLake	Water	Water(0.264)	Wetland(0.254)
Rsrvr	Water	Water(0.506)	

TdlRiv	Water
	Wetland
		Water(0.496)
		Wetland(0.324)	

TdlBay	Water	Water(0.503)	Wetland(0.224)
Dredge	Water	Water(0.227)	HighRes(0.120) Wetland(0.286)

NJ LULC	MATCHED MD LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
Ocean	Water	Water(0.545)	Wetland(0.118)
SlMrsh	Wetland	Wetland(0.370)	Water(0.123)
SlMrshV	Wetland	Wetland(0.346)	Brush(0.104)
FrMrsh	Wetland	Wetland(0.494)	
VegDune	Wetland	Wetland(0.122)	Indstrl(0.216)
PhrgCWt	Wetland	Wetland(0.465)	

DecWdWt	Wetland
	DeciF
		Wetland(0.171)
		DeciF(0.176)
			MxFrst(0.138) Brush(0.170) EvrgrnF(0.173)

ConWdWt	Wetland EvrgrnF	EvrgrnF(0.483)	

CedarWt	EvrgrnF
	Wetland
		EvrgrnF(0.624)
		Wetland(0.187)	

DecBrWt	Brush
	Wetland
		Wetland(0.200)
		Brush(0.200)
			DeciF(0.172) MxFrst(0.100) EvrgrnF(0.159)

ConBrWt	Brush
	Wetland
		Wetland(0.215)
		Brush(0.197)
			DeciF(0.103) EvrgrnF(0.335)

MxBrWtD	DeciF
	MxFrst
	Wetland
		DeciF(0.159)
		Wetland(0.179)
		MxFrst(0.196)
			EvrgrnF(0.213) Brush(0.144)

MxBrWtC	EvrgrnF
	MxFrst
	Wetland
		MxFrst(0.201)
		Wetland(0.231)
		EvrgrnF(0.250)
			Brush(0.142)

HrbWtNT	Brush
	Wetland
		Wetland(0.326)
		Brush(0.146)	

PhrgWet	Wetland	Wetland(0.358)	Brush(0.105) EvrgrnF(0.101)

MxFrWtD	DeciF
	MxFrst
	Wetland
		DeciF(0.148)
		MxFrst(0.201)
		Wetland(0.155)
			EvrgrnF(0.312) Brush(0.139)

MxFrWtC	EvrgrnF
	MxFrst
	Wetland
		MxFrst(0.180)
		Wetland(0.178)
		EvrgrnF(0.416)	

BrndWet	BrGrnd Wetland	Wetland(0.314)	EvrgrnF(0.270)
Beach	Beach	Beach(0.385)	Indstrl(0.119) Water(0.175)
BrGrnd	BrGrnd	BrGrnd(0.150)	
Extr	Extr	Extr(0.130)	Indstrl(0.325) Beach(0.104)
AltLnd			OpenUrb(0.129) Indstrl(0.155) Comm(0.156)
DstrbWt	Wetland	Wetland(0.224)	Brush(0.104)
Transi	OpenUrb BrGrnd		Indstrl(0.231)
Barren	BrGrnd		Indstrl(0.364)

d

NJ LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
HighRes	MultFam	MultFam(0.402)	SinFam(0.215)

NJ LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
	MblHm		
MedRes	SinFam MblHm	SinFam(0.277)	OthrUrb(0.130) MixUrb(0.120)
LowRes	SinFam	SinFam(0.368)	MixRng(0.114)
RurlRes	SinFam	SinFam(0.382)	
MixRes	SinFam MultFam MblHm	MblHm(0.251)	MixUrb(0.114) Watrway(0.250) MixRng(0.199) MixFrst(0.163)

Comm	Retail
	VclAct
	Warehs
	OthrCom
		OthrCom(0.275)
		Retail(0.145)
			MultFam(0.164)

Milit	Inst		Retail(0.238)
Indstrl	Indstrl	Indstrl(0.583)	
Transp			Utility(0.359)
Road			OthrCom(0.268) Indstrl(0.242) MultFam(0.236)
Bridge			BayCove(0.5)
Airport			Frmstd(0.133) OthrUrb(0.116) Warehs(0.184) Indstrl(0.101)
WtlndWa	Wetland	Wetland(0.265)	NtrlLk(0.118) InldSnd(0.172) ShrbRng(0.164)
UpldWaD			ClrCut(0.195) MixUrb(0.158)
UpldWa	ShrbRng MixRng	ShrbRng(0.181)	InldSnd(0.124) ClrCut(0.179)
StrmBas			OthrCom(0.434)

ICCmplx	Indstrl
	OthrCom
		OthrCom(0.189)
		Indstrl(0.176)
			MultFam(0.500)

MixUrb	MixUrb	MixUrb(0.168)	OthrUrb(0.116) MultFam(0.245)
OthrUrb	OthrUrb	OthrUrb(0.199)	OthrCom(0.104) MixUrb(0.172)
Cemet	OthrUrb	OthrUrb(0.139)	MultFam(0.117) SinFam(0.159)
WtCemet	OthrUrb Wetland	Wetland(0.247)	OrchHrt(0.25) Frmstd(0.25)
Phrg			Utility(0.125) MultFam(0.25) MblHm(0.128) OthrUrb(0.212) MixUrb(0.174)
MngWtld	Pasture HerbRng Wetland	Wetland(0.277)	
Recreat	Recreat	Recreat(0.375)	
Athlet	Inst		Recreat(0.131) IdleFld(0.229)
Stadium	Inst		Frmstd(0.223) OthrCom(0.155) MblHm(0.155) Indstrl(0.131)

MngWtRe	Wetland
	Recreat
		Wetland(0.185)
		Recreat(0.158)
			MixUrb(0.109)

CrpPstr	Crop Pasture	Crop(0.436)	TruckCrp(0.104)
AgriWet	Crop Pasture Wetland	Wetland(0.310)	InldSnd(0.102)

FmAgrWt	IdleFld
	Wetland
	HerbRng
	ShrbRng
		Wetland(0.213)
		IdleFld(0.1)
		ShrbRng(0.173)
			MixUrb(0.115) OthrUrb(0.144)

NJ LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
	MixRng		
OrchHrt	OrchHrt	OrchHrt(0.355)	
FeedOp	Feedlot	Feedlot(0.502)	Frmstd(0.181)
OthrAgr	OrchHrt Feedlot OthrAgr	OthrAgr(0.307)	IdleFld(0.184)

DeciF10	DecFrst
	ShrbRng
	MixRng
		ShrbRng(0.100)
		DecFrst(0.225)
			MixFrst(0.134) EvrgrnF(0.153)

DeciF50	DecFrst	DecFrst(0.278)	Wetland(0.129) EvrgrnF(0.142) MixFrst(0.140) ShrbRng(0.108)
ConiF10	EvrgrnF ShrbRng MixRng	EvrgrnF(0.328)	DecFrst(0.128)
ConiF50	EvrgrnF	EvrgrnF(0.481)	DecFrst(0.122)

Plant	OrchHrt
	EvrgrnF
		EvrgrnF(0.379)
		OrchHrt(0.171)
			OthrAgr(0.280)

MxCnF10	MixFrst
	EvrgrnF
		EvrgrnF(0.169)
		MixFrst(0.165)
			DecFrst(0.119) Wetland(0.141) MixRng(0.176)

MxCon50	MixFrst
	EvrgrnF
		EvrgrnF(0.245)
		MixFrst(0.157)
			DecFrst(0.113) MixRng(0.152) Wetland(0.143)

MxDec10	MixFrst
	DecFrst
	ShrbRng
	MixRng
		MixFrst(0.141)
		DecFrst(0.179)
		MixRng(0.162)
			Wetland(0.179) EvrgrnF(0.102)

MxDec50	MixFrst
	DecFrst
		MixFrst(0.145)
		DecFrst(0.175)
			EvrgrnF(0.109) Wetland(0.216) MixRng(0.146)

OldFld	HerbRng MixRng ClrCut		ShrbRng(0.419) IdleFld(0.148)
PhrgOld	HerbRng MixRng	MixRng(0.174)	ShrbRng(0.182) IdleFld(0.205)
DecBrsh	ShrbRng MixRng	ShrbRng(0.322)	DecFrst(0.158)
ConBrsh	ShrbRng MixRng	ShrbRng(0.159)	NtrlLk(0.108) EvrgrnF(0.156) InldSnd(0.160)

MxBrush	HerbRng
	ShrbRng
	MixRng
		ShrbRng(0.197)
		MixRng(0.171)	

BrUplnd	ClrCut Trans		MixFrst(0.131) EvrgrnF(0.319) Wetland(0.208) DecFrst(0.135)
Stream	Watrway	Watrway(0.250)	EvrgrnF(0.125) Wetland(0.166) NtrlLk(0.145)
NatLake	NtrlLk	NtrlLk(0.421)	InldSnd(0.158) Rsrvr(0.189)
Rsrvr	Rsrvr	Rsrvr(0.322)	BayCove(0.102) NtrlLk(0.224) OthrAgr(0.107)
TdlRiv	BayCove Wetland	BayCove(0.483)	Beach(0.105) Watrway(0.118)
TdlBay	BayCove	BayCove(0.780)	
Dredge	Rsrvr Watrway	Watrway(0.525)	Wetland(0.136) MblHm(0.190)
Ocean	BayCove	BayCove(0.583)	Beach(0.175)
SlMrsh	Wetland		MblHm(0.132) BayCove(0.300)
SlMrshV	Wetland		Watrway(0.131)

NJ LULC	MATCHED DE LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
FrMrsh	Wetland	Wetland(0.256)	Watrway(0.113) Beach(0.215)
VegDune	Wetland InldSnd	InldSnd(0.138)	Beach(0.410)
PhrgCWt	Wetland	Wetland(0.424)	Watrway(0.118)

DecWdWt	Wetland
	DecFrst
		Wetland(0.449)
		DecFrst(0.150)
			MixRng(0.120)

ConWdWt	Wetland
	EvrgrnF
		EvrgrnF(0.391)
		Wetland(0.335)	

CedarWt	EvrgrnF
	Wetland
		EvrgrnF(0.446)
		Wetland(0.410)	

DecBrWt	ShrbRng
	Wetland
	MixRng
		Wetland(0.233)
		ShrbRng(0.245)
			EvrgrnF(0.108) DecFrst(0.147)

ConBrWt	ShrbRng
	Wetland
	MixRng
		Wetland(0.205)
		ShrbRng(0.311)
			EvrgrnF(0.274)

MxBrWtD	ShrbRng
	Wetland
	DecFrst
	MixRng
		DecFrst(0.119)
		Wetland(0.228)
		ShrbRng(0.138)
		MixRng(0.146)
			EvrgrnF(0.151)

MxBrWtC	ShrbRng
	Wetland
	EvrgrnF
	MixRng
		ShrbRng(0.148)
		Wetland(0.248)
		EvrgrnF(0.185)
		MixRng(0.178)	

HrbWtNT	Wetland HerbRng	Wetland(0.317)	NtrlLk(0.105) IdleFld(0.132)
PhrgWet	HerbRng Wetland	Wetland(0.467)	MixRng(0.103)

MxFrWtD	Wetland
	DecFrst
	MixFrst
		DecFrst(0.131)
		Wetland(0.372)
			EvrgrnF(0.151) MixRng(0.109)

MxFrWtC	Wetland
	EvrgrnF
	MixFrst
		Wetland(0.323)
		EvrgrnF(0.274)
			MixRng(0.111)

BrndWet	HerbRng Wetland	Wetland(0.679)	NtrlLk(0.250)
Beach	Beach	Beach(0.395)	Indstrl(0.125) BayCove(0.158) InldSnd(0.104)
BrGrnd			JunkYrd(0.181)
Extr	Extr	Extr(0.137)	Feedlot(0.139) Frmstd(0.165) VclAct(0.231)
AltLnd			Rsrvr(0.109) VclAct(0.211)
DstrbWt	Wetland	Wetland(0.229)	Trans(0.168)
Transi	Trans	Trans(0.367)	Indstrl(0.117)
Barren	InldSnd	InldSnd(0.118)	Extr(0.119) VclAct(0.435)

As shown in Table 6.2 and Figure 6.1, the matching result of weighted sum (WtdSum) has better conformance rate (Conf) and precision (Prec) (defined in section 4.3.2) than those of either the remote sensing based method (RS) or the lexical semantics based method (KE-LSA). In the experiments of matching MD LULC to DE LULC, NJ LULC to MD LULC, and NJ LULC to DE LULC, the conformance rate is higher than 90% with a precision at about 50%. This means, combining lexical semantics and remote sensing, our method is capable of finding correct matches for 90% of the categories at a precision of 50%. As a comparison, in its best performed experiment, previous feature-based method (Zhou and Wei 2008) finds correct matches for 50% of the categories at a precision of 30%. Also included is the single measure named F measure that trades off precision versus conformance rate. It is defined as the harmonic mean of conformance rate and precision

$$F = \frac{2PC}{P+C},$$

where *P* is the precision and *C* is the conformance rate(Manning et al. 2008). The F measure confirmed the improvement of performance. Comparing to average the F measure values among human evaluations (0.90 for DE2MD, 0.89 for MD2DE), weighted sum strategy still need substantial improvement.But for the experiment of matching MD LULC to DE LULC, the difference of F measure between weighted sum (0.70) and least agreed evaluation (0.77) is reduced.

As explained in 6.1, from a theoretical point of view, we anticipate the potential advantages from combining two independent data sources. The advantage is two-fold. First, two data sources can be mutually complementary: the missing pieces in one source can be picked up by the other, and the integration should therefore be more complete. Second, as data sources contain noises, having several independent data sources enables the mutual validation, which is important to separate information from noise. Intuitively, same piece of information extracted from multiple independent data sources are much more likely to be true, while a piece of information extracted from one data source but disapproved by other sources more likely turns out to be noise.

Table 6.2 Performance of Weighted Sum

	Conformance Rate			Precision			F Measure		
	KE-LSA	RS	WtdSum	KE-LSA	RS	WtdSum	KE-LSA	RS	WtdSum
DE2MD	0.78	0.84	0.84	0.40	0.35	0.37	0.53	0.50	0.52
MD2DE	0.86	0.82	0.91	0.46	0.34	0.57	0.60	0.48	0.70
NJ2MD	0.92	0.77	0.92	0.48	0.38	0.50	0.63	0.51	0.64
NJ2DE	0.81	0.70	0.90	0.37	0.40	0.45	0.51	0.51	0.59

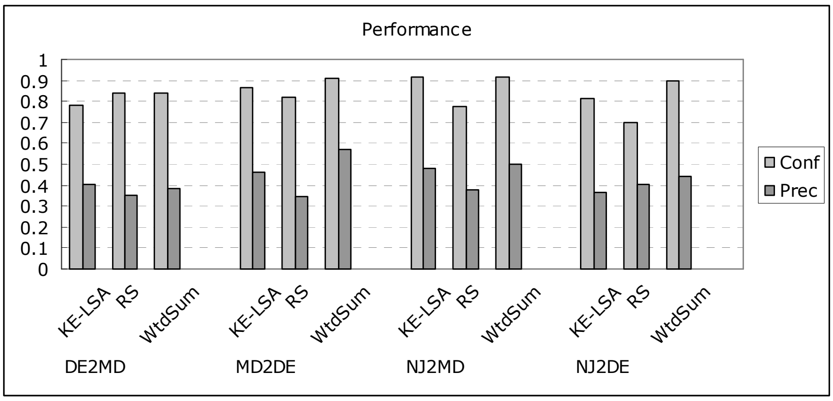


Figure 6.2 Performance of Weighted Sum

By examining the matching result, we find weighted sum, although plain and simple, is capable of realizing these theoretical advantages, and gives better performance than using lexical semantics or remote sensing alone (table 6.2). For example, when matching NJ LULC to MD LULC, the lexical semantic method missed the match from Saline Marshes (Low marsh vegetation) to Wetland. But in remotely sensed data, parcels of these two categories have very similar spectral responses and therefore are matched together at high similarity value. When calculating the weighted sum, despite the low SIM_{Sem}, Saline Marshes (Low marsh vegetation) NJ and Wetland MD are correctly matched together because of the high SIM_{RS}. This is one of several examples of how combining independent data sources can further complete the matching results.

At the same time, we observe more cases in which combining data sources demonstrates the advantage of denoising matching results. For example, when matching NJ LULC to DE LULC, although both lexical semantic method and remote sensing method correctly match category Orchards, Vineyards, Nurseries, Horticultural Areas, and Sod Farms of NJ LULC to Orchards/Nurseries/Horticulture of DE LULC, they both mistakenly include several mismatches, such as Truck Crops picked up by lexical method and Cropland and Pasture picked up by remote sensing method. It is worth noticing mismatching orchards to cropland or pasture in remote sensing is not very surprising, and it is already "warned" in some early work (Anderson 1976): "many of these (Orchard) areas may be included in another category, generally Cropland and Pasture, when identification is made by use of small-scale imagery alone." Anderson also pointed out that "identification (of Orchards) may be aided by recognition of the combination of soil qualities, topography, and local climatological factors needed for these operations." In this research, instead of soil, topography, and local climatology, we will use lexical semantics to aid the separation and improve the accuracy. Then in the matching result for orchard using weighted sum, Truck Crops, Cropland, and Pasture are all excluded as their compound similarities are below the threshold.

From Table 6.2 and Figure 6.2, we can see an obvious improved performance in the matching results, and concludes that weighted sum is an effective strategy to aggregate similarities from different measurements, even without any optimization on weights. In the next section, we will try to figure out whether there is a way to optimize the weights to further improve the matching results.

6.2.2 Optimizing Weights

In 6.2.1, we discussed the matching results using the weighted sum of SIM_{RS} and SIM_{Sem}. Using evenly assigned weight of 0.5, the compound similarity leads to matching results that have an obvious improvement over the results from either of the two. But is there a space for further improvement by adjusting the weights?

Intuitively, we want to assign a higher weight to the similarity leading to better performance. But as shown in previous chapters, both similarity measurements have limitations, and their performance will fluctuate among different LULC categories in different classification systems. Therefore, a fixed or global weighting of the two similarities is not feasible. Instead, a good weighting scheme should be specific to each category.

