MODELLING SEMANTIC UNCERTAINTY OF LAND CLASSIFICATION SYSTEM

XU QIANXIANG

Ph.D

The Hong Kong Polytechnic University

2016

The Hong Kong Polytechnic University

Department of Land Surveying and Geo-Informatics

Modelling Semantic Uncertainty of Land Classification System

XU Qianxiang

A thesis submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy

October 2014

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____________________________(Signed)

______XU QIANXIANG______(Name of student)

ABSTRACT

Land use and land cover information are fundamental for study of earth's ecosystem, global carbon cycling, global climate, atmospheric composition, energy and water balance, biodiversity, ecologically mediated diseases, and other concerns. Several global, regional and national land inventory projects have been carried out to acquire land classification data, e.g., AFRICOVER, CORINE, NLCD, etc. Correspondingly, a vast number of land classification systems have been developed using different philosophical principles to satisfy different purposes. Semantic problems, e.g., semantic overlap, have been identified by few researchers. However, classification systems are usually applied without considering their semantic problems that will cause confusion and could be misleading. Therefore, to address this research gap, this thesis focuses on the semantic uncertainties of classes in classification systems.

First, classes are formalized for quantitative calculation. Based on the characteristics of land class definitions, classes are divided into two types: concept and category. A concept can be rebuilt by applying product operations and union operations, whereas a category can be formalized using an equation set of concepts. A class can always be formalized by applying product operations and union operations.

Second, a reference system is established to uniquely represent all classes. The reference system is set up based on the contrast among classes using a bottom-up method through addition of classes step by step. A reference system is composed of reference concepts, which contain contrast components, not-contrast components, and complement components. All classes can be optimally and economically represented by a combination of reference concepts.

Finally, different models for measuring semantic uncertainties are proposed based on the reference system. These models are divided into three groups: (1) uncertainties between classes, including semantic overlap and semantic similarity; (2) uncertainties between hierarchical levels, including semantic gap and semantic overflow; and (3) semantic interoperability between different classification systems. Characteristics of these models are also analysed in this thesis.

Throughout the thesis, the National Land Cover Database Classification Systems (NLCD CS) of the United States are used for demonstration. The results reveal that the proposed theories and models are feasible and that semantic uncertainties are widespread in the NLCD CSs.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and support of many people. I am deeply grateful to all of these people but can only mention a subset of them here.

First and foremost, I would like to express my sincerest gratitude to my supervisor Dr. Wenzhong Shi, who gave me the chance to study in his research group and motivated me to conduct research on semantic uncertainty. Dr. Shi provided discussions and comments throughout the research. Without his guidance and support, I would not have finished this thesis.

Next, I want to express my thanks to Dr. Mingsheng Liao and Dr. Hui Sheng for encouraging me to study at PolyU. Additionally, I thank Mr. Mingcheng Du for his understanding, trust and support.

Many thanks are given for the financial, academic and technical support from PolyU and its staff. I am thankful to all of the academic and general staff in LSGI for their support and assistance during the period of my Ph.D. study. Special thanks are offered to the research carrel in the library where I wrote this thesis.

I would also like to take this opportunity to thank several colleagues and close friends. To Ling Zhang, Rui Xu, Susu Deng, Mi Jiang, Jian Guo, Jue Wang, Gang Pang, Qi Zhou, Yiliang Wan, Anshu Zhang, Han Xu, Bin Wang, Zelang Miao, Ming Hao, Dr. Haiyong Ding, Dr. Hua Zhang and Dr. Mingyuan Hu, etc. I was so lucky to meet you in Hong Kong. To Chunying Niu, Libo Zhang, Qilong Sun, Xing Liang, Jun Chen, Feng Chen, Rulin Xiao, and Haitao Zhao, etc. in mainland China, thank you for your caring and your support.

Last but not least, I wish to express my deepest thanks to my family: my parents, my parents-in-law, my little sister, and especially my charming wife and my little Mira, for their love, support and encouragement for always.

Qianxiang XU

PolyU, Hong Kong

October, 2015

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

C-concept	Complex-concept
CL	Leaf Concept (Reference Concept)
CS	Classification System
DSRS	Dynamic Semantic Reference System
J-concept	Join-concept
LCCS	Land Cover Classification System
LCS	Land Classification System
NLCD	National Land Cover Database
S-concept	Single-concept
SRS	Semantic Reference System
U-concept	Union-concept

CHAPTER 1 INTRODUCTION

1.1 Motivations

Land use and land cover information are fundamental for study of earth's ecosystem, global carbon cycling, global climate, atmospheric composition, energy and water balance, biodiversity, ecologically mediated diseases, and other concerns (Foley et al., 2005). For example, land use and land cover play a first-order role in affecting climate (Feddema et al., 2005). Since 1850, approximately 35% of anthropogenic CO_2 emissions have been directly generated from land use (Houghton and Hackler, 2001). Changes in land use and land cover are primarily affected by urbanization (Grimm et al., 2008), which leads to an increase in fragmented patches (Cadenasso et al., 2007). It has been estimated that 1/3 to 1/2 of earth's surface has been transformed by human development (NASA news feature, 2005).