By comparing to human evaluation, we define Net Conforming Similarity (NCS) to measure a method's performance on each LULC category. For each given category c_i, the equation is defined as follows

$$NCS(c_i) = \sum_{c_j \in C_{conf}} s_{ij} - \sum_{c_j \in C_{non}} \frac{s_{ij}}{2^{l_j}},$$

where C_{conf} is the collection of all conforming matches of the given category c_i, C_{non} is the collection of all non-conforming matches, S_{ij} is the similarity for category c_i and its match category c_j, and l_j is the level of minimal common upper category of non-conforming match c_j and any one of the matches given by human evaluation. By its definition, given a LULC category NCS measures the conformity (with human evaluation) of matches found by an integration algorithm. For conforming matches, their contribution to NCS is positive, weighted by their similarity. For non-conforming matches, their contribution is negative, weighted by similarity and how deviant from human evaluation each non-conforming match is. For example, in table 6.3, for Multiple Family DE, algorithm found four matches in MD LULC, in which Medium Density Residential and High Density Residential are conforming to human evaluation and Commercial and Low Density Residential are not. The first part of NCS is the sum of conforming similarity, calculated as 0.181 plus 0.278 equals to 0.459. For the second part, we first find the minimal common upper category for each non-conforming match and matches from human evaluation. For category Commercial, it shares minimal common upper category Urban and Built-up with Medium Density Residential from human evaluation. Urban and Built-up is a level 1 category. Therefore category Commercial's contribution is its similarity 0.165 divided by 2^1 equals to 0.0825. As the minimal common upper category of Low Density Residential and either of the human matches is Residential, a level 2 category, its contribution is its similarity 0.280 divided by 2^2 equals 0.07. Then the NCS for Multiple Family Residential in DE LULC is calculated as 0.3065.

Table 6.3 Matching Multiple Family DE to MD LULC

DE LULC	**MATCHED MD LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**

MultFam	MedRes
	HighRes
		MedRes(0.181)
		HighRes(0.278)
			Comm(0.165)
			LowRes(0.280)

Now it seems we could assign different weight to SIM_{Sem} and SIM_{RS} according to their NCS values for each LULC category; higher weight goes to the similarity with a higher NCS value. However, in the process of matching LULC classification systems, complete human matching is not present; otherwise the matching process is redundant. The approach of weighting similarities by directly comparing their resultant matches to human evaluation is not feasible. Instead, we need to find some indicator to predict the performance of similarity measurements without human evaluations.

Data Approach

We attempted two approaches to find an indicator. First, we tried to summarize the indicator from experimental data. Within this approach, two hypotheses that may lead to indicators are made and tested. The first hypothesis is as a high similarity suggests more confidence from the algorithm in this match, it is more likely to be true (conforming to human evaluations). As all similarity values of matching a source category to all its targets sum up to 1, a higher similarity among leads to a higher variance. This hypothesis predicts that a higher variance in the similarity values indicates a better performance from the algorithm.

We tested this hypothesis in all the matching experiments, and the results only support this hypothesis to a limited extent. In Figure 6.3, x axis is the variance, y axis is the NCS, and each source category in matching represents a dot in the graph. There seems to be a faint correlation between the variance and NCS, but the correlation is not strong enough to be discovered statistically or to be used as an indicator of potential performance.

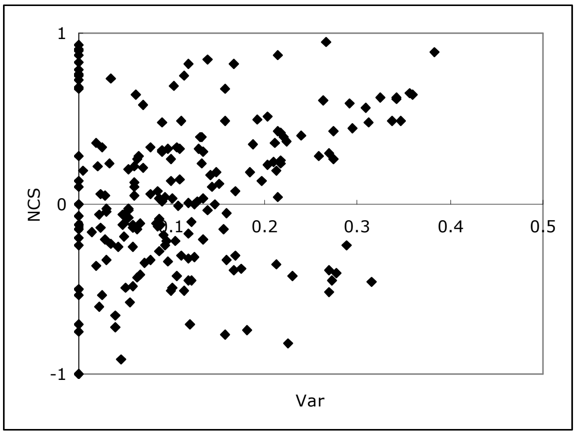


Figure 6.3 Distribution of NCS along variance

A second potential indicator is designed for SIM_{RS} only. In previous section 3.2.8, to find typical parcels to represent its category and serve as the input to train a SVM classifier, we first removed parcels that are outcasts in spectral space, and then performed a cross validation and keep those unaffected: a parcel's label assigned in the cross validation is same as its original label. Parcels that can survive these two filtering processes are considered representative and constitute the training set. Our hypothesis is then the higher percentage of representative parcels (survived the two steps), the better chance a similarity measurement could work on this category (high NCS).

However, this hypothesis is also difficult to be supported by experiments. In Figure 6.4, x axis is the percentage of representative parcels, y axis is the NCS, and each source category in matching represents a dot in the graph. As we can observe and quantitatively test, there is no obvious trend or correlation in the graph. Comparing to the variance of output similarities, the percentage of representative parcels is even less likely to be related to the performance. A possible explanation to this non-correlation is that the process of finding representative parcels is a "quality before quantity" process, and many less representative parcels that are not calculated in the percentage, are actually classified to conforming matches and still benefit NCS.

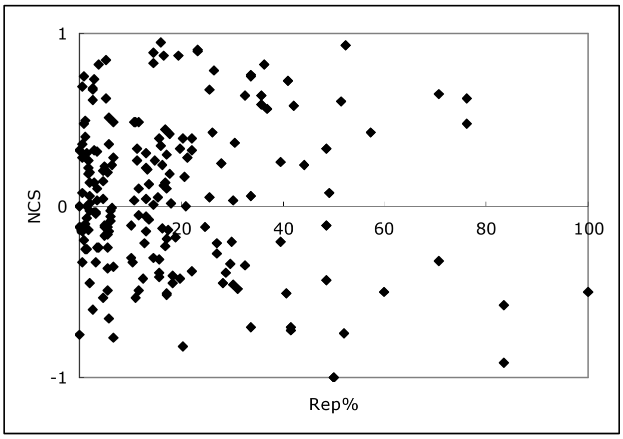


Figure 6.4 Distribution of NCS along the percentage of representative parcels

Based on the above discussion, trying to predict the performance by exploring input data or output similarities is difficult. However, at this point, we do not rule out the possibility that solid correlation might emerge in future matching attempts, but in this research, we need to find an alternative strategy to optimize the weights.

Domain Knowledge Approach

Different from data approach, a knowledge approach tries to apply domain knowledge to our weighting problem. As explained in previous chapters, remote sensing has a limitation when discriminating urban land uses, because the use of artificial parcels is usually difficult to decide from observations alone. At the same time, the semantic heterogeneity is a serious problem for non-urban categories, where a remote sensing method and a lexical semantic method are both indispensable. Based on this consideration, we make a hypothesis that urban land use categories are more accurately matched by lexical semantic methods, while the remote sensing method will contribute more on non-urban categories. Hence, we propose an uneven weighting scheme that will assign a higher weight to the SIM_{Sem} and a lower weight to SIM_{RS} when matching urban categories, and will assign the original weighting to both similarities when matching non-urban categories.

The experiment of matching the NJ LULC to the MD LULC shows the performance is improved when assigning the weight of 0.6 to SIM_{Sem} and the weight of 0.4 to SIM_{RS} for urban categories. We tested this weighting scheme in other experiments, and observed better performance in matching NJ LULC to DE LULC and DE LULC to MD LULC, but performance dropped when matching MD LULC to DE LULC (Table 6.4). In order to understand under what circumstances uneven weighting will work, we compare lexical semantic and remote sensing methods' performances on each category when matching NJ LULC to MD LULC and MD LULC to DE LULC.

Table 6.4 Performance using domain knowledge driven weighting (uneven) and even weighting

	Conformance rate		Precision		F Measure	
	Even	Uneven	Even	Uneven	Even	Uneven
DE2MD	0.84	0.84	0.37	0.38	0.52	0.53
MD2DE	0.91	0.86	0.57	0.5	0.70	0.63
NJ2MD	0.91	0.93	0.5	0.5	0.64	0.65
NJ2DE	0.9	0.9	0.44	0.45	0.59	0.6

Figure 6.5 compares the NCS values using two methods to match (a) NJ LULC to MD LULC and (b) MD LULC to DE LULC, in which a dark grey bar corresponds to an urban category and a light grey bar corresponds to a non-urban category. When the remote sensing method (RS) performs better than lexical semantic method (Sem), the bar is upward, and otherwise it is downward. The height of the bar corresponds to the difference of NCS between two methods on the category.

From the figures, we can see matching NJ LULC to MD LULC follows our assumption very well. The lexical semantic method has obvious advantages over the remote sensing method in matching urban categories; while the two methods perform equally well for non-urban categories. In this case, the uneven weighting leads to an improvement in performance. As for matching MD LULC to DE LULC, however, the hypothesis cannot stand because the remote sensing method has an overall better performance than the lexical semantic method, and an uneven weighting is no longer needed and will lead to a performance drop.

The causation behind rejecting the hypothesis is quite straightforward. As listed in Appendix I, MD LULC has the fewest categories and the simplest urban classification, which can be handled well by remote sensing. But if a LULC classification system has more detailed urban categories, such as DE LULC and NJ LULC, uneven weighting is more likely to be beneficial.

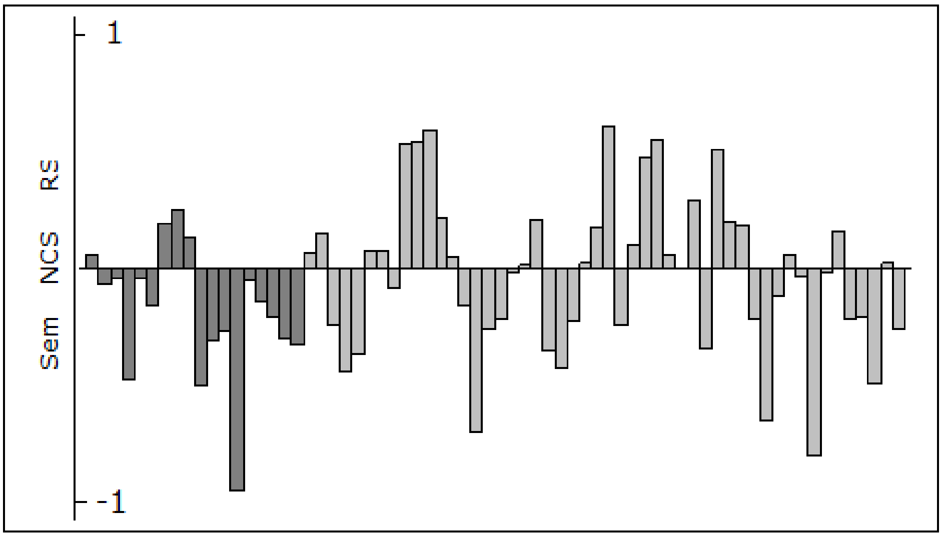


Figure 6.5 (a) Comparing NCS using remote sensing and lexical semantics to match NJ LULC to MD LULC

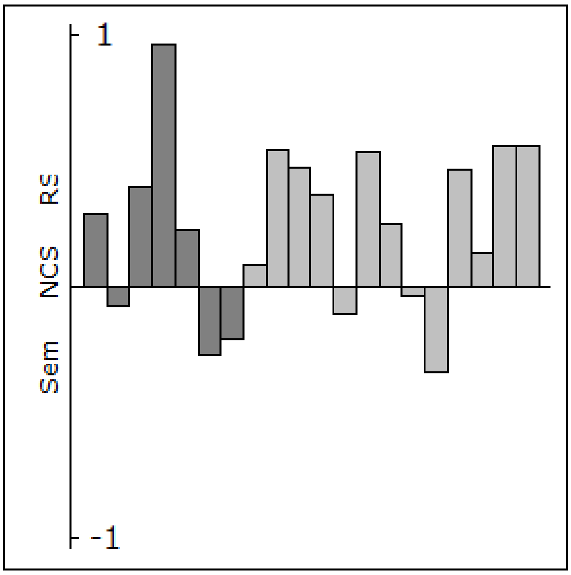


Figure 6.5 (b) Comparing NCS using remote sensing and lexical semantics to match MD LULC to DE LULC

6.3 Balancing Completeness and Accuracy

The methods introduced in this dissertation, either based on lexical semantics or remote sensing, are both completely automated, which means human input or interference are not involved before and during the matching process. But different applications of LULC data, e.g. environmental resource management or urban planning, have different emphases on integrating classification systems. Hence, human expertise, absent in previous stages, will be needed to evaluate integrated classification systems and make adjustment in the matching results to make it suitable to specific needs. To accommodate human adjustments, it is useful to expanding the pool of candidate matches, which should be larger than the matching provided by weighted sum.

As aforementioned, two independent sources can be used as mutual complementation, and this leads to an accommodative scenario of aggregating SIM_{Sem} and SIM_{RS}, in which we pick the higher value from the two to maximize the completeness Then in the matching result, this strategy works as logical disjunction: if either SIM_{Sem} or SIM_{RS} is greater than the threshold, the match will be recognized. Unsurprisingly, the results (Table 6.5) show a high conformance rate (91% - 97%) but a low precision (27% - 35%) (Table 6.6).

Table 6.5 Matching Results Using Logical Disjunction and Conjunction (in bold)

a

DE LULC	**MATCHED MD LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
SinFam	LowRes MedRes	**LowRes MedRes**	Pasture HighRes OpenUrb
MultFam	MedRes HighRes	MedRes HighRes	Comm LowRes
MblHm	HighRes	**HighRes**	LowRes Comm **OpenUrb** Indstrl MedRes
Retail	Comm	**Comm**	HighRes Indstrl
VclAct	Indstrl Comm	Comm Indstrl	Extr
JunkYrd	Indstrl Comm	**Indstrl Comm**	HighRes LowRes **MedRes**
Warehs	Indstrl Comm	Comm **Indstrl**	AgrBldg Inst
OthrCom	Comm	**Comm**	Brush AgrBldg LowRes OrchHrt Indstrl
Indstrl	Indstrl	**Indstrl**	Comm
Utility	OpenUrb Indstrl	**Indstrl**	Pasture Comm Crop
MixUrb	Any urban except OpenUrb	Indstrl **Inst** Comm	MxFrst OpenUrb AgrBldg
OthrUrb	BrGrnd OpenUrb	OpenUrb	Inst Pasture Crop AgrBldg
Inst	Inst	**Inst**	Indstrl Comm HighRes
Recreat	OpenUrb	OpenUrb	OrchHrt GdnCrop Inst **Comm** AgrBldg Indstrl Water
Crop	Crop	**Crop**	FeedOp Pasture
Pasture	Pasture	**Pasture**	Crop AgrBldg Brush
IdleFld	OpenUrb Brush	Brush	LowRes Crop Pasture
OrchHrt	OrchHrt	OrchHrt	LowRes Crop Pasture DeciF
Feedlot	FeedOp	**FeedOp**	Indstrl Extr
Frmstd	AgrBldg	**AgrBldg**	Pasture FeedOp Inst Indstrl
OthrAgr	Crop AgrBldg	Crop	Pasture Indstrl FeedOp Brush
HerbRng	Pasture Brush	Pasture	
ShrbRng	Brush Pasture	**Brush**	LowRes DeciF
MixRng	Brush Pasture	Brush	DeciF LowRes MedRes MxFrst
DecFrst	DeciF	**DeciF**	EvrgrnF MxFrst Brush LowRes
EvrgrnF	EvrgrnF	**EvrgrnF**	MxFrst DeciF
MixFrst	MxFrst	**MxFrst**	Brush **EvrgrnF DeciF** LowRes
ClrCut	BrGrnd Brush	**Brush**	FeedOp Pasture Crop
Watrway	Water	**Water**	
NtrlLk	Water	**Water**	EvrgrnF AgrBldg Wetland
Rsrvr	Water	**Water**	
BayCove	Water	**Water**	

DE LULC	**MATCHED MD LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
Wetland	Wetland	**Wetland**	Brush DeciF MxFrst
Beach	Beach	**Beach**	Water Indstrl
InldSnd	BrGrnd		Crop Indstrl **Beach** Water
Extr	Extr		Beach **Indstrl** Water Comm Crop
Trans	Crop Brush BrGrnd		**FeedOp** GdnCrop OpenUrb Indstrl

b

MD LULC	**MATCHED DE LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
LowRes	SinFam	**SinFam**	JunkYrd MultFam MixRng
MedRes	MultFam SinFam	**SinFam** MultFam	MixRng JunkYrd
HighRes	MultFam MblHm	**MblHm MultFam**	**SinFam**
Comm	Retail OthrCom MixUrb	MixUrb OthrCom **Retail**	JunkYrd
Indstrl	Indstrl JunkYrd Warehs	Warehs JunkYrd **Indstrl**	Retail
Inst	Inst	Inst	Warehs SinFam **MixUrb** OthrUrb
Extr	Extr	Extr	SinFam VclAct Feedlot
OpenUrb	Recreat	Recreat	SinFam MixUrb OthrUrb MblHm
Crop	Crop OthrAgr	**Crop**	Extr TruckCrp SinFam
Pasture	Pasture	**Pasture**	SinFam Crop
OrchHrt	OrchHrt	**OrchHrt**	Crop TruckCrp Trans Pasture OthrCom SinFam
FeedOp	Feedlot	**Feedlot**	Crop Frmstd
AgrBldg	Frmstd	**Frmstd**	Crop OthrCom Feedlot OthrUrb Pasture Warehs
GdnCrop	TruckCrp	TruckCrp	Trans Pasture Crop
DeciF	DecFrst	**DecFrst**	**MixFrst** ShrbRng EvrgrnF MixRng
EvrgrnF	EvrgrnF	**EvrgrnF**	NtrlLk ShrbRng **MixFrst** DecFrst
MxFrst	MixFrst	**MixFrst**	ShrbRng **MixRng** DecFrst EvrgrnF
Brush	ShrbRng MixRng IdleFld ClrCut HerbRng Trans	ClrCut IdleFld **ShrbRng**	OthrAgr EvrgrnF
Water	Watrway Rsrvr NtrlLk BayCove	Watrway **Rsrvr BayCove**	Extr Beach
Wetland	Wetland	Wetland	MixRng Watrway Rsrvr
Beach	Beach	**Beach**	Retail BayCove Extr InldSnd Indstrl
BrGrnd	InldSnd VclAct		Extr Pasture MixRng

c

NJ LULC	MATCHED MD LULC		
	by HUMAN	by ALGORITHM	
		CONFORMING	NON-CONFORMING
HighRes	HighRes	**HighRes**	Indstrl Comm MedRes LowRes
MedRes	MedRes	**MedRes**	Comm LowRes **HighRes**
LowRes	LowRes	**LowRes**	HighRes Brush **MedRes**
RurlRes	LowRes	**LowRes**	**MedRes** Brush HighRes
MixRes	LowRes MedRes HighRes	HighRes	**OpenUrb** Wetland MxFrst
Comm	Comm	**Comm**	OpenUrb HighRes Indstrl
Milit	Inst	Inst	Indstrl HighRes OpenUrb Brush Beach
Indstrl	Indstrl	**Indstrl**	OpenUrb HighRes Comm
Transp			Wetland OpenUrb Indstrl HighRes Comm
Road			Brush Comm HighRes Wetland MedRes
Bridge			**Water** OpenUrb Brush
Airport			Inst HighRes AgrBldg Indstrl
WtlndWa	Brush Wetland	**Wetland** Brush	OpenUrb LowRes DeciF
UpldWaD			Inst LowRes HighRes **OpenUrb** Comm
UpldWa	Brush	Brush	Pasture LowRes OpenUrb
StrmBas			Indstrl Brush HighRes **Comm**
ICCmplx	Indstrl Comm	Comm Indstrl	HighRes AgrBldg
MixUrb	LowRes MedRes HighRes Comm Indstrl Inst	Comm HighRes	MxFrst OpenUrb
OthrUrb	BrGrnd OpenUrb	OpenUrb	HighRes Comm Indstrl
Cemet	OpenUrb	**OpenUrb**	Pasture AgrBldg
WtCemet	OpenUrb Wetland	OpenUrb Wetland	AgrBldg Pasture
Phrg			OpenUrb Wetland HighRes Comm
MngWtld	Pasture Wetland	Wetland Pasture	HighRes OpenUrb Comm
Recreat	OpenUrb	OpenUrb	Inst HighRes Comm Indstrl Brush
Athlet	Inst	**Inst**	Comm Brush **Indstrl** FeedOp
Stadium	Inst	Inst	HighRes Indstrl AgrBldg
MngWtRe	OpenUrb Wetland	Wetland **OpenUrb**	AgrBldg Inst Pasture
CrpPstr	Crop Pasture	Pasture **Crop**	FeedOp GdnCrop
AgriWet	Crop Pasture Wetland	Crop Wetland Pasture	OpenUrb FeedOp **AgrBldg**
FmAgrWt	Wetland Brush	Wetland Brush	DeciF AgrBldg LowRes Pasture

NJ LULC	**MATCHED MD LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
OrchHrt	OrchHrt	OrchHrt	AgrBldg FeedOp Crop Indstrl Pasture
FeedOp	FeedOp	FeedOp	MxFrst Comm Indstrl
OthrAgr	FeedOp AgrBldg GdnCrop	**AgrBldg** FeedOp	Brush Pasture OpenUrb Indstrl
DeciF10	DeciF Brush	DeciF Brush	MedRes EvrgrnF MxFrst LowRes
DeciF50	DeciF	**DeciF**	MxFrst Brush LowRes HighRes EvrgrnF
ConiF10	EvrgrnF Brush	**EvrgrnF** Brush	DeciF MxFrst
ConiF50	EvrgrnF	**EvrgrnF**	HighRes Brush MxFrst DeciF
Plant	OrchHrt EvrgrnF	**EvrgrnF**	DeciF Brush
MxCnF10	MxFrst EvrgrnF Brush	MxFrst **EvrgrnF** Brush	DeciF
MxCon50	MxFrst EvrgrnF	**EvrgrnF** MxFrst	DeciF Brush HighRes
MxDec10	MxFrst DeciF Brush	DeciF Brush MxFrst	**EvrgrnF**
MxDec50	MxFrst DeciF	MxFrst DeciF	EvrgrnF Brush
OldFld	Brush	**Brush**	Comm Pasture
PhrgOld	Brush	**Brush**	EvrgrnF OpenUrb Crop MedRes Wetland
DecBrsh	Brush	**Brush**	MxFrst LowRes **DeciF**
ConBrsh	Brush	**Brush**	EvrgrnF
MxBrush	Brush	**Brush**	DeciF LowRes MxFrst
BrUplnd	BrGrnd		**Brush** MxFrst EvrgrnF OpenUrb
Stream	Water	**Water**	EvrgrnF Wetland **Brush** OpenUrb
NatLake	Water	**Water**	EvrgrnF Wetland
Rsrvr	Water	**Water**	Wetland
TdlRiv	Water Wetland	**Water Wetland**	
TdlBay	Water	**Water**	**Wetland** OpenUrb
Dredge	Water	**Water**	HighRes **Wetland**
Ocean	Water	**Water**	OpenUrb Wetland
SlMrsh	Wetland	**Wetland**	Water Brush LowRes MxFrst
SlMrshV	Wetland	Wetland	Brush EvrgrnF HighRes
FrMrsh	Wetland	**Wetland**	Water
VegDune	Wetland	Wetland	Indstrl OpenUrb Comm EvrgrnF
PhrgCWt	Wetland	**Wetland**	
DecWdWt	Wetland DeciF	Wetland **DeciF**	MxFrst Brush EvrgrnF
ConWdWt	Wetland EvrgrnF	**EvrgrnF** Wetland	Brush DeciF MxFrst
CedarWt	EvrgrnF Wetland	**EvrgrnF** Wetland	Brush
DecBrWt	Brush Wetland	**Wetland Brush**	DeciF MxFrst EvrgrnF
ConBrWt	Brush Wetland	**Wetland Brush**	DeciF **EvrgrnF**
MxBrWtD	DeciF MxFrst	DeciF **Wetland**	EvrgrnF Brush

NJ LULC	**MATCHED MD LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
	Wetland	MxFrst	
MxBrWtC	EvrgrnF MxFrst Wetland	MxFrst **Wetland EvrgrnF**	Brush DeciF
HrbWtNT	Brush Wetland	**Wetland Brush**	MxFrst
PhrgWet	Wetland	**Wetland**	Brush EvrgrnF
MxFrWtD	DeciF MxFrst Wetland	DeciF MxFrst Wetland	**EvrgrnF** Brush
MxFrWtC	EvrgrnF MxFrst Wetland	MxFrst Wetland **EvrgrnF**	Brush DeciF
BrndWet	BrGrnd Wetland	**Wetland**	MxFrst EvrgrnF Brush OpenUrb
Beach	Beach	**Beach**	HighRes Indstrl Water
BrGrnd	BrGrnd	BrGrnd	
Extr	Extr	Extr	Indstrl Beach
AltLnd			OpenUrb Indstrl Comm
DstrbWt	Wetland	**Wetland**	OpenUrb Indstrl Brush
Transi	OpenUrb BrGrnd	OpenUrb	Indstrl Comm
Barren	BrGrnd		Beach Indstrl OpenUrb Brush

d

NJ LULC	**MATCHED DE LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
HighRes	MultFam MblHm	**MultFam** MblHm	Indstrl SinFam
MedRes	SinFam MblHm	**SinFam** MblHm	OthrUrb MultFam MixRng MixUrb
LowRes	SinFam	**SinFam**	OthrUrb MixRng
RurlRes	SinFam	**SinFam**	MixRng MultFam
MixRes	SinFam MultFam MblHm	MblHm	MixUrb Watrway MixRng MixFrst
Comm	Retail VclAct Warehs OthrCom	OthrCom **Retail**	MultFam Frmstd Indstrl
Milit	Inst		**Retail** Rsrvr InldSnd MultFam
Indstrl	Indstrl	**Indstrl**	Warehs Retail
Transp			MixUrb Utility OthrUrb Indstrl MultFam Retail
Road			SinFam OthrCom Indstrl MultFam
Bridge			IdleFld BayCove VclAct
Airport			Frmstd OthrUrb MultFam Warehs Feedlot MixUrb Indstrl
WtlndWa	Wetland	**Wetland**	MixRng NtrlLk InldSnd ShrbRng EvrgrnF
UpldWaD			MixRng Feedlot InldSnd ClrCut SinFam MblHm OthrUrb **MixUrb** Frmstd ShrbRng

NJ LULC	**MATCHED DE LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
UpldWa	ShrbRng MixRng	MixRng ShrbRng	OthrUrb InldSnd ClrCut NtrlLk
StrmBas			OthrCom MultFam
ICCmplx	Indstrl Retail VclAct Warehs OthrCom	OthrCom Indstrl	MultFam
MixUrb	MixUrb	MixUrb	OthrUrb MixFrst MixRng MultFam
OthrUrb	OthrUrb	OthrUrb	Recreat OthrCom MixUrb MultFam
Cemet	OthrUrb	**OthrUrb**	MultFam Frmstd MixUrb **SinFam**
WtCemet	OthrUrb Wetland	OthrUrb Wetland	SinFam OrchHrt Frmstd MultFam
Phrg			Utility MultFam MblHm OthrUrb MixUrb
MngWtld	Pasture HerbRng Wetland	HerbRng Wetland Pasture	Indstrl OthrCom InldSnd
Recreat	Recreat	**Recreat**	VclAct IdleFld
Athlet	Inst		MblHm Indstrl Feedlot **Recreat** OthrCom IdleFld
Stadium	Inst		Frmstd OthrCom MblHm Indstrl InldSnd MultFam
MngWtRe	Wetland Recreat	Wetland Recreat	MixUrb Pasture OthrUrb
CrpPstr	Crop Pasture	Pasture **Crop**	TruckCrp
AgriWet	Crop Pasture Wetland	Crop Wetland Pasture	InldSnd NtrlLk
FmAgrWt	IdleFld Wetland HerbRng ShrbRng MixRng	Wetland IdleFld MixRng ShrbRng	MixUrb OthrUrb
OrchHrt	OrchHrt	OrchHrt	Crop Trans OthrCom
FeedOp	Feedlot	**Feedlot**	**Frmstd** Indstrl
OthrAgr	OrchHrt Feedlot OthrAgr	Feedlot OthrAgr	IdleFld Frmstd
DeciF10	DecFrst ShrbRng MixRng	ShrbRng MixRng DecFrst	MixFrst EvrgrnF Wetland
DeciF50	DecFrst	DecFrst	Wetland EvrgrnF MixFrst MixRng ShrbRng
ConiF10	EvrgrnF ShrbRng MixRng	**EvrgrnF** ShrbRng	Wetland NtrlLk DecFrst MixFrst InldSnd
ConiF50	EvrgrnF	**EvrgrnF**	MixFrst InldSnd DecFrst
Plant	OrchHrt EvrgrnF	EvrgrnF OrchHrt	OthrAgr
MxCnF10	MixFrst EvrgrnF	**EvrgrnF** MixFrst	DecFrst ShrbRng Wetland MixUrb **MixRng**
MxCon50	MixFrst EvrgrnF	**EvrgrnF** MixFrst	DecFrst ShrbRng MixUrb MixRng Wetland
MxDec10	MixFrst DecFrst ShrbRng MixRng	ShrbRng MixFrst DecFrst **MixRng**	Wetland MixUrb EvrgrnF
MxDec50	MixFrst DecFrst	MixFrst DecFrst	EvrgrnF MixUrb ShrbRng Wetland **MixRng**
OldFld	HerbRng MixRng ClrCut	MixRng ClrCut	ShrbRng IdleFld
PhrgOld	HerbRng MixRng	MixRng	**ShrbRng** IdleFld
DecBrsh	ShrbRng MixRng	MixRng **ShrbRng**	NtrlLk InldSnd DecFrst Wetland

NJ LULC	**MATCHED DE LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
ConBrsh	ShrbRng MixRng	**ShrbRng** MixRng	NtrlLk EvrgrnF InldSnd
MxBrush	HerbRng ShrbRng MixRng	**ShrbRng MixRng**	DecFrst MixFrst InldSnd Wetland
BrUplnd	ClrCut Trans		MixFrst **EvrgrnF** Wetland DecFrst
Stream	Watrway	Watrway	EvrgrnF Wetland Rsrvr OthrUrb MixUrb NtrlLk
NatLake	NtrlLk	**NtrlLk**	InldSnd MixRng **Rsrvr**
Rsrvr	Rsrvr	**Rsrvr**	BayCove NtrlLk Watrway OthrAgr
TdlRiv	BayCove Wetland	**BayCove**	InldSnd Rsrvr NtrlLk Beach Watrway
TdlBay	BayCove	**BayCove**	Watrway
Dredge	Rsrvr Watrway	**Watrway**	Wetland MblHm BayCove
Ocean	BayCove	**BayCove**	Beach Indstrl
SlMrsh	Wetland		EvrgrnF VclAct MblHm MixUrb OthrUrb Watrway BayCove
SlMrshV	Wetland		SinFam EvrgrnF MultFam Watrway Crop Extr MixRng
FrMrsh	Wetland	Wetland	Watrway MixRng BayCove Beach
VegDune	Wetland InldSnd	InldSnd	Indstrl Beach MultFam
PhrgCWt	Wetland	Wetland	Watrway EvrgrnF MixRng
DecWdWt	Wetland DecFrst	**Wetland** DecFrst	MixRng EvrgrnF
ConWdWt	Wetland EvrgrnF	EvrgrnF **Wetland**	MixFrst MixUrb MixRng
CedarWt	EvrgrnF Wetland	EvrgrnF Wetland	ShrbRng
DecBrWt	ShrbRng Wetland MixRng	**Wetland ShrbRng** MixRng	EvrgrnF DecFrst
ConBrWt	ShrbRng Wetland MixRng	Wetland ShrbRng	EvrgrnF
MxBrWtD	ShrbRng Wetland DecFrst MixRng	DecFrst **Wetland ShrbRng MixRng**	EvrgrnF MixFrst
MxBrWtC	ShrbRng Wetland EvrgrnF MixRng	ShrbRng **Wetland** EvrgrnF **MixRng**	MixUrb MixFrst
HrbWtNT	Wetland HerbRng	Wetland	ShrbRng NtrlLk EvrgrnF IdleFld MixRng
PhrgWet	HerbRng Wetland	Wetland	Watrway EvrgrnF ShrbRng MixRng
MxFrWtD	Wetland DecFrst MixFrst	DecFrst MixFrst **Wetland**	EvrgrnF MixRng
MxFrWtC	Wetland EvrgrnF MixFrst	**Wetland** EvrgrnF MixFrst	MixRng MixUrb
BrndWet	HerbRng Wetland	**Wetland**	NtrlLk
Beach	Beach	**Beach**	Watrway Indstrl **BayCove** InldSnd
BrGrnd			MblHm JunkYrd Frmstd
Extr	Extr	Extr	Indstrl Feedlot InldSnd Frmstd VclAct
AltLnd			MixUrb OthrUrb Rsrvr VclAct
DstrbWt	Wetland	Wetland	EvrgrnF Trans InldSnd

NJ LULC	**MATCHED DE LULC**		
	by HUMAN	**by ALGORITHM**	
		CONFORMING	**NON-CONFORMING**
Transi	Trans	Trans	OthrCom Feedlot Indstrl
Barren	InldSnd	InldSnd	Extr VclAct Trans

On the contrary, as independent sources can be used to validate each other, in a tight scenario, we pick the lower similarity of the two to maximize the accuracy. In the matching result, this strategy works as logical conjunction: only if both SIM_{Sem} and SIM_{RS} are greater than the threshold, the match will be recognized. This time the results (bold in Table 6.5) show a rather high precision (77% - 80%) with a small sacrifice of the conformance rate. Adopting the F measure defined as the harmonic mean of conformance rate (*C*) and precision (*P*)

$$F = \frac{2PC}{P + C},$$

the logical conjunction strategy beats both lexical semantic and remote sensing methods and all other aggregation schemes (Figure 6.6), because it offers evenly high conformance rate and precision. Having more improved the F measure (0.60-0.75) than all other aggregating strategies, the remote sensing method, and the information retrieval method alone, a logical conjunction (adopting the minimum similarity) is the recommended method to achieve both high completeness and high exactness when matching LULC classifications. Especially, when matching the MD LULC to DE LULC, a logical conjunction has achieved a F measure of 0.75, which is very close to the least agreed human evaluation at 0.77. However, although it is a better solution than previous works, conformance rate at 64% with precision at 78% suggests there are still substantial improvements in our algorithm needed for this fully automated algorithm to be totally reliable in a real world task.

On the other hand, despite algorhtmic improveiments that are progressively harder to achieve as performance gets better, only a small amount of human inputs in the algorithm can provide a more reliable matching. Having a pool of candidates in the accommodative scenario, human experts will be able to refine the matching results according to their requirements. At the same time, they could be much more confident on the matches recognized by both lexical semantics and remote sensing methods.