The Group on Earth Observation identifies land use and land cover as one of the most important sources of information for all areas of societal benefits in the GEOSS 10-year implementation plan (GEOSS, 2005). Currently, a number of global land use and land cover products are available, e.g., the IGBP DISCover1-km dataset (17-class legend, 1992-1993) (Loveland et al., 2000), the 1-km land cover product of the University of Maryland (UMD) (14-class legend, 1992-1993) (Hansen et al., 2000), the Global Land Cover 2000 (GLC2000) 1-km dataset (23-class legend, 1999-2000) (Bartholome and Belward, 2005), the GeoCover 30m dataset (13-class legend, 1998-2000) (Nelson and Robertson, 2007), and the Global Land Cover (GlobCover) 300m product (22-class legend, 2005-2006) (Bicheron et al., 2008).

Large-scale land inventories have been routinely conducted in certain regions and countries, including the European Union, the United States, China, etc. The CORINE (Coordination of Information on the Environment) Land Cover (CLC) of the European Union consists of the CLC1990, CLC2000, CLC2006 and

CLC2012. The National Land Cover Database (NLCD) products of the United States primarily consist of NLCD1992, NLCD2001, NLCD2006 and NLCD2011. China conducts the National Land Survey (NLS) twice and generates two National Land Survey Dataset (NLSD): NLSD1996 (1984-1996) and NLSD2009 (2007-2009).

To perform these land inventory projects, different land classification systems (LCS) have been established. A 3-level hierarchical classification nomenclature, which contains 44 level-III classes, is established for CORINE (EPA, 2003, 2006). Although the same nomenclature is applied, the interpretation of the classes varies slightly. For example, areas with the presence of small enclaves (<25 ha) of trees and shrub vegetation, swamps, etc. are interpreted as class 243 (land principally occupied by agriculture, with significant areas of natural vegetation) in CLC2000, whereas they were previously interpreted as a homogeneous class (e.g., class 21, arable land) in CLC1990 (Büttner et al., 2004). The NLCD1992 applies a 21-class classification scheme, which represents a merge of the NOAA Coastal Change Analysis Program (C-CAP) classification protocol and the Anderson Land Cover System (Vogelmann et al., 2001). The NLCD2001, NLCD2006 and NLCD2011 apply a 16-class classification scheme by slightly modifying the NLCD 1992 classification scheme. The NLSD1996 adopts a 46-class classification scheme (Lin and Ho, 2003), and the NLSD2009 adopts a 38-class classification scheme (GB/T 21010-2007, 2007).

A classification system is vital and is a prerequisite for a land inventory project. The quality of the classification system significantly affects the accuracy of the classification products. Motivated by facts listed above, this research concentrates on studies of semantic uncertainties in land classification systems.

1.2 Research Gaps and Research Objectives

1.2.1 Practical Problems

It is expected that a LCS is able to comprehensively and exhaustively describe a domain, and every entity in this domain can be denoted by a term. The same terms imply identical meanings and different terms refer to different meanings. However, many problems have been found in the existing classification systems. Selected examples are listed for demonstration.

Semantic overlap and semantic gap exist within a classification system. For example, there are overlapping definitions of crown cover parameters and modifiers in the LCCS (Jansen et al., 2008). A gap exists in the EarthSat GeoCover Global Land Cover legend in that no terms are defined to categorize land with tree height >3 meters and tree cover <35% (Herold and Schmullius, 2004).

Among different classification systems, certain issues of semantic uncertainty are more serious primarily because different land inventory projects are usually designed independently to suit the requirements of different national and international initiatives (Herold et al., 2008). For the purpose of interoperability, matching relationships should be set up between different systems. For example, Fritz and See (2005) established a matching table between the GLC-2000 land cover classes and IGBP land cover classes for comparison of land cover maps. It is common that the same terms from different systems have different meanings. For example, a forest is defined using a tree cover percentage >60% and a height exceeding 2 meters in the IGBP legend, but the GLC2000 defines a forest as a tree cover percentage >15% and a height exceeding 3 meters (Herold and Schmullius, 2004).