Table 6.6 Performance in accommodative (Max) and tight (Min) scenarios

	Conformance rate		Precision		F	
	Min	Max	Min	Max	Min	Max
DE2MD	0.68	0.92	0.77	0.288912	0.72	0.42
MD2DE	0.73	0.95	0.78	0.31	0.75	0.47
NJ2MD	0.65	0.97	0.78	0.35	0.71	0.52
NJ2DE	0.49	0.91	0.80	0.33	0.60	0.48

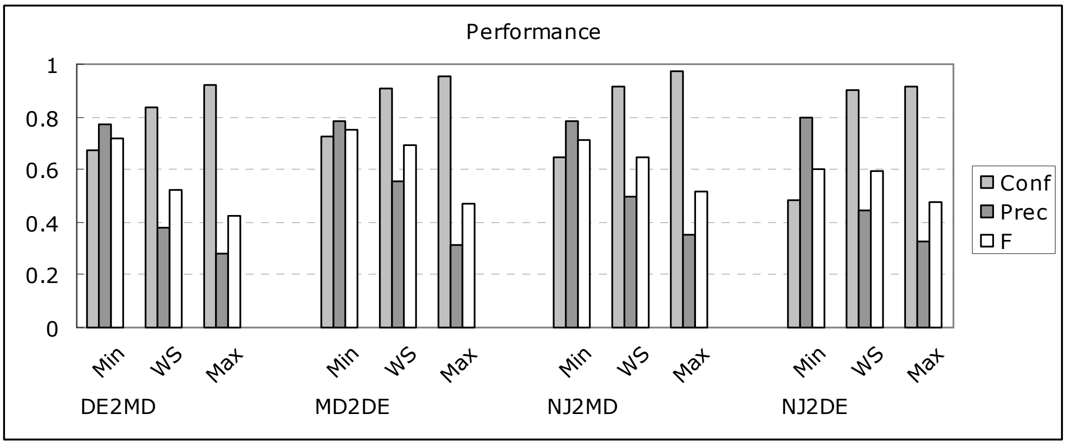


Figure 6.6 Performance of different compound similarities, i.e. logical conjunction (Min), logical disjunction (Max), and weighted sum (WS)

6.4 Awareness of Semantic Heterogeneities

Previously in this chapter, we developed different ways to aggregate SIM_{Sem} and SIM_{RS}, and then compared their resultant matching with human evaluation to decide how well each aggregation method performs. However, as mentioned multiple times in previous chapters, using lexical semantics to match LULC categories, including human evaluation, are potentially vulnerable to semantic heterogeneities. Therefore, in addition to providing a matching result that is more conforming to human evaluation (which we achieved in 6.2 and 6.3), we need to take the other advantage of possessing two independent sources to locate semantic heterogeneities.

There are two main types of semantic heterogeneities, which are naming conflicts and confounding conflicts (Goh 1997). Naming conflicts happen when LULC categories under different labels represent same or similar LULC from observation, while confounding conflicts happen when labels and descriptions of two LULC categories seem to describe same or similar LULC status, but in reality they are different. As discussed in section 2.3, both types have an impact on matching LULC classification systems.

Fortunately, at this point we already have the information needed to discover confounding conflicts and reconcile naming conflicts. In chapter 2, we compared every two LULC categories using lexical semantics, and the similarity reflects how similar two categories are in text (names and descriptions). Then in chapter 4, we compared each two LULC categories using remote sensing, and the similarity reflects how similar two categories are in terms of actual land cover status. By comparing the two results, we could know if LULC categories have names and descriptions consistent to their actual LULC status. Inconsistency usually indicates semantic heterogeneity, and locating inconsistency is important to discover semantic heterogeneities and prevent severe conceptual mistakes in future applications.

The inconsistency between LULC categories' textual definition and actual status will be captured by calculating the difference between SIM_{Sem} and SIM_{RS}. Table 6.7 listed 10 most inconsistent matches (5 positive and 5 negative) for each experiment, in which the difference is calculated as SIM_{RS} minus SIM_{Sem}. A positive value means the two categories in matching in more similar in actual status than in text, which may fulfill a naming conflict, and a negative value means the opposite – a potential confounding conflict, two categories in matching in more similar in text than in actual status. A high value represents greater disparity between actual status and text on that match.

In Table 6.7, we can see many pairs of semantically heterogeneous categories that have been discussed in previous chapters. Given detailed discussions in chapter 2 and 3, we are not surprised to see LULC concepts that are well known to be complicated, such as Wetland and Forest, dominated the list.

Among matches whose SIM_{RS} is higher than SIM_{Sem}, naming conflicts can explain many of those, such as Vehicle Related Activities DE to Industrial MD, Natural Lake DE to Water MD, and Plantation NJ to Evergreen Forest MD. Other matches root in the limitation of remote sensing, where lexical semantic method can make a finer discrimination. For example, Industrial and Commercial Complexes NJ is matched to Multi Family Dwellings DE due to highly similar land cover status from observation, while their land use is actually different. The confusion between forested wetland (e.g. Coniferous Wooded Wetlands and Atlantic White Cedar Wetlands in NJ) and forest (e.g. Evergreen Forest in DE) is included in this type of inconsistency. Surprisingly, remote sensing saw more similarity in Extractive MD to

Extraction DE than the lexical semantic method did. This results from a drawback of using latent semantics rather than "real" semantics. Despite their obvious semantic relation, term extractive and term extraction do not share occurrences in current context and therefore are not considered related in latent semantics.

Table 6.7 Difference between $\mathrm{SIM}_{\mathrm{RS}}$ and $\mathrm{SIM}_{\mathrm{Sem}}$

a

DE LULC	**MD LULC**	**$\mathrm{SIM}_{\mathrm{RS}}$ – $\mathrm{SIM}_{\mathrm{Sem}}$**	**DE LULC**	**MD LULC**	**$\mathrm{SIM}_{\mathrm{RS}}$ – $\mathrm{SIM}_{\mathrm{Sem}}$**
InldSnd	Indstrl	0.600	OrchHrt	OrchHrt	-0.921
VclAct	Indstrl	0.555	OthrAgr	Brush	-0.892
NtrlLk	Water	0.546	MixRng	MxFrst	-0.859
Extr	Indstrl	0.482	Wetland	Wetland	-0.845
OthrAgr	FeedOp	0.469	IdleFld	Brush	-0.749

b

MD LULC	**DE LULC**	**$\mathrm{SIM}_{\mathrm{RS}}$ – $\mathrm{SIM}_{\mathrm{Sem}}$**	**MD LULC**	**DE LULC**	**$\mathrm{SIM}_{\mathrm{RS}}$ – $\mathrm{SIM}_{\mathrm{Sem}}$**
Extr	Extr	0.215	Pasture	Pasture	-0.834
Wetland	MixRng	0.214	Wetland	Wetland	-0.802
OpenUrb	SinFam	0.213	FeedOp	Feedlot	-0.626
Beach	Retail	0.200	Extr	VclAct	-0.609
Beach	BayCove	0.200	OrchHrt	OrchHrt	-0.549

c

NJ LULC	**MD LULC**	**$\mathrm{SIM}_{\mathrm{RS}}$ – $\mathrm{SIM}_{\mathrm{Sem}}$**	**NJ LULC**	**MD LULC**	**$\mathrm{SIM}_{\mathrm{RS}}$ – $\mathrm{SIM}_{\mathrm{Sem}}$**
ICCmplx	HighRes	0.978	MxCnF10	MxFrst	-0.335
ConWdWt	EvrgrnF	0.695	MxDec10	MxFrst	-0.324
Barren	Indstrl	0.683	MxBrWtD	MxFrst	-0.324
Plant	EvrgrnF	0.630	DeciF50	MxFrst	-0.319
Extr	Indstrl	0.610	MxBrWtC	MxFrst	-0.313

d

NJ LULC	DE LULC	$\mathbf{SIM_{RS} - SIM_{Sem}}$	NJ LULC	DE LULC	$\mathbf{SIM_{RS} - SIM_{Sem}}$
Bridge	BayCove	1.000	Barren	VclAct	-0.870
ICCmplx	MultFam	0.998	PhrgWet	Wetland	-0.834
CedarWt	EvrgrnF	0.893	PhrgCWt	Wetland	-0.823
ConWdWt	EvrgrnF	0.783	VegDune	Beach	-0.772
Plant	EvrgrnF	0.759	CedarWt	Wetland	-0.769

As for matches whose SIM_{Sem} is higher than SIM_{RS}, confounding conflict is the main cause, especially for those involving wetland or forest. Confounding conflicts originated from different understanding of same LULC concepts, and can be too subtle to be captured by lexical semantic method. Although sharing same or similar names and descriptions, some categories in different classification systems have very different land cover statuses, e.g. Wetland MD and Wetland DE, which can be observed in remote sensing.

This list, again, reminds us of semantic heterogeneities in LULC classification systems that need to be overcome before LULC data from different sources become interoperable. While comparing the textual definitions of these categories is just telling one side of the story, remote sensing will serve as an important second source to differentiate or reconcile them.

6.5 Conclusions and Recommendations

Textual description and remote sensing are two different angles to understanding a LULC category, and methods using either one of the two data sources were implemented to match LULC classifications. But as we found out in previous chapters, both methods have limitations. Lexical semantic method is vulnerable to semantic heterogeneities, while differentiating some LULC categories is simply beyond the capability of remote sensing.

To overcome the limitations of each method when used alone, we took the advantages of combining the two together. The advantage is two-fold. First, two methods can be mutually complementary: the missing pieces in one can be picked up by the other, and make the matching more complete. Second, two methods enables the mutual validation, which is important to separate information from noise.

Semantic integration aims at determining the relations between concepts, which is based on the measurement of their semantic similarities. The integration of the two methods can be achieved by aggregating the output similarities. We first adopted weighted sum as the aggregating strategy, and by examining the matching result, we find weighted sum gives higher f measure than the better performed lexical semantics or remote sensing method alone (10.0% maximum, 5.3% on average). Benefits of mutual complementation (9.0% maximum, and 3.7% on average increase in conformance rate) and mutual validation (11% maximum, 4.3% on average increase in precision) are both observed in weighted sum's improved matching results, even without any optimization on weights.

After attempts to optimize weighting by exploring methods' input data and output similarities, we realized better weighting should involve domain knowledge. Given remote sensing's limitation on discriminating urban land uses, we proposed an uneven weighting scheme that assigned a higher weight to the similarity of semantics and a lower weight to similarity of remote sensing when matching urban categories, and kept an even weighting (0.5) to both when matching non-urban categories. The performance is slightly improved (1%) under this weighting in majority of cases, although exceptions may happen when a LULC classification has very few and simple urban categories, which can be separated well in remote sensing.

If weighted sum is the straightforward answer to the question what is the integrated matching result, a pool of candidate matches with recommendations is the detailed answer, based on which human adjustments according to specific requirements can be made. Mutual complementation between independent sources leads to an accommodative scenario, in which the higher value from the two similarities is chosen to maximize the completeness, and a highly conforming (94% on average) but less precise (32% on average) matching result is achieved. On the contrary, in a tight scenario lead by mutual validation, the lower similarity is chosen to ensure the accuracy, and the result shows a high precision (78% on average) with only small sacrifice in the conformance rate (63% on average). Either the weighted sum or a tight scenario achieves improved matching results over the Last but not least, we listed categories with largest differences between similarity of semantics and similarity of remote sensing. Many semantic heterogeneous categories discussed in previous chapters are on the list, which once again reminds us of semantic heterogeneities in LULC classification systems that must to be overcome before LULC data from different sources become interoperable.

Chapter 7: Conclusions and Future Directions

7.1 Reason of Integrating LULC Information

In GI Science, one of a most important interoperability issue needs to be resolved for LULC data, LULC is of vital importance because of its environmental impacts on many aspects of the Earth system throughout the globe (Foley et al. 2005),. Accordingly, this research aims to address the interoperability of LULC information derived by different authorities using very different classificatory approaches.

Interoperability is impaired by six levels of heterogeneity (Sheth 1999). Among them, semantic heterogeneity is the main challenge. Semantic heterogeneity originated from the different conceptualization of physical existence. Ontology is the theory of physical existence, and serves as a base to which heterogeneous semantic expressions can return. It makes semantic integration possible.

In this dissertation, we adopted Barry Smith's definition of ontology as the representation of universals (Smith 2004). Inspired by considering LULC categories as universal, we calculated the similarity between categories not only by comparing their semantic expression (e.g. text and semantic taxonomy), but also by comparing their individual parcels, which are directly monitored by modern earth observing technologies, such as remote sensing.

7.2 Summary of Methodology and Findings

LULC data are described by LULC classification systems. The interoperability of LULC data depends on the semantic integration of LULC classification systems. A typical LULC classification system organizes LULC categories in a taxonomy structure, in which each category is defined by a name (label) and often followed by a textual description.

Existing works on semantically integrating LULC classification systems has a major drawback: it is rather difficult to define comparable semantic representations from textual descriptions. To tackle this problem, we borrowed the method of comparing documents in information retrieval, and applied it to the semantic integration of LULC classification system. We tried to compare and match LULC classification systems using lexical information contained in category name and descriptions. Our method is based on bag-of-words model and cosine similarity enhanced by LSA or WordNet, and the results showed large improvement (about 70%) compared to previous feature-based models.

However, this is just solving half of the problem. Lexical semantic methods are not able to solve the semantic heterogeneities happening between different classificatory approaches. In LULC classification systems, confounding conflict happens when labels and descriptions of two LULC categories seem to represent same or similar LULC status, but in reality different. Confounding conflicts are widely observed in complicated land use concepts, such as wetland. Due to variations in vegetation species and coverage, water table height and period, parcels labeled as wetland in different areas can be way different in actual land cover status. Lexical semantic methods, however, are easily disguised by same label and similar concept description to discover this type of semantic heterogeneity. The naming conflict happens when LULC categories under different labels represent same or similar

LULC type from observation. Without confirmation of their actual land cover status, lexical semantic method cannot achieve reliable matching.

To discover confounding conflicts and reconcile naming conflicts, we not only need to know if two LULC categories are seemingly different or similar, but also one step forward, we need to know if they are different or similar in reality, which is why a method of applying remote sensing to the integration of LULC classification systems is important.

Remote sensing is a means of observation on actual LULC status of individual parcels. We calculated parcel level statistics from spectral and textural data, and used these statistics of parcels from different areas in a SVM classifier as training and testing respectively. Then an extensional similarity measurement is adopted to calculate category similarity from parcel level matches. The matching results showed this remote sensing based approach fulfilled its goal – to overcome semantic heterogeneities and achieve more reliable and accurate matching between LULC descriptions in the majority of situations.

The generality of remote-sensing-based integration of LULC classification systems depends on two factors: the availability and applicability of remote sensing data and the comparability of LULC categories. As for data availability, the LULC classifications in comparison must be in areas covered by same/similar type of remotely sensed data. This explains the reason of using Landsat data: it has a global coverage. As for data applicability, remotely sensed data in use must represent a consistent relation between land cover and reflectance values, which means if multiple scenes are involved, the effects of the atmosphere, sensor, and sun on land surface reflectance must be corrected. The experiments suggest our remote sensing method has the geospatial generality in neighboring areas. To deal with these larger areas, the remote-sensing-based method has the potential to serve as a translation between classification systems in distant areas, but a strategy of adaptation is needed to be developed in future works.

The integration method based on remote sensing is limited by the capability of remote sensing, which cause 34%-42% non-conformance (compared to human evaluation) in our matching result. In order to overcome the limitation, we refined the input data of the integration method to only include parcels that are large enough to contain 50 or more Landsat ETM+ pixels. The result shows using large parcels has increased the conformance rate (9.0% maximum, and 3.7% on average) and/or the precision (11% maximum, 4.3% on average) for most LULC categories, except for a few of urban LULC categories that do not have any large parcels. All major forest and wetland categories, which are of special interest in environment analysis and resource management, contain qualified large parcels and were included in the matching process.

Moreover, by using only large parcels, we effectively refined parcel level statistics and made a more reliable data representation of LULC categories, based on which the extraction of underlining semantic relations is achieved.

To overcome the limitations of either lexical semantic or remote sensing based method, we combined the two together by aggregating their output similarities. We first adopted weighted sum as the aggregating strategy, and by examining the matching result, we find weighted sum gives better performance than the better performed lexical semantics or remote sensing method alone (10.0% maximum, 5.3% on average increase in f measure). Benefits of mutual complementation (9.0% maximum, and 3.7% on average increase in conformance rate) and mutual validation (11% maximum, 4.3% on average increase in precision) are both observed in weighted sum's improved matching results. Then in order to accommodate human adjustments according to specific requirements, we introduced an accommodative scenario, in which the higher value from the two similarities is chosen to maximize the completeness, and a tight scenario, in which the lower similarity is chosen to ensure the accuracy. As expected, a highly conforming (94%) but less precise (32%) matching result is achieved in an accommodative scenario, while a high precision (78%) with only small sacrifice of the conformance rate (63%) is achieved in a tight scenario. In our best experiment, the F measure achieved in a tight scenario (0.75) is close to that of the least agreed human evaluation (0.77). This indicates our methodology's effectiveness.

Last but not least, we studied LULC categories with largest differences between lexical semantics and remote sensing, many of which are discussed in previous chapters. This, once again, reminds us of semantic heterogeneities in LULC classification systems that must to be overcome before LULC data from different sources become interoperable and serve as the key to understanding Earth system and global change.

7.3 Future Direction – Potential Global LULC Data Interoperability

Just recently (June 2001), the Global Land Cover Facility (GLCF) at the University of Maryland College Park has launched the first global surface reflectance dataset (GLS SR) based on the Landsat series of satellites (GLCF and GSFC 2011).

Retrieved from remotely sensed images, surface reflectance is corrected for atmospheric effects to approximate the reflectance just above the Earth's surface. Without any artifacts from the atmosphere or illumination and viewing geometry, surface reflectance not only provides an accurate observation on land cover, but also largely increases the consistency between remotely sensed images at different times and locations.

Surface reflectance has been a standard for MODIS. GLS SR is the very first global surface reflectance product based on Landsat data, at a spatial resolution 8 times finer than that of previous MODIS surface reflectance.

GLS SR will provide "the primary input for essentially all higher-level surface geophysical parameters" (GLS SR website: http://www.glcf.umd.edu/data/gls_SR/), and open doors to many new applications, including the semantic integration of LULC information introduced in this dissertation. With GLS SR data's global availability and consistency, our remote-sensing-based method has the potential to serve as a translation between semantic heterogeneous classifications in distant areas, even they are defined in different languages.

However, it is very possible that LULC status, either at a continental scale or for distant areas around the globe, will be too heterogeneous to be integrated directly.

Therefore, in the future we will look for an expansion strategy, such that more comparable classifications from neighboring regions should be first compared and integrated: the integration will then progressively incorporate more comparable classifications and gradually grow to a continental scope.

Appendices

Appendix I

====MARYLAND CATEGORIES====	
Low-density residential(LowRes)	Detached single-family/duplex dwelling units, yards and associated areas. Areas of more than 90 percent single-family/duplex dwelling units, with lot sizes of less than five acres but at least one-half acre (.2 dwelling units/acre to 2 dwelling units/acre).
Medium-density residential(MedRes)	Detached single-family/duplex, attached single-unit row housing, yards, and associated areas. Areas of more than 90 percent single-family/duplex units and attached single-unit row housing, with lot sizes of less than one-half acre but at least one-eighth acre (2 dwelling units/acre to 8 dwelling units/acre).
High-density residential(HighRes)	Attached single-unit row housing, garden apartments, high-rise apartments/condominiums, mobile home and trailer parks. Areas of more than 90 percent high-density residential units, with more than 8 dwelling units per acre.
Commercial(Comm)	Retail and wholesale services. Areas used primarily for the sale of products and services, including associated yards and parking areas.
Industrial(Indstrl)	Manufacturing and industrial parks, including associated warehouses, storage yards, research laboratories, and parking areas.
Institutional(Inst)	Elementary and secondary schools, middle schools, junior and senior high schools, public and private colleges and universities, military installations (built-up areas only, including buildings and storage, training, and similar areas), churches, medical and health facilities, correctional facilities, and government offices and facilities that are clearly separable from the surrounding land cover.
Extractive(Extr)	Surface mining operations, including sand and gravel pits, quarries, coal surface mines, and deep coal mines. Status of activity (active vs. abandoned) is not distinguished.
Open urban land(OpenUrb)	Urban areas whose use does not require structures, or urban areas where non-conforming uses characterized by open land have become isolated. Included are golf courses, parks, recreation areas (except areas associated with schools or other institutions), cemeteries, and entrapped agricultural and undeveloped land within urban areas.
Cropland(Cropland)	Field crops and forage crops.
Pasture((Pasture)	Land used for pasture, both permanent and rotated; grass.
Orchards/vineyards/horticulture (OrchVineHort)	Areas of intensively managed commercial bush and tree crops, including areas used for fruit production, vineyards, sod and seed farms, nurseries, and green houses.
Feeding operations(FeedOp)	Cattle feed lots, holding lots for animals, hog feeding lots, poultry houses.
Agricultural building(AgrBldg)	Agricultural building breeding and training facilities, storage facilities, built-up areas associated with a farmstead, small farm ponds, commercial fishing areas.
Row and garden crops (RowGdnCrop)	Intensively managed truck and vegetable farms and associated areas.
Deciduous forest(DeciFrst)	Forested areas in which the trees characteristically lose their leaves at the end of the growing season. Included are such species as oak, hickory, aspen, sycamore, birch, yellow poplar, elm, maple, and cypress.
Evergreen forest(EvrgrnFrst)	Forested areas in which the trees are characterized by persistent foliage throughout the year. Included are such species as white pine, pond pine, hemlock, southern white cedar, and red pine.

====MARYLAND CATEGORIES====	
Mixed forest(MixFrst)	Forested areas in which neither deciduous nor evergreen species dominate, but in which there is a combination of both types.
Brush(Brush)	Areas which do not produce timber or other wood products but may have cut-over timber stands, abandoned agriculture fields, or pasture. These areas are characterized by vegetation types such as sumac, vines, rose, brambles, and tree seedlings.
Water(Water)	Rivers, waterways, reservoirs, ponds, bays, estuaries, and ocean.
Wetlands(Wetland)	Forested or non-forested wetlands, including tidal flats, tidal and non-tidal marshes, and upland swamps and wet areas.
Bare ground(BareGrnd)	Areas of exposed ground caused naturally, by construction, or by other cultural processes.

====DELAWARE CATEGORIES====
Single Family Dwellings(SinFam)
Multi Family Dwellings(MultFam)
Mobile home Parks/Courts(MblHm)
Retail Sales/Wholesale/Professional Services(Retail)
Vehicle Related Activities(VehicleAct)
Junk/Salvage Yards(JunkYard)
Warehouses and Temporary Storage(Warehouse)
Other Commercial(OtherComm)
Industrial(Indstrl)
Utilities(Utility)
Mixed Urban or Built-up Land(MixUrb)
Other Urban or Built-up Land(OtherUrb)
Institutional/Governmental(Inst)
Recreational(Recreate)
Cropland(Cropland)
Pasture(Pasture)
Idle Fields(IdleFld)
Truck Crops(TruckCrop)
Orchards/Nurseries/Horticulture(OrchNursHorti)
Confined Feeding Operations/Feedlots/Holding(Feedlot)
Farmsteads and Farm Related Buildings(Farmstead)
Other Agriculture(OtherAgr)
Herbaceous Rangeland(HerbRng)
Shrub/Brush Rangeland(ShrubRng)
Mixed Rangeland(MixRng)
Deciduous Forest(DeciFrst)
Evergreen Forest(EvrgrnFrst)
Mixed Forest(MixFrst)
Clear-cut(ClrCut)
Waterways/Streams/Canals(Waterway)
Natural Lakes and Ponds(NtrlLk)
Man-made Reservoirs and Impoundments(Rsrvr)
Bays and Coves(BayCove)
Wetlands(Wetland)
Beaches and River Banks(Beach)
Inland Natural Sandy Areas(InlandSand)

====DELAWARE CATEGORIES====
Extraction(Extr)
Transitional(Trans)

==== New Jersey Categories ====	
1110 HighRes Residential (High Density or Multiple Dwelling)	This category contains either high-density single units or multiple dwelling units on 1/8 to 1/5-acre lots. These areas are found in the densely populated urban zones and generally are characterized by impervious surface coverage of ~65%.
1120 MedRes Residential (Single Unit Medium Density)	This category is comprised of residential urban/suburban neighborhoods greater than 1/8 acre and up to and including 1/2 acre lots. These areas generally contain impervious surface areas of ~30-35%.
1130 LowRes Residential (Single Unit Low Density)	This category contains single unit residential neighborhoods with areas greater than 1/2 acre up to and including 1-acre lots. These areas generally contain impervious surface areas of ~20-25%.
1140 RurlRes Residential (Rural Single Unit)	This category contains single unit residential neighborhoods with areas between 1 acre and up to and including 2-acre lots. These areas generally contain impervious surface areas of ~15-20% or less. This type is found in sparsely populated regions surrounded by or adjacent to forested or agricultural lands. Also included are estates or modern sub-divisions with large lot sizes providing a density less than or equal to 1 dwelling unit per acre. Impervious surface areas in the more rural settings can be as low as 5%.
1150 MixRes Mixed Residential	The mixed residential category is used for an area where various residential uses occur and the individual uses cannot be separated at mapping scale (1 acre). Where more than 1/3 intermixture of other residential use or uses occurs in a specific area, it is classified as mixed residential. Where the inter-mixtures of other residential land use or uses total less than 1/3 of the specified area, the dominant land use category is applied. Impervious surface coverage in these areas can vary significantly.
1200 Comm Commercial and Services	Areas that contain structures predominantly used for the sale of products and services are classified as Commercial and Services. The main building, secondary structures and supporting areas such as parking lots, driveways and landscaped areas are also placed under this category, (unless the landscaped areas are greater than 1 acre in size in which case they are put into a separate category). Sometimes non-commercial uses such as residential or industrial intermix with commercial uses making it difficult to identify the predominant land use. These categories are not separated out; but, if they exceed 1/3 of the total commercial area, the Mixed Urban category (16) is used. Often, specific uses of some commercial and services buildings cannot be easily identified from photography alone. Some supplemental information is required. These areas generally have a high percentage of impervious surface coverage. Any of the specific uses listed below may be included in the 1200 category, with the exception of Military Installations which are delineated separately under the code 1211.
1211 Milit Military Installations	Included in this category are portions of former military installations, that have been de-commissioned and sold. New development of these areas has not yet begun, so particular use can be determined from the photography. Many of the undeveloped portions of these former military sites may remain as preserved undeveloped open space. Developed areas may be re-developed for other uses. However, in all cases, the new intended use is not discernible from the latest available photography, or other ancillary data.
1300 Indstrl Industrial	This category encompasses a great variety of structure types and land uses. Light and heavy industry are comprised of land uses where manufacturing, assembly or processing of products takes place. Power generation is included here because of its similarity to heavy industry. These areas generally have a

==== New Jersey Categories ====	
	high percentage of impervious surface coverage.
1400 Transp Transportation/Communication/Utilities	The transportation, communication, and utilities land uses are often associated with the other Urban or Built-up categories, but are often found in other categories. However, they often do not meet minimum mappable size and are considered an integral part of the land use in which they occur. The presence of major transportation routes, utilities such as sewage treatment plants and power lines, power substations, and communication facilities greatly influence both the present and potential uses of an area. These areas generally have a high percentage of impervious surface coverage.
1410 Road Major Roadway	Major roadways include limited access highways that typically contain at least two lanes in each direction, separated by a concrete barrier or median strip. There are usually no cross streets or traffic lights, and access is limited to ramps. Included in this category are service (rest) areas, right-of-ways, interchanges, maintained hillsides, other service and terminal facilities and portions of local roads. Examples are interstates, U.S. highways and freeways. Limited access highways are characterized by 'diamon' and 'clover-leaf' patterns of ramps, crossroads intersecting via underpasses or overpasses, and the lack of adjacent residential, commercial or industrial development with direct connections to the highway. Limited access highways right-of-way are often bounded by fences or drainage paths.
1419 Bridge Bridge Over Water(WATER)	Bridges over water are characterized by having significance in the delineation of watercourses flowing below. Any bridge or portion of roadway constructed over a mappable open water body has been identified and characterized as water. Although the bridge surface is impervious, the structure does not impact or alter the impervious nature of the water flowing below.
1440 Airport Airport Facilities	Airport facilities are characterized by the presence of long, linear runway surfaces and adjacent areas cleared of vegetation and other obstructions. Typical moderate to large-sized airports contain parallel primary runways, smaller parallel taxi strips, intervening land, aircraft parking aprons, hangars, terminals, service buildings, navigation aids, fuel storage areas, parking lots, and limited buffer zones. This category also includes heliports and land associated with seaplane bases. It does not include other built-up land of small airports.
1461 WtlndWa Wetland Rights-of-Way(WETLANDS)	Included in this category are rights-of-way that exist in former wetland areas, and which still exhibit evidence of soil saturation on the photography. Because of alterations associated with creating the rights-of-way and the periodic clearing, these areas may not support the typical natural wetland vegetation found in adjacent unaltered natural areas. They may, however, support shrubby forms of the surrounding vegetation. They do, however, exist in areas shown on the Natural Resources Conservation Service soil surveys to have hydric soils, and exhibit the darker tonal signatures associated with saturated soils on the photography. Colors of these areas will vary generally from blue-gray to black on winter CIR film and dark gray to black on panchromatic film. Textures will generally be smooth to slightly rough depending on whether the dominant vegetation is low herbaceous species or taller shrubs.
1462 UpldWaD Upland Rights-of-Way Developed	Included in this category are Rights-of-Way in uplands that exist in developed areas. These areas looked well maintained, usually in mowed grass, but are not part of adjacent land use. It should include areas adjacent to urban or agricultural areas but not visibly used in connection with any agricultural or urban land use. They may contain access roads and have a clear separation from surrounding land use. Because of alterations associated with maintaining these portions of the rights-of-way, these areas may not support typical natural vegetation. Textures will generally be smooth due to the lack of low herbaceous species or taller shrubs.

==== New Jersey Categories ====	
1463 UpldWa Upland Rights-of-Way Undeveloped	Included in this category are Rights-of-Way in uplands that usually exist in undeveloped non-urban areas. They typically support shrubby forms of the surrounding vegetation, which may be periodically cut or mowed back. Because of alterations associated with creating the rights-of-way, these areas may support the natural vegetation found in adjacent unaltered natural areas. It should also include areas adjacent to agricultural areas but not visibly used in connection with any agricultural or urban land use. Textures will generally be smooth to slightly rough depending on whether the dominant vegetation is low herbaceous species or taller shrubs.
1499 StrmBas Stormwater Basin	This category consists of stormwater management basins or areas identified as serving the function of a surface water collection site. They are typically associated with new commercial and residential areas. They may contain water and show varying degrees of management or vegetation.
1500 ICCmplx Industrial and Commercial Complexes	The Industrial and Commercial Complexes category includes those industrial and commercial land uses that typically occur together or in close proximity. These areas are commonly referred to as 'Industrial or Commercial Parks.' The major types of business establishments located in these planned industrial and commercial parks are light manufacturing, administration offices, research and development facilities, and computer systems companies. Also found here are facilities for warehousing, wholesaling, retailing and distributing. Industrial and Commercial Complexes are usually located in suburban or rural areas. The key identifying feature is the planned layout of buildings exhibiting the same or very similar construction. Other identifying features include well kept lawns and landscaped areas, ample parking areas and common roadways connecting buildings that also provide access to major highways. The lack of smokestacks, storage tanks, raw materials or finished products, and waste signifies that no heavy industries are present. These areas generally have a high percentage of impervious surface coverage (~85%) and some may be up to 100%.
1600 MixUrb Mixed Urban or Built-up Land	This category includes those urban or built-up areas for which uses cannot be separated into individual categories at the mapping scale employed. Areas are identified under the mixed urban category when more than one-third intermixture of another use or uses is evident. Uses considered in mixed urban include primarily residential, commercial/service, industrial and transportation/communication/utility. Not included in the category are areas considered part of a definable commercial strip as described under 1202. In addition, open land that could be classified for any agricultural use would not be included in the mixed urban category. Level 3 divisions of the Mixed Urban category involve separating the mixed areas based on the predominant use in the intermixture, if one is evident.
1700 OthrUrb Other Urban or Built-up Land	Included are undeveloped, open lands within, adjacent to or associated with urban areas. Some structures may be visible, as in the case of abandoned residential or commercial sites that have not yet been redeveloped. The land cover in these areas may be brush-covered or grassy. Large, managed, maintained lawns common to some residential areas, and those open areas of commercial/service complexes, educational installations, etc., are also included. Undeveloped, but maintained lawns in urban parks are also part of this category, if a specific recreational use is not evident. In addition, areas that have been partially developed or redeveloped but remain unfinished are included. Cemeteries were included in this category in 1986 & 1995, but were separated out for 2002.
1710 Cemet Cemetery	These areas represent large tracts of primarily open land within urban areas. Large cemeteries can be identified by layout of driveways, lots, mausoleums and marking stones. Cemeteries associated with small towns, individual

==== New Jersey Categories ====	
	churches or family estates may not be easily identifiable. Supplemental information is often needed to identify these smaller cemeteries.
1711 WtCemet Cemetery on Wetland(WETLANDS)	These areas represent large tracts of primarily open land within urban or rural areas on land identified as wetland. Large cemeteries can be identified by layout of driveways, lots, mausoleums and marking stones. Cemeteries associated with small towns, individual churches or family estates may not be easily identifiable. Supplemental information is often needed to identify these smaller cemeteries.
1741 Phrg Phragmites Dominate Urban Area	This category contains urban areas where the common reed, Phragmites australis dominates. The photographic signatures for these areas are rough and puffy and range in color from tan to silvery pale white.
1750 MngWtld Managed Wetland in Maintained Lawn Green space(WETLANDS)	Included in this category are former natural wetland areas that now are part of an altered managed landscape, but which still exhibit signs of soil saturation on the imagery. These areas do not support typical wetland vegetation, but are vegetated primarily by grasses and other planted vegetation that may be routinely mowed. Examples of this category would be maintained open lawns and storm water swales in residential, commercial or industrial areas. None of the wetlands included in this category are routinely inundated, although the swales may be on occasion. These altered wetlands exist on areas shown on the US Soil Conservation Service soil surveys to have hydric soils.
1800 Recreat Recreational Land	Under this category are included those areas which have been specifically developed for recreational activities, if these areas are open to the general public. Any facilities that are part of a resort complex and open only to patrons of the hotel or motel are not mapped under category 18, but under Commercial and Services category. Facilities mapped as recreational land may charge user fees to the public, such as public golf courses; or, they may be free to the public, such as ball fields on public school grounds. Level III divisions of this category involve identifying the predominant recreational uses of the areas.
1804 Athlet Athletic Fields (Schools)	Included in this category are a variety of recreational facilities which are not part of established parks, such as baseball fields, tennis courts, basketball courts, and playgrounds. These may be associated with schools. Industrial and commercial firms, or a community housing development.
1810 Stadium Stadium Theaters Cultural Centers and Zoos	Included in this category is any entertainment facility that is developed for public use. Stadiums, outdoor concert halls, racetracks (horse and car), drive-in theaters, amusement parks, and zoos are the primary facilities involved. Such facilities are primarily commercial, although some public recreation areas may be found. Not included are similar facilities on private property, such as horse tracks within private farms, that are open to the public. Parking areas, driveways, and support buildings are mapped in this category.
1850 MngWtRe Managed Wetland in Built-up Maintained Rec Area(WETLANDS)	Included in this category are former natural wetland areas that now are part of an altered managed recreational area, but which still exhibit signs of soil saturation on the imagery. These areas do not support typical wetland vegetation, but are vegetated primarily by grasses and other planted vegetation that may be routinely mowed. Examples of this category would be saturated portions of golf courses, and fields used for baseball and other sports in designated recreation areas. None of the wetlands included in this category are routinely inundated, although portions may be on occasion. These altered wetlands exist on areas shown on the US Soil Conservation Service soil surveys to have hydric soils.
2100 CrpPstr Cropland and Pastureland	This Level II category contains agricultural lands managed for the production of both row and field crops and for the grazing of cattle, sheep and horses. Also included in this category are croplands left fallow or planted with soil improvement grasses and legumes. Cropland and pastureland can easily be distinguished from other land uses with large-scale imagery.