1.2.2 Research Gaps

Although certain semantic problems have been noted in selected references, systematic studies on semantic uncertainties are not found in the literature. The LCS is still adopted for classification and without concern as to whether it is sufficient or perfect. Research gaps on semantic uncertainties are described as follows:

(1) Semantic similarity models are primarily developed in disciplines of psychology, linguistics, and information science. Models for spatial information are seldom developed by experts in GIScience.

(2) Semantic overlap and semantic gap have been recognised, but no models have been developed to quantify their magnitudes for a classification system.

(3) Semantic uncertainty models are not yet based on a common foundation.

1.2.3 Research Objectives

To fill the research gaps in studies of semantic uncertainty and resolve practical problems in land classification systems, the ultimate goal of this research is to systematically and quantitatively assess the semantic uncertainties of classes in classification systems based on a common foundation.

To achieve the ultimate goal, the following specific objectives are stated.

OBJECTIVE 1: To analyse the characteristics of land class definitions and to formalize classes for quantitative measurement.

OBJECTIVE 2: To set up a reference system for unique representation of the concerned classes on a common foundation.

OBJECTIVE 3: To quantitatively evaluate the semantic uncertainties of classes based on the reference system.

1.3 Semantic Uncertainties Studied in this Thesis

Semantics relate symbols in a language to their meanings (Gärdenfors, 2000). “Semantic descriptions ought to be an important adjunct, filling out the labels and codings of classes and providing justification for measurements” (Comber et al., 2008b).

Semantic uncertainties in this thesis study the uncertainties of meanings of class names in land classification systems. Semantic uncertainties include semantic overlap, semantic similarity, semantic gap, semantic overflow, and semantic interoperability. Semantic overlap and semantic similarity are used to evaluate uncertainties between any two classes. Semantic gap and semantic overflow are used to evaluate uncertainties between different levels of a classification system. Semantic interoperability is used to evaluate uncertainties between two different classification systems.

Semantic overlap measures whether the meanings of two classes have commonality. Semantic overlap is preferred if some entities actually belong to both classes, such as a class and its subordinate class. Semantic overlap should be avoided between classes at the same level in a classification system because any entity should be assigned to a certain class at a level. Semantic similarity measures how close the meanings of two classes are. The more similar of their meanings, the more close they are.

Semantic gap and semantic overflow measure whether the sums of all class meanings at two levels are identical in a classification system. Semantic gap occurs when an entity can be assigned to a class at a lower level but cannot be assigned to any class at a higher level in a classification system. Semantic overflow occurs when an entity can be assigned to a class at a higher level but cannot be assigned to any class at a lower level in a classification system. Semantic gap and semantic overflow must be problematic and should be tackled.

Semantic interoperability measures whether classes in a classification system can be entirely transformed to classes in another classification system. Semantic interoperability is meaningful and compulsory when communications are carried out between different classification systems for data share and interchange. Theoretically, in a fixed region, land classes in a classification system could be wholly transformed to one or more land classes in another classification system.

This thesis will propose models to measure these five types of semantic uncertainties in a quantitative way which is different from other researches. And more over, all measurements will be calculated based on a common foundation, which is very meaningful that measurements can be compared in a wide scope.

1.4 Research Impacts and Benefits

1.4.1 Impacts on Uncertainty

Sources of uncertainty (see Figure 1.1) in spatial data consist of inherent uncertainties in nature, uncertainties in human cognition, measurement errors, and the propagation of uncertainty in spatial analysis or data processes (Shi, 2008). A suitable idea for resolving inherent uncertainties in nature may not involve changing the entities in nature but instead aiming to understand them deeply. In GIScience, the first step in spatial data capture is geospatial cognition (Shi, 2008). A LCS is a type of recognition of land for classification, and therefore, it is an essential first step in classification.

Currently, research on uncertainty focuses on uncertainties in spatial analysis and data processes. The work in this research primarily focuses on semantic uncertainties, which are uncertainties in human cognition and carries out basic studies of uncertainty to fill in the gaps in this area.

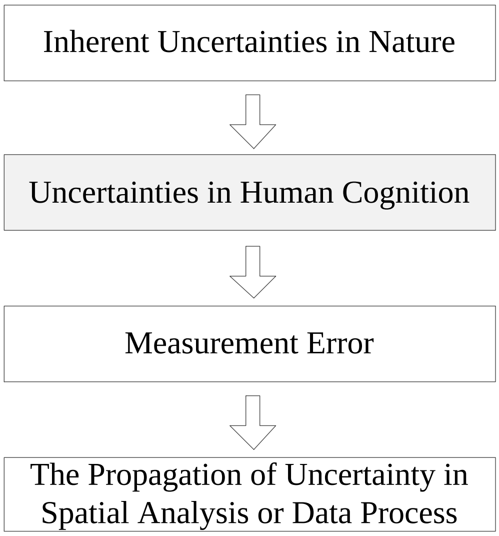


Figure 1.1 Sources of Uncertainty (Shi, 2008) and Research Positions of this Study