==== New Jersey Categories ====	
2140 AgriWet Agricultural Wetlands (Cranberry Farms & Modified Uplands)(WETLANDS)	Included in this category are lands under cultivation that are modified former wetland areas, and which still exhibit evidence of soil saturation on the photography. These lands will exhibit the textural signature characteristics described for the other agricultural categories, but will have darker color and tonal signatures. Colors will range from blue-gray to black on winter CIR film and dark gray to black on panchromatic film. In addition, these agricultural wetlands also exist in areas shown on soil surveys of the Natural Resources Conservation Service to have hydric soils. In the 2002 update all Cranberry farmland have been combined into this code, regardless of the presence of water.
2150 FmAgrWt Former Agricultural Wetlands (Becoming Shrubby not Built-up)(WETLANDS)	This category was added to identify areas coded as 2140 in the baseline data set, but which do not appear to be under active cultivation in subsequent years. These areas have not undergone any other alterations, such as filling, grading or development, and may again be returned to the 2140 category if the farmland is again place under cultivation. However, these wetlands may continue to develop into a scrub/shrub wetland area if active cultivation is not resumed. As areas in a state of flux, they have been given a separate code.
2200 OrchHrt Orchards Vineyards Nurseries Horticultural Areas Sod Farms	This Level II category contains agricultural areas, which are intensively managed for production of fruits, trees, ornamental plants, and vegetable seedlings. Wholesale greenhouses where plants are grown are also included in this category as are orchards, nurseries, cranberry farms and blueberry farms vineyards, sod and seed farms, and commercial greenhouses. Areas delineated include actively cultivated lands as well as land associated with the operations as, uncultivated lands, dirt roads, dikes, etc.
2300 FeedOp Confined Feeding Operations	This Level II category contains specialized livestock and poultry production enterprises and other specialty farms. These operations have high populations in relatively small areas, resulting in a concentration of waste material. Since this concentrated animal waste is a critical environmental concern, these areas warranted a specific Level II category. Normal structures [barns] associated with a farmstead are not mapped in this category.
2400 OthrAgr Other Agriculture	This category contains other miscellaneous agricultural areas, including experimental fields, horse farms and isolated dikes and access roads.
4110 DeciF10 Deciduous Forest (Low Crown Closure)	This category contains deciduous forest stands that have crown closure greater than 10%, but less than 50%. Crown closure is the percentage of a forest area occupied by the vertical projections of tree crowns. Crown closure percentages provide a reasonable estimate of stand density. An ocular estimate of percent crown closure is made while viewing the area stereoscopically. The ocular judgement is a reliable estimate since the category levels for closure are relatively broad: 10-50% and > 50%. This procedure will also be followed to determine percent crown closure in the other categories.
4120 DeciF50 Deciduous Forest (High Crown Closure)	This category contains deciduous stands with crown closures greater than 50%. The majority of the deciduous forests in New Jersey will be in this category.
4210 ConiF10 Coniferous Forest (Low Crown Closure)	This category contains natural coniferous stands with crown closure> 10%, but less than 50%. Context: This Level II category includes forested lands which contain coniferous tree species. The stand must be 20 feet high and must be stocked by at least 75% conifers to be labeled as a coniferous stand. Coniferous species are those trees commonly known as evergreens. They do not lose their leaves (needless) at the end of the growing season but retain them through the year. Conifers can easily be distinguished from deciduous trees on wintertime color infrared photography because of their high infrared reflectance due to their leaf retention.
4220 ConiF50 Coniferous Forest (High	This category contains natural coniferous stands with crown closure > 50%. Context: This Level II category includes forested lands which contain

==== New Jersey Categories ====	
Crown Closure)	coniferous tree species. The stand must be 20 feet high and must be stocked by at least 75% conifers to be labeled as a coniferous stand. Coniferous species are those trees commonly known as evergreens. They do not lose their leaves (needless) at the end of the growing season but retain them through the year. Conifers can easily be distinguished from deciduous trees on wintertime color infrared photography because of their high infrared reflectance due to their leaf retention.
4230 Plant Plantation	This category contains conifer stands that have been artificially planted. These include stands planted for timber harvesting or aesthetics. Crown closure estimates will not be determined for plantations. Plantations appear as uniform blocks (usually rectangular) of conifers.Other planted stands of conifers, such as Christmas tree farms, will not be included in this category but in the nursery category under Agriculture.
4311 MxCnF10 Mixed Forest (More Coniferous with Low Crown Closure)	This category contains stands of mixed coniferous and deciduous trees with the coniferous species > 50% and with crown closures between 10% and 50%. Context: This category contains stands of mixed coniferous and deciduous trees. The percentage of coniferous trees is higher than the deciduous (>50% of the stand) but the coniferous species do not dominate the stand (<75%).
4312 MxCon50 Mixed Forest (More Coniferous with High Crown Closure)	This category contains stands of mixed coniferous and deciduous trees with the coniferous species > 50% and with crown closures > 50%. Context: This category contains stands of mixed coniferous and deciduous trees. The percentage of coniferous trees is higher than the deciduous (>50% of the stand) but the coniferous species do not dominate the stand (<75%).
4321 MxDec10 Mixed Forest (More Deciduous with Low Crown Closure)	This category contains stands of mixed deciduous and coniferous trees with the deciduous species > 50% and crown closures between 10% and 50%. Context: This category contains stands of mixed deciduous and coniferous trees. The percentage of deciduous trees is higher than the coniferous (> 50%), but the deciduous species do not dominate the stand (< 75%).
4322 MxDec50 Mixed Forest (More Deciduous with High Crown Closure)	This category contains stands of mixed deciduous and coniferous trees with the deciduous species > 50% and crown closures > 50%. Context: This category contains stands of mixed deciduous and coniferous trees. The percentage of deciduous trees is higher than the coniferous (> 50%), but the deciduous species do not dominate the stand (< 75%).
4410 OldFld Old Field (Low Brush Covered)	This category includes open areas that have less than 25% brush cover. The predominant cover types are grasses, herbaceous species, tree seedlings and/or saplings. Old fields are distinguished from inactive farmland (2130) by the amount of brush cover. If a field contains few woody stems (<5%), it should be placed in the inactive farmland category. An area should be placed in the Old Field category if the amount of brush cover requires extensive brush removal before plowing. In some cases, it may not be established that the previous use was agricultural. Context: BRUSH/SHRUBLAND
4411 PhrgOld Phragmites Dominate Old Field	This category contains open fields where the common reed, Phragmites australis dominates. The photographic signatures for these areas are rough and puffy and range in color from tan to silvery pale white. Context: Brush Shrubland
4420 DecBrsh Deciduous Brush/Shrubland	This category contains natural forested areas with deciduous species less than 20 feet in height. An area must have greater than 25% brush cover to be placed in this category. This category also contains inactive agricultural areas that have been grown over with brush. There are photographic signature differences between brushland and the pole or saw-timber stage trees (Categories 4100, 4200, 4300). Besides the obvious height difference visible on stereo viewing, larger trees display much larger crown diameters than brushland areas.
4430 ConBrsh	This category contains natural forested areas with coniferous species less than

==== New Jersey Categories ====	
Coniferous Brush/Shrubland	20 feet high. This category is for natural areas; therefore, Christmas tree farms should be placed in the Nursery category (223).
4440 MxBrush Mixed Deciduous/Coniferous Brush/Shrubland	This category contains natural forested areas less than 20 feet in height with a mixture of coniferous and deciduous trees.
4500 BrUplnd Severe Burned Upland Vegetation	Included in this category are naturally vegetated upland areas which have been altered by intense burning. These burned areas have not re-vegetated sufficiently on the photography, or at the time of any field inspection undertaken to support a mapping effort, to make a determination of the type of vegetation that will re-appear in the burned area. The pre-burn cover type may be any of those listed above in the 4000 series. Where sufficient re-vegetation has occurred to determine a post-burn cover type, the burned area is given the appropriate land cover code. However, where the re-vegetation has been insufficient, the 4500 code has been applied. Note that many different upland forest types may be included in this category.
5100 Stream Streams and Canals	This category includes river, creeks, canals and other linear water bodies that have a minimum width of 80 feet. For watercourses interrupted by control structures, the impoundments are placed in other appropriate water categories (see below), and the impoundment structures are included in the Urban or Built-up category. Remote sensing of these features is not difficult. Colors on infrared photography range from light blue to black, and on the black & white photography the tones range from medium gray to black. The signature can be smooth or rippled depending on the conditions at the time of the photography. The greatest difficulty occurs when overhanging vegetation or shadows obscure the extent of the watercourse.
5200 NatLake Natural Lakes	Water bodies larger than three acres that are non-flowing and naturally enclosed, including regulated natural lakes but excluding reservoirs, are placed in this category. Islands less than three (3) acres are included in the water area. To identify this feature accurately, it is important to remember natural lakes are the results of ground water seepage and surface run-off of precipitation, whereas reservoirs are the result of man-made impoundments and are maintained primarily by linear watercourses. Remote sensing of this feature, once again is simple. The signatures and attendant problems are discussed under category 5100.
5300 Rsrvr Artificial Lakes	Artificial impoundments of water larger than three (3) acres used for irrigation, flood control, municipal water supplies, recreation, landscaping and hydro-electric power or the result of an active extractive operation are included in this category. Dams, bulkheads, spillways and other water control structures should be evident and are critical for accurately identifying these features. Also important to remember is that artificial lakes and reservoirs are charged primarily through linear WCs. Photo identification should key on the non-linear shapes of these features, the water control structures, and the signatures discussed in category 5100. All water reservoirs supporting cranberry operations will be included; however, water within dikes will be included in the agriculture codes for the 2002 update.
5410 TdlRiv Tidal Rivers Inland Bays and other Tidal Waters	Included in this category are the tidal portions of watercourses, enclosed tidal bays, and other tidal water bodies such as tidal pools, ponds and natural lagoons. The tidal watercourses may include everything from smaller entirely tidal features commonly draining tidal marsh systems, to the tidal portions of intermediate and large features such as the Mullica River, the Raritan River, and even the Delaware River. Enclosed tidal bays are those open water tidal features existing commonly behind barrier island systems. These bays generally have a restricted opening to larger tidal features such as Delaware

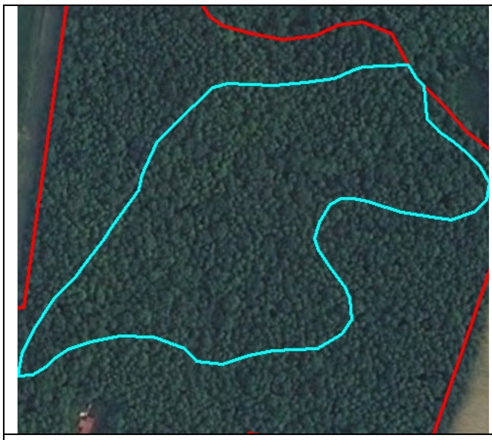
==== New Jersey Categories ====	
	Bay or the Atlantic Ocean. While these features are regularly flushed, portions of these enclosed bays may have complex flushing patterns due to the relatively small outlets. These small bays provide important finfish, shellfish and waterfowl habitat, as well as important recreational potential. Tidal pools and ponds generally will be found in the interior portions of regularly flowed tidal marshes, but these water bodies themselves may not be flooded on every tidal cycle.
5411 TdlBay Open Tidal Bays	Included in this category are large tidal water bodies such as Delaware and Raritan Bays, which have large unrestricted openings directly to the Atlantic Ocean.
5420 Dredge Dredged Lagoon	Artificial dredged lagoons are networks of rectangular dredged areas, containing water, usually associated with residential development or mobile home development. Dredged lagoons are generally in sites of former wetlands and have characteristically bulkheaded shorelines. They usually feed into a central dredged waterway that gives access to open tidal water.
5430 Ocean Atlantic Ocean	This category includes only open water off areas of the Atlantic Ocean. (It was added to identify open ocean offshore waters from those of tidal bays and rivers for water quality analyses).
6111 SlMrsh Saline Marshes (Low marsh vegetation)	This category contains herbaceous vegetation dominated by Spartina alternifloria where the height is <1 foot and is primarily flooded throughout. The photographic signature for these areas range in color from blues to red.
6112 SlMrshV Saline Marshes (High marsh vegetation)	This category contains herbaceous vegetation dominated by Spartina patens where the height is 1 foot to 3 feet. The photographic signature for these areas range in color from red to pink or pale white.
6120 FrMrsh Freshwater Tidal Marshes	These marshes are co-dominated by annual and perennial herbaceous vegetation on substrates associated with tidal waters with salinities less than 1 0/00. Freshwater marsh species are characterized by Nuphar lutea, Peltandra virginica, Pontederia cordata, Zizania aquatica, Polygonum punctatum, Bidens laevis, and Typha latifolia. Marshes exhibiting this cover are found on the tidal Delaware River and tributaries downstream of Trenton to Salem and upstream of the saline marshes on the Atlantic drainage watercourses. Non-tidal marshes are listed under interior wetlands. The photographic signatures for these areas are both smooth-and rough-textured with little elevation. The colors range from dark grey to pink on summer infrared photographs.
6130 VegDune Vegetated Dune Communities	These are areas near the coast that are between saline marsh and open beach. The dominant vegetation can be Ammophila breviligulata, Prunus maritimus, Rhus radicans, Juniperus virginicus, and Acer rubrum. The areas have open to partly closed canopied signatures that are rough in texture and exhibit a red to red brown color on summer infrared photographs.
6141 PhrgCWt Phragmites Dominate Coastal Wetlands	This category contains saline marsh areas where the common reed, Phragmites australis dominates. The photographic signatures for these areas are rough and puffy and range in color from tan to silvery pale white. Freshwater wetlands will have a cowardin code present in the attributes while saline marshes will have no cowardin code.,
6210 DecWdWt Deciduous Wooded Wetlands	These wetlands are closed canopy swamps dominated by deciduous trees normally associated with watercourses, edges of marshes, and isolated wetlands. The important canopy species includes Acer rubrum, Nyssa sylvatica, Fraxinus pennsylvanica, Salix nigra, Quercus bicolor, Q. phellos, Q. falcata, Liquidambar styraciflua, and Platanus occidentalis. These species combine to form a series of mixed hardwood lowland habitats throughout the entire state. These species have photographic signatures that exhibit height, rough texture, and are dark blue-gray to dark gray or black on winter infrared, and gray to dark gray on panchromatic film.
6220 ConWdWt	These wetlands are closed canopy, dominated by coniferous tree species

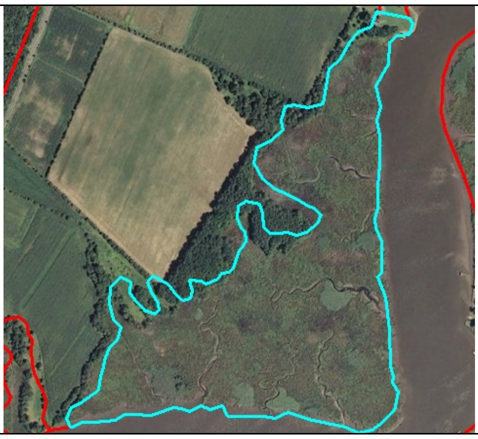
==== New Jersey Categories ====	
Coniferous Wooded Wetlands	associated with watercourses, seeps, and low topographic land. The northern areas will support Tsuga canadensis, Larix laricina, and Picea mariana as monotypic stands or mixed communities. The southern portion of the State has Pinus rigida and P. taeda in monotypic communities or co-dominate with Acer rubrum. Other species such as Nyssa sylvatica and Chamaecyparis thyoides may also be present. These species have photographic signatures that are varied in texture and are red to dark red on winter infrared film and dark gray to black on winter panchromatic film.
6221 CedarWt Atlantic White Cedar Wetlands	These wetlands are predominantly closed canopy, seasonally flooded wetlands of southern New Jersey dominated by Atlantic White-cedar, Chamaecyparis thyoides. Some other trees such as Acer rubrum and Nyssa sylvatica, and shrubs such as Vaccinium corymbosum may also be present. The dense cedar cover, however, generally precludes a heavy herbaceous layer.
6231 DecBrWt Deciduous Scrub/Shrub Wetlands	This brush category will include communities composed primarily of young samplings of deciduous tree species such as Acer rubrum, A. negundo, Liquidamber stryaciflua, Alnus serrulata, Cornus stolonifer, and C. amomum; and woody shrubs such as Vaccinium corymbosum, V. macrocarpon, Spirea alba, Viburnum dentatum, Rosa palustris, Myrica pennsylvania, M. gale, Clethra alnifolia, Cephalanthus occidentalis and Rhododendron viscosum, among others.
6232 ConBrWt Coniferous Scrub/Shrub Wetlands	This brush category will include communities composed primarily of young samplings of coniferous tree species such as Pinus rigida, Larix larcinia, Tusga canadensis, and Picea mariana, and shrubs such as Chamaedaphne calyculata, and Kalmia angustifolia.
6233 MxBrWtD Mixed Scrub/Shrub Wetlands (Deciduous Dom.)	Included in this category are brush and bog wetlands with a mixture of deciduous and coniferous species, with the deciduous species > 50% but < 75%. Species will be similar to those described under 6231 and 6232.
6234 MxBrWtC Mixed Scrub/Shrub Wetlands (Coniferous Dom.)	Included in this category are brush and bog wetlands with a mixture of deciduous and coniferous species, with the coniferous species > 50% but < 75%. Species will be similar to those described under 6231 and 6232.
6240 HrbWtNT Herbaceous Wetlands (Non-Tidal)	These are wetlands dominated by various herbaceous species that are not connected or associated with tidal waters. Lake edges, open flood plains and abandoned wetland agricultural fields are locations for this cover type. Leersia oryzoides, Phalaris arundinacea, Nuphar lutea, Polygonum arifolium, P. sagittatum, Typha latifolia and Phragmites are species that may dominate this cover type. Bog herbaceous vegetation will be covered by this section includes numerous Cyperaceae genera, Juncus sp. and the carnivorous genera of Drosera and Sarracenia. This cover type will have a similar photographic signature as 6120, varied texture, and light blue-gray or tan color on winter infrared and light gray on the panchromatic photograph.
6241 PhrgWet Phragmites Dominate Interior Wetlands	This category contains fresh marsh areas where the common reed, Phragmites australis dominates. The photographic signatures for these areas are rough and puffy and range in color from tan to silvery pale white. Freshwater wetlands will have a cowardin code present in the attributes while saline marshes will have no cowardin code.
6251 MxFrWtD Mixed Forested Wetlands (Deciduous Dom.)	Included in this category are brush and bog wetlands with a mixture of deciduous and coniferous species, with the deciduous species > 50% but < 75%. Species will be similar to those described under 6231 and 6232.
6252 MxFrWtC Mixed Forested Wetlands (Coniferous Dom.)	Included in this category are brush and bog wetlands with a mixture of deciduous and coniferous species, with the coniferous species > 50% but < 75%. Species will be similar to those described under 6231 and 6232.
6500 BrndWet	Included in this category are naturally vegetated wetland areas which have