1.4.2 Impacts on Classification

The major steps in remote-sensing classification (see Figure 1.2) include selection of remotely sensed data, determination of a suitable classification system, selection of training samples, data preprocessing, feature extraction, selection of suitable classification approaches, post-classification processing and accuracy assessment (Lu and Weng, 2007). This research contributes to the evaluation of the performance of a classification system. A classification system should be informative, exhaustive, and separable (Landgrebe, 2003). However, these requirements are not evaluated in practice, and semantic problems have been noted in operation of classification systems (see Section 1.2.1). The results of this thesis can be used as a tool to quantify the performance of a classification system, and the classification accuracy can be improved by resolving the semantic uncertainties.

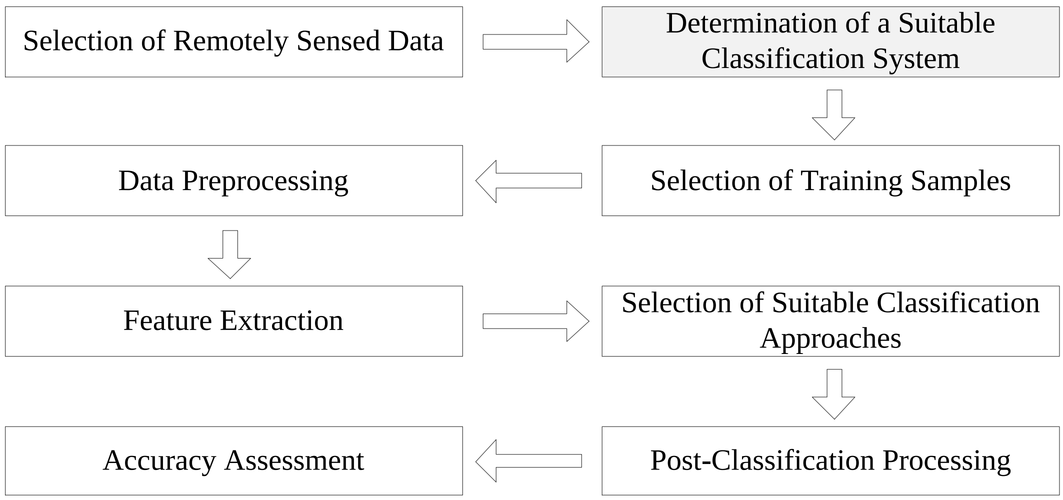


Figure 1.2 Remote-Sensing Classification Process and Research Positions of this Study (highlighted)

1.5 Organization of the Thesis

To achieve the research objectives put forth in Section 1.2.3, I design this work to accomplish the following research contents shown in Figure 1.3.

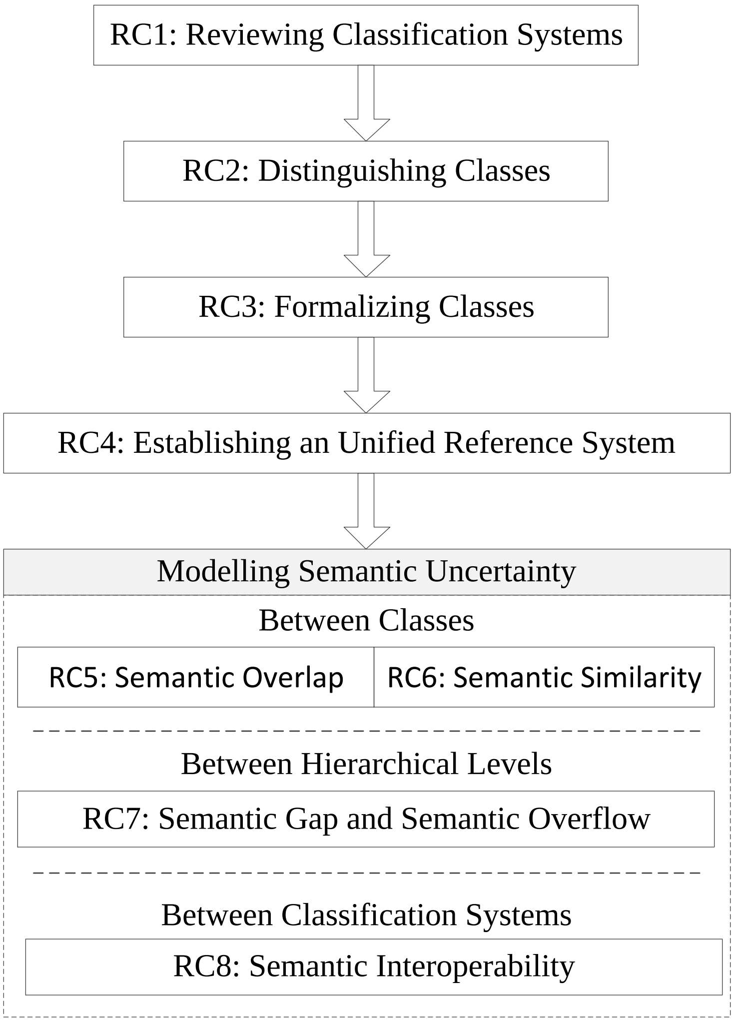


Figure 1.3 Research Contents

The first research content RC1 conducts a literature review of existing prominent classification systems. Next, RC2 through RC4 engage in establishing theoretical foundations for this research, and RC5 through RC8 systematically evaluate the semantic uncertainties between classes, between hierarchical levels, and between different classification systems. In addition to the introduction in Chapter 1 and the conclusion in Chapter 10, these research contents are organized into eight chapters. The logical structure among the chapters, research contents, and research objectives are illustrated in Figure 1.4.

Chapter 2 reviews the existing land classification systems and discusses four systems established by China, the United States, the European Union, and the

United Nations. Semantic problems in the current systems are also identified in this chapter.

Chapter 3 makes a novel distinction of the class definitions in terms of concepts and categories based on complete symbols and incomplete symbols. This new distinction is more suitable for calculation.

Chapter 4 proposes a model for formalization of concepts using product operations and union operations. The characteristics of the model are also analysed in this chapter.

Chapter 5 establishes a unified reference system known as the Dynamic Semantic Reference System (DSRS). Instead of establishing a top-level ontology similar to the Semantic Reference System (SRS) and Land Cover Classification System (LCCS), the DSRS is a dynamic system built by a bottom-up method. All classes of interest can be represented by the DSRS. The NLCD1992 CS and NLCD2001 CS are employed in construction of a DSRS for demonstration.

Chapter 6 proposes a model used to measure semantic overlaps. Semantic overlaps between the NLCD2001 CS classes are systematically evaluated based on the DSRS.

Chapter 7 proposes a model used to measure the semantic gap and semantic overflow that occur between different hierarchical levels. Semantic gaps and semantic overflows in the NLCD2001 CS are evaluated based on the DSRS.

Chapter 8 proposes a model used to measure semantic similarity. Although many models are available for semantic similarity, this model is based on the DSRS. The semantic similarities between NLCD2001 CS classes are evaluated for demonstration.

Chapter 9 employs the semantic overlaps between classes for semantic interoperability between different classification systems. Instead of producing a qualitative matching table, a quantitative transforming table is obtained. The NLCD1992 CS and NLCD2001 CS are used for demonstration.