==== New Jersey Categories ====	
Severe Burned Wetlands	been altered by intense burning at the time of the land cover analysis. These burned areas have not re-vegetated sufficiently on the photography, or at the time of any field inspection undertaken to support a mapping effort, to make a determination of the type of vegetation that will re-appear in the burned area. The pre-burn cover type may be any of those listed above in the 6200 series. Where sufficient re-vegetation has occurred to determine a post-burn cover type, the burned area is given the appropriate land cover code. However, where the re-vegetation has been insufficient, the 6500 code has been applied. Note that many different wetland types may be included in this category.
7100 Beach Beaches	Beaches are predominantly composed of sand and may occur at the land-water interface of oceans, bays and estuaries. Beaches are generally elongated non-vegetated buffering systems subject to the action of waves and tides.
7200 BrGrnd Bare Exposed Rock Rockslides etc.	Areas lacking vegetation and composed of rock or rock faces are included in this category. Exposed rock from highway construction is not included in this category.
7300 Extr Extractive Mining	Extractive operations include a wide variety of mining activities, both surface and subsurface. Included are stone quarries, gravel, sand and clay pits, and limestone quarries to mention a few. Extractive industries are characterized by disturbed ground usually with depth, extractive machinery, buildings and roads for and with heavy equipment. Open mining areas frequently contain water. Extractive mining areas may be large as stone quarries or small as borrow pits.
7400 AltLnd Altered Lands	Altered lands are areas outside of an urban setting that have been changed due to man's activities other than for mining.
7430 DstrbWt Disturbed Wetlands (Modified)	Included in this category are former natural wetlands that have been altered by some form of clearing, leveling, grading, filling and/or excavating, but which still exhibit obvious signs of soil saturation on the imagery. Because of the alterations, these areas do not generally support typical wetland vegetation, and may in fact be unvegetated. They do, however, exist in areas shown on the US Soil Conservation Service soil surveys to have hydric soils, and exhibit the darker tonal signatures associated with saturated soils on the photography. Colors of these areas will vary from gray to blue-gray to black on winter CIR film and gray to black on panchromatic film. These areas may be in transition to a use or associated with a transitional development.
7500 Transi Transitional Areas (sites under construction)	This category encompasses lands on which site preparation for a variety of development types has begun. However, the future land use has not been realized. Included are residential, commercial and industrial areas under construction. Also, areas that are under construction for unknown use and abandoned structures are included. These areas are usually sparsely vegetated.
7600 Barren Undifferentiated Barren Lands	Undifferentiated barren lands encompass cleared lands that have no apparent site preparation or any indication of past activities. Such areas vary in shape and size but generally possess little vegetation, exposing the soil or surface material only. Ancillary information also gives no indication of former uses.

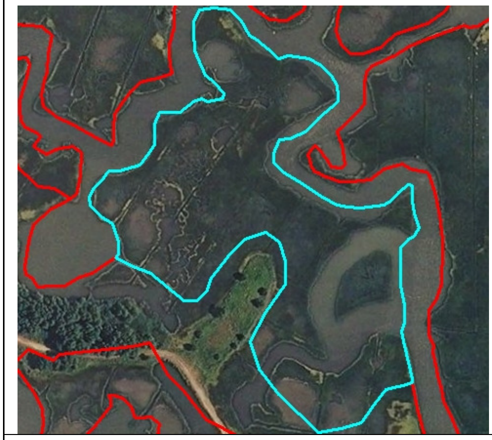
Appendix II

1	2

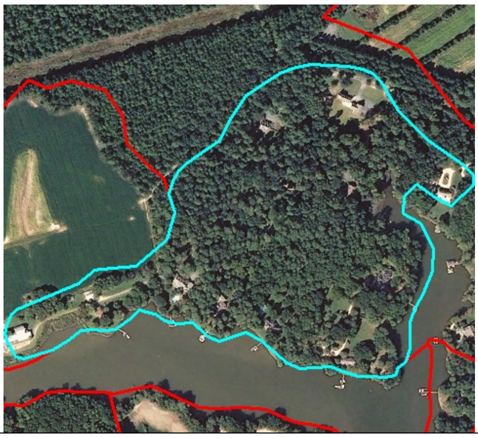




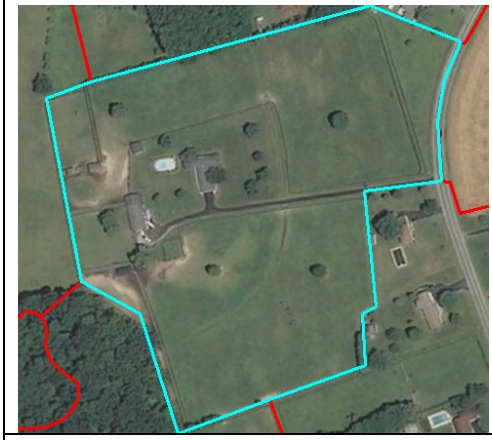
3



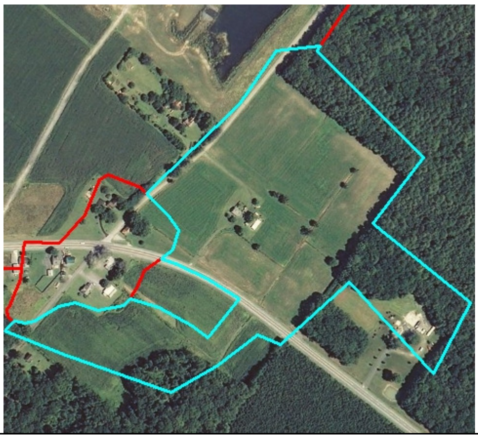
4



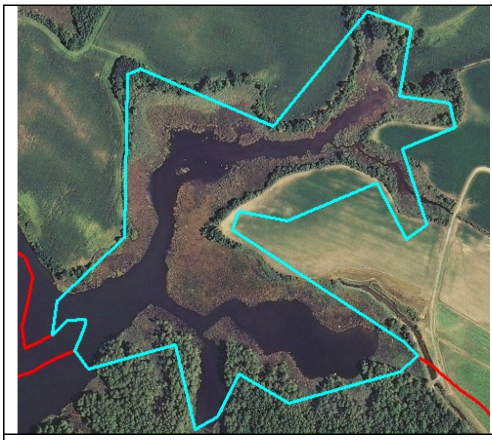
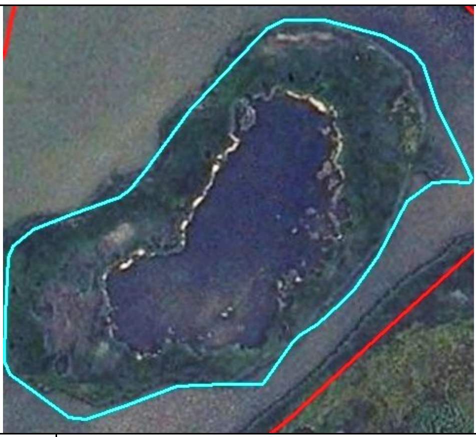
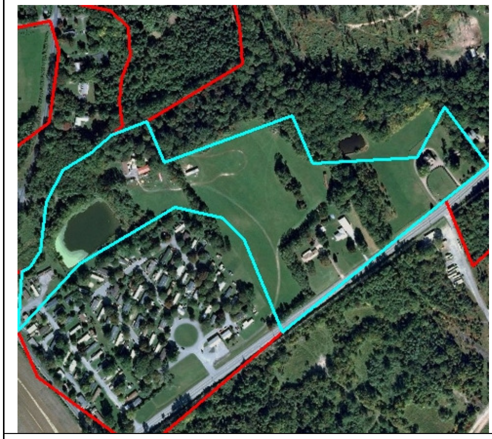
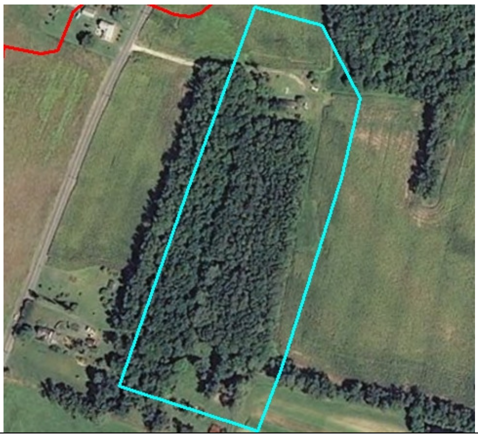
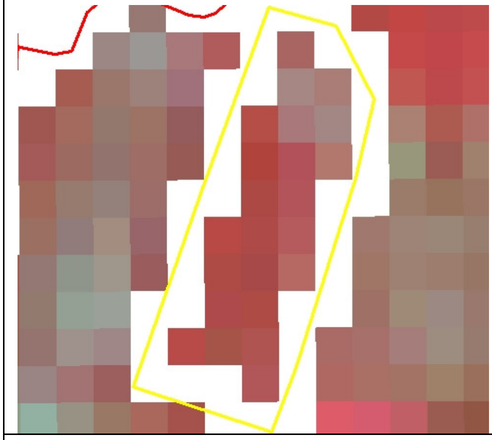
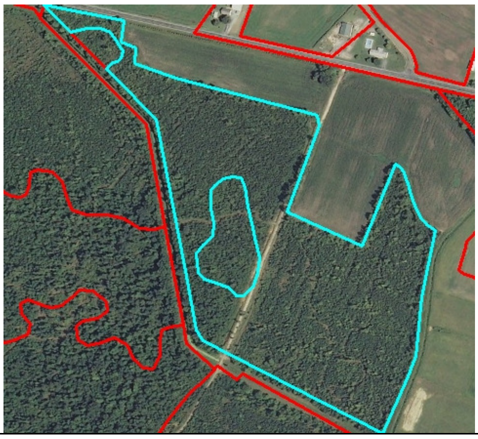
5



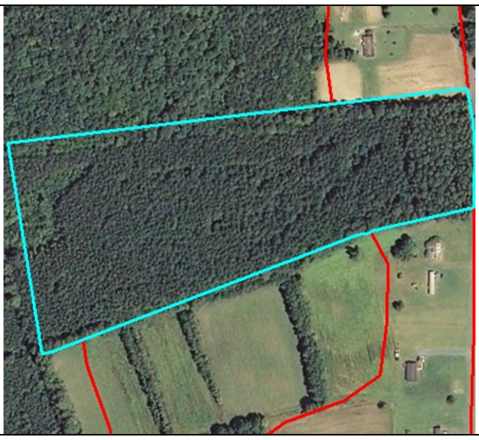
6



7

	
8 	9 
10 	11 
12	13

14



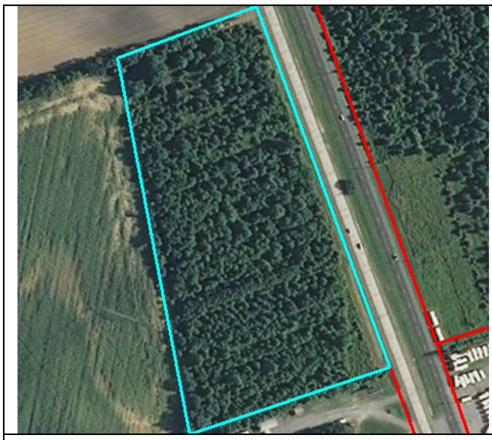
15

16

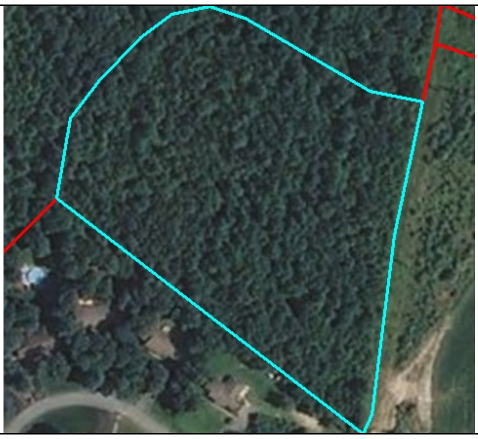
17

18

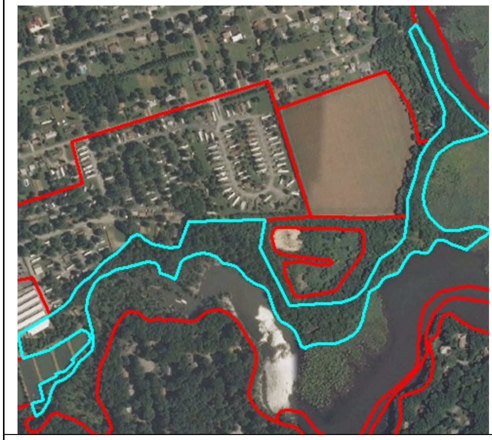
19



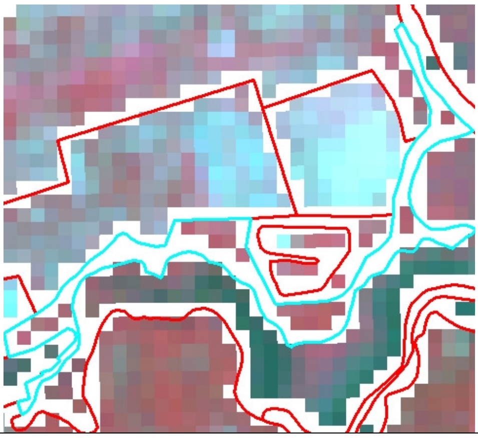
20



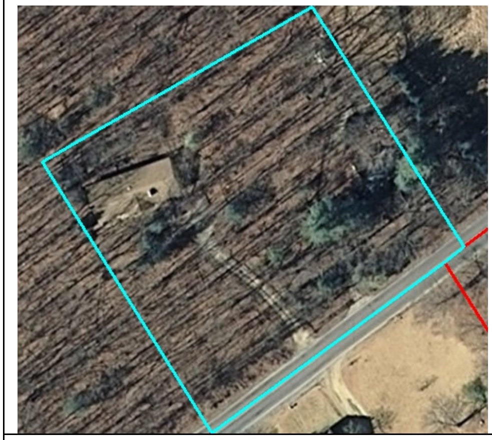
21



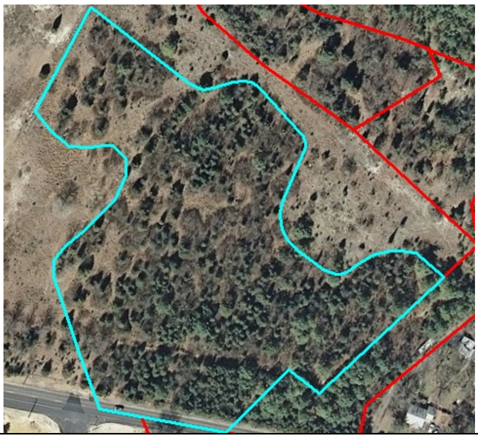
22



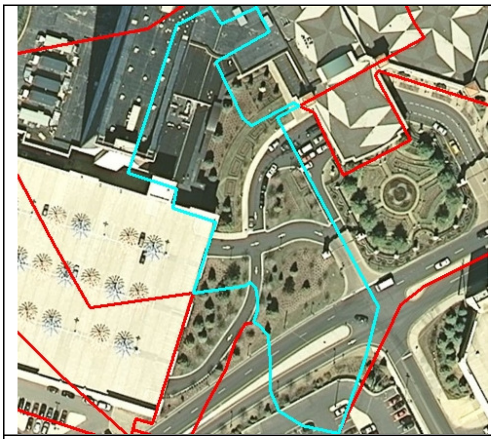
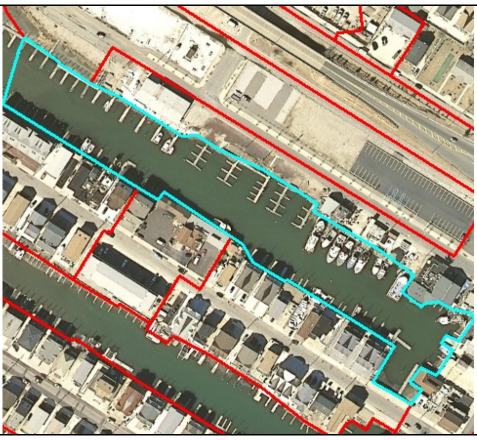
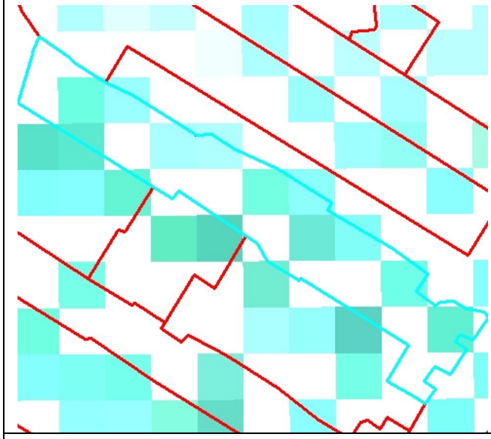
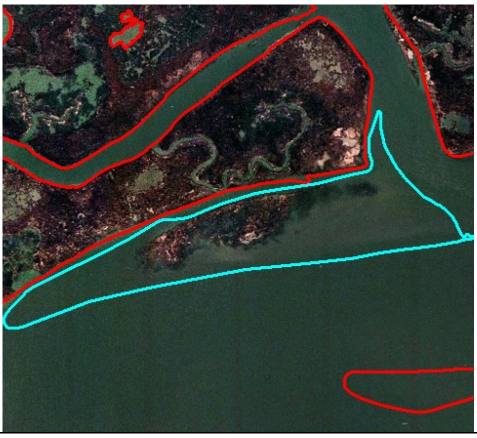
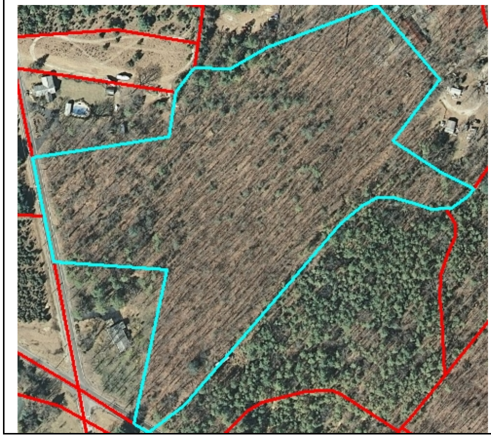
23



24



25

	
26 	27 
28 	

Bibliography

Agarwal, P., 2005, Ontological consideration in GIScience. *International Journal of Geographic Information Science*, 19, pp. 501-536.