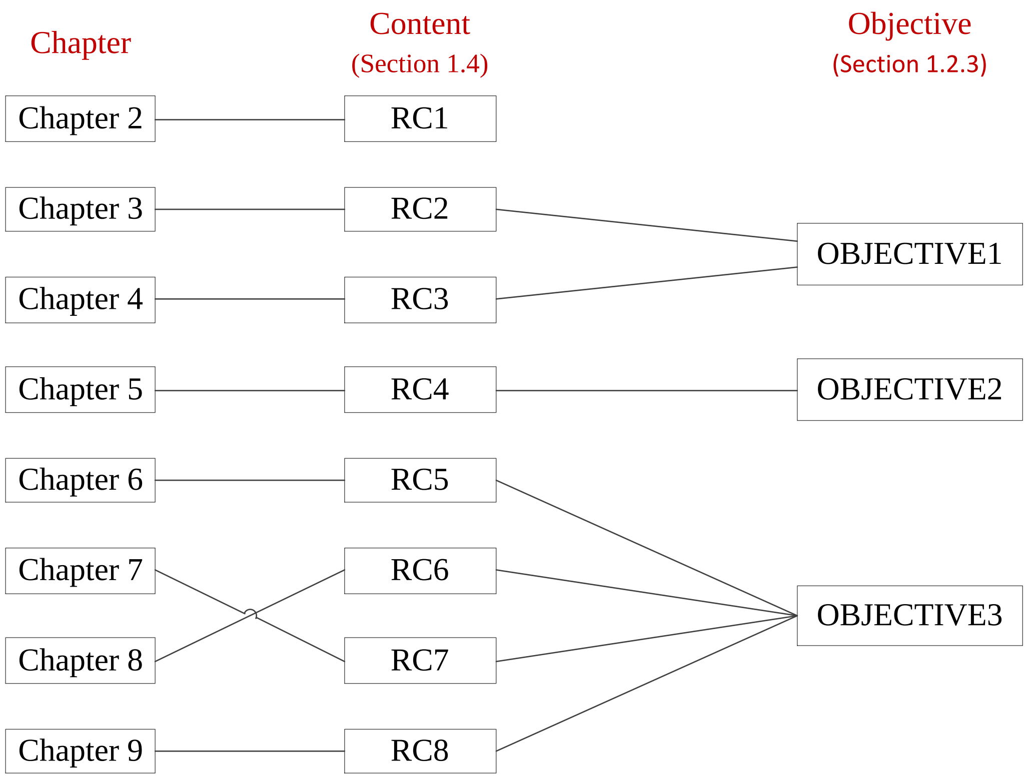


Figure 1.4 Logical Structure of this Thesis

CHAPTER 2 LAND CLASSIFICATION SYSTEMS: A REVIEW

2.1 Introduction

'Science can extend only so far as the power of accurate classification extends' (Ritter, 1916). 'Current mapping techniques of land cover would not be possible today without milestones such as James Anderson's 1976 publication entitled 'A Land Cover Classification System for Use with Remote Sensor Data'' (USGS, http://landcover.usgs.gov/usgslandcover.php, accessed on 27 August 2014).

Classification is a fundamental activity to most sciences (Crovello, 1970; Sokal, 1974). It orders entities into groups on the basis of their features and relationships to achieve economy of memory. A land classification system (LCS) should systematically assign land parcels to land classes. A land parcel is a particular extent of land on the earth (e.g. the area of PolyU located in Hong Kong), and a land class represents an abstraction of real-world land (e.g. a building class). The LCS usually consists of names, codes and definitions of classes in a (nominal) structured framework (e.g. a hierarchical tree). The word 'nominal' is used in this work because most LCSs are not rigorously structured, which will be demonstrated in this thesis.

Land surface is heterogeneous and the LCSs used to acquire, represent, and report land characteristics are as diverse as the land surface itself (Herold et al., 2006b). A vast number of different LCSs have been developed using different philosophical principles (DRDLR, 2013, p.13). The importance and divergence of LCSs are motivations for performing a review of LCSs.

The remainder of this chapter is structured as follows. Section 2.2 introduces the type of roles played by LCSs. Section 2.3 divides LCSs into different types. Four typical land classification projects and their LCSs are introduced in Section 2.4.

Section 2.5 notes potential problems within a LCS and between LCSs. Finally, conclusions are presented in Section 2.6.

2.2 Roles of Classification Systems

Classification systems affect the content, quality and acquisition method of land products. The LCS, which plays a central role in production and application of land use and land cover data, is commonly established and released as a specification or standard at the beginning of a project. The function of the LCS is manifold.

First, a classification system is set up by an organization to meet the legal and regular requirements for retrieving specific information, which is often the motivation to implement a classification project. The LCS is designed to serve a general purpose or a particular purpose. For a general purpose, a LCS should consist of various general classes and neglect the differences within a general class, whereas for a particular purpose, a LCS always emphasizes the existence of a certain type of land and makes fine divisions for this land type. A LCS could be suited to global, continental, national, or local surveys, as determined by whether the classes of the LCS are able to cover all types of land in that area. For example, the CORINE (Coordination of Information on the Environmen) land cover classification system is a continental classification system of Europe designed for a general purpose, and the FRA (Forest Resources Assessment) classification system (FAO, 2001) is a particular system intended for reporting statements and conditions of the world's forest.

Second, a classification system is the foundation used to identify and understand entities. The classification system standardizes producers' activities to generate products of the same type with common characteristics, facilitates the assignment of a large project to multiple manufacturers, and guarantees that final products can be perfectly integrated. Entities with certain commonalities are related and grouped together. Users' and producers' understandings of a common class are matched as predefined in the classification system. Using classification, it is also possible to cut storage and backup costs and speed up entity searches.

Information from individual entities is abstracted into classes, which reduce the information volume for memory and storage.

Third, a classification system describes the entity features and relationships and eases the manipulation required to reveal the laws governing the behaviour of these entities. The classification system acts as a philosophical ground leading to the emergence of the right methodology to fit the principle (Kosolapoff, 1945). Classification simplifies the relationships of the constituent objects, and general statements can be offered relative to a class of entities instead of individual entities. The chaos present in individual entities is ordered by the classes. The primary interesting characteristics are retained via classification, and the other characteristics are ignored.

Fourth, a classification system provides the foundation for information sharing and exchange. What is exchanged is not the data but the meaning within the data, as specified in the class definition. Without commonly accepted definitions of classes, communication cannot be conducted effectively and misinterpretation is prone to occur. For example, a nod of the head indicates ‘yes’ in most countries, but in India, it is assigned the opposite meaning of ‘no’. Land use and land cover information is basic information for various academic and practical applications in different fields. A comprehensive and accurate understanding of the meaning behind the adopted data is the first and foremost precondition leading to a reasonable and scientific solution and conclusion.

Fifth, by adopting a correct classification system, many controversies could be narrowed and even avoided, points under contention could be refined, and scientific work will be facilitated. Data with the same name in different databases might not have the same meaning, and data with different names in different databases might actually mean the same thing (Hunter, 2002). For example, roads can be defined as the physical pavement or the entire road reserve and are represented by line segments, polygons or land parcel boundaries for different application purposes. Differences in a same name create obstacles to direct communication between these departments.

Sixth, classifications are not limited to a reflection of the current situation because classification results can be instrumental in changing the existing circumstances (Jansen and Di Gregorio, 2002). Land cover results are largely employed by environmental researchers to reveal the mechanisms of environmental change. Additionally, these discoveries are used to suggest and guide social-economic activities that will transform the current land cover status.

2.3 Types of Classification Systems

2.3.1 Land Cover or Land Use

The terms of land cover and land use are often used interchangeably, which could lead to confusion and ambiguity. Many definitions of these terms have been put forth by various sources (see Table 2.1). All definitions assume that land cover can be determined directly by observation, whereas determination of land use requires additional information to confirm the purpose for which the land is used, e.g., a statement from the owner. For example, the class ‘forest’ can be defined from a land cover perspective by the vegetation life forms, vegetation height and vegetation density or from a land use perspective by timber production, recreation and conservation of biodiversity. The relationship between land use and land cover is not fixed. Certain forms of land use might take place on only one land cover pattern, e.g., agriculture, but others may take place on more than one land cover, e.g., business or commercial areas, and several land uses might occur on the same piece of land as well, e.g., a forest used for both hunting and timber production.

Land cover is of primary interest to scientific researchers and plays an important role in development of physical and environmental models. Existing applications include climate change, biodiversity, production of statistics for planning and investment, forest and rangeland monitoring, and desertification control (Di Gregorio and Jansen, 1998). At a certain level, land cover acts as a common ground and provides the platform with which to link information from different disciplines (Herold et al., 2006a).

Table 2.1 Definitions of Land Cover and Land Use from Different Sources

Land Cover	Land Use
'Land cover is the observed (bio)physical cover on the earth's surface.' (Di Gregorio, 2005, p.3)	'Land use is characterized by the arrangements, activities and inputs people undertake in a certain land cover type to produce, change or maintain it.' (Di Gregorio, 2005, p.3)
'Land cover is the physical material on the surface of the earth.' (DRDLR, 2013, Definitions)	'Land use means the purpose to which the land is committed.' (DRDLR, 2013, Definitions)
Land cover is 'what can be seen on land when viewed from above' (http://cfpub.epa.gov/roe/indicator.cfm?i=49 accessed on 29 July, 2014)	Land use 'represents the economic and cultural activities that are practiced at a place' (http://cfpub.epa.gov/roe/chapter/land/use.cfm accessed on 29 July, 2014)
Land cover is 'the vegetational and artificial constructions covering the land surface' (Burley, 1961)	Land use is caused by 'man's activities on land which are directly related to the land' (Clawson and Stewart, 1965)
Land cover is 'the composition and characteristics of land surface elements' (Cihlar, 2000)	Land use is caused by 'a series of operations on land, carried out by humans, with the intention to obtain products and/or benefits through using land resources.' (Wyatt et al., 1997, p.10)

Land use is of interest to the planning sector (DRDLR, 2013, p.15). The current land use information is used for management, e.g., rezoning and consent uses, and future land use is concerned with ensuring support for the vision, goals, and objectives of planning. Other applications of land use are focused on evaluating certain aspects of selected projects, e.g., environmental impact and financial value. Fifteen major land use purposes have been listed in the Spatial Planning and Land Use Management Bill (SPLUMB, 2013, p.40).

2.3.2 A Priori or A Posteriori

An a priori classification system is established based on experience and knowledge before data collection occurs, whereas an a posteriori classification system is deduced from characteristics of the collected data (ISO 19144, 2009;

Di Gregorio, 2005, p.5). The main advantage of an a priori system is that the classes can be consistently produced to conform to an existing system. The advantage of an a posteriori system is its flexibility in meeting the study objectives (Wyatt et al., 1997, p.14). The disadvantage of an a priori system is that certain entities in the field may be not easily assigned to a predefined class, whereas the disadvantage of an a posteriori system is limited to a specific area and cannot be used elsewhere (Di Gregorio, 2005, p.6).