Ahlqvist, O., 2008, Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: A study of 1992 and 2001 U.S. National Land Cover Database changes. *Remote sensing of environment*, 112, pp. 1226-1241.

Ahlqvist, O., Keukelaar, J. and Oukbir, K., 2003, Rough and fuzzy geographical data integration. *International Journal of Geographic Information Science*, 17, pp. 223-234.

Ahlqvist, O., Keukelaar, J. and Oukbir, K., 2003, Rough and fuzzy geographical data integration. *International Journal of Geographic Information Science*, 17, pp. 223-234.

Albertoni, R. and Martino, M.D., 2008, Asymmetric and Context-Dependent Semantic Similarity among Ontology Instances In *Journal on Data Semantics X*, pp. 1-30 (Berlin / Heidelberg: Springer).

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. USGS.

Baglioni, M., Masserotti, M.V., Renso, C. and Spinsanti, L., 2007, Building geospatial ontologeis from geospatial databases. In *GeoSpatial Semantics*, pp. 195-209 (Berlin / Heidelberg: Springer).

Bhogal, J., Macfarlane, A. and Smith, P., 2007, A review of ontology based query expansion. *Information Processing and Management*, 43, pp. 866-886.

Bishr, Y., 1998, Overcoming the Semantic and Other Barriers to GIS Interoperability. *International Journal of Geographic Information Science*, 12, pp. 299-314.

Bishr, Y., Pundt, H., Kuhn, W. and Radwan, M., 1999, Probing the concept of information communities - a first step towards a road of semantic interoperability. In *Interoperating Geographic Information Systems*, M.F. Goodchild, M. Egenhofer, R. Fegeas and C.A. Kottman (Eds.), pp. 55-71 (Boston, MA: Kluwer Academic Publishers).

Buscaldi, D., Rosso, P. and Arnal, E.S., 2005, A wordnet-based query expansion method for geospatial information retrieval. In *7th Workshop of the Cross-Language Evaluation Forum* (Alicante, Spain) p. 998.

Butenuth, M., Gösseln, G., Tiedge, M., Heipke, C., Lipeck, U. and Sester, M., 2007, Integration of heterogeneous geospatial data in a federated database. *International Journal of Photogrammetry and Remote Sensing*, 62, pp. 328-346.

Chapman, W., Bridewell, W., Hanbury, P., Cooper, G. and Buchanan, B., 2001, A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34, pp. 301-310.

Chu, W., Liu, Z. and Mao, W., 2002, Textual Document Indexing and Retrieval via Knowledge Sources and Data Mining. In *Communication of the Institute of Information and Computing Machinery (CIICM)* (Taiwan).

Collins, A.M. and Quillian, M.R., 1969, Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, pp. 240-248.

Comber, A., Fisher, P. and Wadsworth, R., 2004, Integrating land-cover data with different ontologies: identifying change from inconsistency. *International Journal of Geographical Information Science*, 18, pp. 691-708.

Comber, A., Fisher, P. and Wadsworth, R., 2005, You know what land cover is but does anyone else?... an investigation into semantic and ontological confusion. *International Journal of Remote Sensing*, 26, pp. 223-228.

Cortes, C. and V., V., 1995, Support-Vector Networks. *Machine Learning*, 20, pp. 273-297.

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V., 2002, GATE: a framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th anniversary meeting of the association for computational linguistics (ACL'02)* (Philadelphia).

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R., 1990, Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, pp. 391-407.

Deng, Y. and Manjunath, B.S., 2001, Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on pattern analysis and machine intelligence*, 23, pp. 800-810.

Doshi, P., Kolli, R. and Thomas, C., 2009, Inexact matching of ontology graphs using expectation-maximization. *Web Semantics: Science, Services, and Agents on the World Wide Web*, 7, pp. 90-106.

Duckham, M. and Worboys, M., 2005, An algebraic approach to automated geospatial information fusion. *International Journal of Geographic Information Science*, 19, pp. 537-557.

Durbha, S.S. and King, R.L., 2005, Semantics-Enabled framework for knowledge discovery from earth obervation data archives. *IEEE transactions on geoscience and remote sensing*, 43.

Durbha, S.S., King, R.L., Shah, V.P. and Younan, N.H., 2009, A framework for semantic reconciliation of disparate earth observation thematic data. *Computers and Geosciences*, 35, pp. 761-773.

Ehrig, M., Haase, P., Hefke, M. and Stojanovic, N., 2005, Similarity for ontologies - a comprehensive framework In *ECIS 2005* (Regensburg, Germany).

Elwood, S., 2008, Grassroots groups as stakeholders in spatial data infrastructures: challenges and opportunities for local data development and sharing. *International Journal of Geographic Information Science*, 22, pp. 71-90.

Euzenat, J. and Shvaiko, P., 2007, *Ontology Matching* (Verlag, Berlin Heidelberg: Springer).

Fan, R., Chen, P. and Lin, C., 2005, Working set selection using the second order information for training SVM. *Journal of Machine Learning Research*, 6, pp. 1889-1918.

FAO, 1995, Planning for Sustainable Use of Land Resources. *FAO Land and Water Bulletin 2. Rome: Food and Agriculture Organization of the United Nations.*

Fellbaum, C.E., 1998, *WordNet - An Electronic Lexical Database* (Cambridge: MIT Press).

Feng, C. and Flewelling, D., 2004, Assessment of semantic similarity between land use/land cover classification system. *Computers, environment and urban systems*, 28, pp. 229-246.

Foley, J., DeFries, R., Asner, G., Barford, C., Bonan, G., Carpenter, S., Chapin, F., Coe, M., Daily, G., Gibbs, H., Helkowski, J., Holloway, T., Howard, E., Kucharik, C., Monfreda, C., Patz, J., Prentice, I., Ramankutty, N. and P., S., 2005, Global consequences of land use. *Science*, 309, pp. 570-574.

Fonseca, F. and Egenhofer, M., 1999, Ontology-driven geographic information system. In *Proceedings of the 7th International Symposium on Advances in Geographic Information Systems* (Kansas City, MO, USA: ACM).

Fonseca, F., Egenhofer, M., Agouris, P. and Câmara, G., 2002, Using Ontologies for Integrated Geographic Information Systems. *Transactions in GIS*, 6, pp. 231-257.

Gardenfors, P., 2000, *Conceptual Spaces - the Geometry of Thought* (Cambridge, MA: The MIT Press).

Global Land Cover Facility (GLCF), Goddard Space Flight Center (GSFC) (2011), Landsat Surface Reflectance , Landsat TM & ETM+, Global Land Cover Facility University of Maryland, College Park.

Goh, C.H., 1997, Representing and reasoning about semantic conflicts in heterogeneous information sources. MIT.

Goldstone, R.L., 1994, Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology*, 20, pp. 3-28.

Goldstone, R.L. and Son, J.Y., 2005, Similarity. In *The Cambridge Handbook of Thinking and Reasoning*, K.J. Holyoak and R.G. Morrison (Eds.) (Cambridge UK: Cambridge University Press).

Goodchild, M.F., 1997, Towards a geography of geogrpahic information in a digital world. *Computers, environment and urban systems*, 21, pp. 377-391.

Gregorio, A.D. and Jansen, L.J.M., 1998, Land Cover Classification System (LCCS): Classification Concepts and User Manual. (Rome: FAO).

Guarino, N., 1995, Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies*, 43, pp. 625-640.

Guarino, N., 1998, formal ontology and information systems. In *FOIS*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I., 2009, The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11.

Haralick, R., Shanmugam, K. and Dinstein, I., 1973, Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3, pp. 610-620.

Harris, Z., 1954, Distributional Structure. *Word*, 10, pp. 146-162.

Huang, C., Davis, L.S. and Townshend, J.R.G., 2002, An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23, pp. 725-749.

Jackendoff, R., 1983, *Semantics and Cognition* (Cambridge, MA: MIT Press).

Kashyap, V., 2001, Design and creation of ontologies for environmental information retrieval. In *Proc. of the 12th Workshop on Knowledge Acquisition, Modeling and Management* (Rome.

Kashyap, V. and Sheth, A., 1998, Semantic heterogeneity in global information sytem: the role of metadata, context, and ontologies. In *Cooperative Information Systems: Current Trends and Directions*, P. Papazoglou and G. Schlageter (Eds.), pp. 139-178 (London: Academic Press).

Kavouras, M., 2005, A unified ontological framework for semantic integration. In *Next Generation Geospatial Information*, P. Agouris and A. Croitoru (Eds.), pp. 147-155 (London: Taylor and Francis Group).

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

Kavouras, M., Kokla, M. and Tomai, E., 2005, Comparing categories among geographic ontologies. *Computers and Geosciences.*

Klein, D. and Manning, C., 2001, Parsing with Treebank Grammars: Empirical Bounds, Theoretical Models, and the Structure of the Penn Treebank. In *39th Annual Meeting on Association for Computational Linguistics (ACL 2001)* (Toulouse, France).

Kokla, M. and Kavouras, M., 2005, Semantic Information in Geo-Ontologies: Extraction, Comparison, and Reconciliation. In *Journal on Data Semantics III*, pp. 125-142 (Berlin / Heidelberg: Springer).

Kruskal, J., 1964, Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis. *Psychometrika*, 29, pp. 1-27.

Kuhn, W., 1995, *Semantics of Geographic Information* (Vienna: Department of Geoinformation Technical University Vienna.

Kuhn, W., 2005, Geospatial Semantics: Why, of What, and How? In *Journal on Data Semantics III*, pp. 1-24.

Lin, D., 1998, An Information-Theoretic Definition of Similarity. In *Proceedings of International Conference on Machine Learning* (Madison, Wisconsin).

Lin, K. and Ludascher, B., 2003, A system for semantic integration of geologic maps via ontology.

Manning, C., Raghavan, P. and Schütze, H., 2008, *Introduction to Information Retrieval* (Cambridge University Press).

Marcus, M., Santorini, B. and Marcinkiewicz, M., 1993, Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, 19, pp. 313-330.

Mather, P., 2004, *Computer Processing of Remotely-Sensed Images: An Introduction*, John Weily and Sons.

Mihalcea, R. and Moldovan, D., 2001, A highly accurate bootstrapping algorithm for word sense disambiguation. *Artificial Intelligence Tools*, 10, pp. 5-21.

Miller, G., 1990, Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography*, 3, pp. 245-264.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K., 1990, Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3, pp. 235-244.

Mücher, C.A., Stomph, T.J. and Fresco, L.O., 1993, Proposal for a Global Land Use Classification. *FAO/ITC/Wageningen Agricultural University, Rome.*

Navigi, R. and Velardi, P., 2003, An analysis of ontology-based query expansion strategies. In *Workshop on adaptive text extraction and mining (ATEM 2003) in 14th European conference on machine learning (ECML 2003)*.

Peng, F. and McCallum, A., 2006, Information extraction from research papers using conditional random fields. *Information Processing & Management*, 42, pp. 963-970.

Potter, M., 1980, An algorithm for suffix stripping. *Program*, 14, pp. 130-137.

Ranganathan, S.R., 1967, *Prolegomena to library classification* (New York: Asia Publishing House).

Resnik, P., 1995, Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448-453.

Resnik, P., 1999, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11, pp. 95-130.

Rips, L.J., Shoben, E.J. and Smith, E.E., 1973, Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*.

Rodriguez, A. and Egenhofer, M., 2004, Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, 18, pp. 229-256.

Rodriguez, A., Egenhofer, M. and Rugg, R., 1999, Assessing semantic similarities among geospatial feature class definitions. In *Interoperating Geographic Information Systems, Second International Conference, Interop '99*, A. Vckovski, K. Brassel and H.-J. Schek (Eds.) (Zurich, Switzerland: Springer-Verlag), pp. 189-202.

Rodriguez, M.A. and Egenhofer, M.J., 2003, Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on knowledge and data engineering*, 15, pp. 442-456.

Rousseeuw, P.J., 1990, Robust Estimation and Identifying Outliers. In *Handbook of Statistical Methods for Engineers and Scientists*, H.M. Wadsworth (Ed.), pp. 16.11-16.24 (New York: McGraw-Hill).

Sboui, T., Bédard, Y., Brodeur, J. and Badard, T., 2007, A Conceptual Framework to Support Semantic Interoperability of Geospatial Datacubes. In *Advances in Conceptual Modeling - Foundation and Applications*, pp. 378-387 (Berlin / Heidelberg: Springer).

Schwering, A., 2008, Approaches to semantic similarity measurement for geo-spatial data: a survey. *Transactions in GIS*, 12, pp. 5-29.

Schwering, A. and Raubal, M., 2005, Spatial Relations for Semantic Similarity Measurement In *Perspectives in Conceptual Modeling*, pp. 259-269 (Berlin / Heidelberg: Springer).

Schwering, A. and Raubal, M., 2005, Measuring Semantic Similarity Between Geospatial Conceptual Regions. In *GeoSpatial Semantics*, pp. 90-106 (Berlin / Heidelberg: Springer).

Sheth, A., 1999, Changing focus on interoperability in information systems: form system, syntax, structure to semantics. In *Interoperating Geographic*

Information Systems, M.F. Goodchild, M.J. Egenhofer, R. Fegeas and C. Cottman (Eds.), pp. 5-29 (Boston: Kluwer Academic Publishers).

Smith, B., 2004, Beyond Concepts: Ontology as Realilty Representation. In *International Conference on Formal Ontology and Information Systems*, A. Varzi and L. Vieu (Eds.) (Turin, Italy).

Smith, B. and Mark, D.M., 2001, Geographical categories: an ontological inverstigation. *International Journal of Geographic Information Science*, 15, pp. 591-612.

Soergel, D., 2005, Thesauri and ontologies in digital libraries. *Digital Libraries, 2005. JCDL apos;05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, 7, pp. 421-421.

Song, K., 2010, Tacking uncertainties and errors in the satellite monitoring of forest cover change. PhD thesis, University of Maryland College Park.

Sowa, J.F., 2000, *Knowledge Representation. Logical, Philosophical, and Computational Foundations* (Pacific Grove, CA: Brooks/Cole).

Spiteri, L., 1998, A simplified model for facet analysis. *Journal of information and library science*, 23, pp. 1-30.

Sunna, W. and F.Cruz, I., 2007, Structure-based methods to enhance geospatial ontology alignment. In *GeoSpatial Semantics*, pp. 82-97.

Torres, M., Quintero, R., Levachkine, S., Moreno, M. and Guzman, G., 2009, Geospatial Information Integration Approach Based on Geographic Context Ontologies. In *Information Fusion and Geographic Information Systems, Lecture Notes in Geoinformation and Cartography*, V.V. Popovich (Ed.), pp. 177-192 (Berlin Heidelberg: Springer-Verlag).

Turner, B., Skole, D., Sanderson, S., Fischer, G., Fresco, L. and Leemans, R., 1995, Land-Use and Land-Cover Change; Science/Research Plan. In *IGBP Report No.35, HDP Report No.7.* (Stockholm and Geneva: IGBP and HDP).

Tversky, A., 1977, Featreus of Similarity. *Psychological Review*, 84, pp. 327-352.

Uitermark, H., Oosterom, P.v., Mars, N.J.I. and Molenaar, M., 1999, Ontology-based geographic data set integration. In *Proceedings of International Workshop on Spatio-Temporal Database Management*, pp. 60-78.

Vink, A.P.A., 1975, *Land Use in Advancing Agriculture* (New York Heidelberg Berlin: Springer-Verlag.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. and Hübner, S., 2001, Ontology-based integration of information - a survey of existing approaches. In *IJCAI-01 Workshop: Ontologies and Information Sharing*, H. Stuckenschmidt (Ed.), pp. 108-117.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. and Hübner, S., 2001, Ontology-based integration of information - a survey of existing approaches. In *International Joint Conferences on Artificial Intelligence, Workshop: Ontologies and Information Sharing*, H. Stuckenschmidt (Ed.) (Seattle, Washington, pp. 108-117.

Walter, V., 2004, Object-based classification of remote sensing data for change detection. *ISPRS Journal of Photogrammetry & Remote Sensing*, 58, pp. 225-238.

Woods, W., 1975, What's in a Link: Foundations for Semantic Networks. In *Representation and Understanding: Studies in Cognitive Science*, D. Bobrow and A. Collins (Eds.) (New York: Academic Press).

Worboys, M. and Duckham, M., 2002, Integrating spatio-thematic information. In *Proceedings of the Second International Conference on Geographic Information Science* M. Egenhofer and D. Mark (Eds.) (Boulder, Colorado: Springer), pp. 346-361.

Wu, S.-S., Qiu, X., Usery, E.L. and Wang, L., 2009, Using geometrical, textual, and contextual information of land parcels for classification of detailed urban land use. *Annals of the Association of American Geographers*, 99, pp. 76-98.

Wynar, B.S. and Taylor, A.G., 1992, *Introduction to Cataloging and Classification 8th ed* (Englewood, Colorado: Libraries Unlimited).

Xu, J. and Croft, W., 1996, Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zurich, Switzerland: ACM Press).

Yarowsky, D., 1995, Unsupervised word sense disambiguation rivaling supervised methods. In *ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (Cambridge, Massachusetts: the Association for Computational Linguistics).

Zhou, N. and Wei, H., 2008, Semantic-based Data Integration and Its Application to Geospatial Portals. In *Fifth International Conference on Geographic Information Scienc* (Park City, Utah: Springer).