Most classification systems are a priori systems that provide standards for data acquisition, sharing, interchange, and communication. An example of an a posteriori classification is the Braun-Blanquet method used in vegetation science (Di Gregorio, 2005, p.5). It is more difficult to compare data classified by an a posteriori system because classifiers may be arbitrarily selected and combined. Although it is impossible to predefine standard classes in an a posteriori system, certain standard rules for establishing classes can be set (ISO 19144, 2009).

2.3.3 General or Specific

From the perspective of contents, classification systems can be general (e.g., for applications of multiple disciplines), or specific (e.g., to meet the particular inquires of a domain). The classes of a general system are the results of a consultancy among disciplines that emphasizes commonalities and discriminates differences. A specific system concentrates only on a certain aspect of land, e.g., forest, soil, or impervious surface. The definitions of classes in the former are more complex than in the latter because sophisticated attributes are employed. Normally only one dominant characteristic is chosen in a specific system, e.g., impervious values in the NLCD2001 impervious surface.

From the perspective of extent, classification systems may be designed for a partial area in a zone or for coverage of the entire zone. Generally, an urban area is treated as a special zone for classification because of the socio-economic importance of this area and its distinct coverage characteristics in contrast with those of a natural area. Land fosters human beings, and human beings exploit and develop land. The same cover materials play different roles in artificial and natural areas. Trees in artificial areas are mostly used for recreation and fruit production but are used most for timber production and hunting in natural areas. Therefore, it is reasonable to separate the classes of artificial areas from natural areas without concerns related to cause semantic gaps.

2.3.4 Name or Classifier

Classes can be defined based on names or classifiers. Most of current classification systems are based on names, which are explained by definitions in natural language. For indexing, processing with computers, and explicitly reflecting hierarchical relationships, codes are normally used in a classification system. One possible reason for the vast number of classification systems based on names is that it is customary to communicate in this style. However, it is highly difficult to manage land data using only class names and their separate definition descriptions using modern GIS techniques (Di Gregorio, 2011). These names act as ‘black boxes’ to anyone outside of the immediate group involved in their preparation. It is difficult to completely and clearly extract the attributes that are applied to characterize these classes. Additionally, interpretations of these definitions may change over time and among different users. These terms also lack systematic and formal principles, which can cause uncertainties. The number of classes of this type in a LCS is normally limited to less than one hundred.

A classifier is a type of attribute used to characterize an aspect of a class, e.g., physiognomic, environmental, or technical. A complete property range is partitioned into discrete attributes in a classification system. Classes are built up using a combination of different classifiers. The advantages of the classifier approach are many. A consistent understanding of classes is easy to acquire because classifiers are usually distinct, clarified, and quantitative. The system can be rendered truly hierarchical and free from internal overlaps. The number of potential classes is exponentially increased with the addition of new classifiers. New automatic classification algorithms may emerge as the characteristics of a class are explicitly reflected by its composite classifiers. The accuracy of the end product can be validated by class or by the classifiers that form the class. Interoperability between systems is directly possible if these terms are defined based on unified classifiers.

2.3.5 Hierarchical or Non-hierarchical

Classification systems may or may not be hierarchically structured (Wyatt et al., 1997, p.9). In a hierarchical classification system, generalized classes are ordered at a lower level, whereas detailed classes are ordered at a higher level. A hierarchical system adapts to record data over a range of scales. The number of classes increases from the lower level to the higher level. A non-hierarchical classification system only contains a single level. Most land classification systems are hierarchically structured (Wyatt et al., 1997, p.14)

Hierarchical systems are a good method for organizing vast amounts of information (Silla and Freitas, 2010) and can be defined by a partially order set $(C, \prec)$, where C is a classification system and $\prec$ represents the relationship. A hierarchical system can be illustrated using a tree or a DAG (Directed Acyclic Graph). Classes structured in a DAG can own more than one parent class, but every class can only have one parent in a tree (except for the root which does not have a parent).

The hierarchical structure can be constructed using a set of local classifiers or global classifiers (Silla and Freitas, 2010). A local classifier acts only on a portion of the hierarchy, whereas a global classifier is used over the entire class hierarchy. There are three types of local classifiers: a local classifier per node, a local classifier per parent node, and a local classifier per level.

2.4 Classification Systems of Large Area Land Inventory Projects

Several global, regional and national land inventory projects have been implemented to acquire data on land and changes of land. The available products include IGBP DISCOVER, the MODIS land cover product, the University of

Maryland (UMD) land cover product, GLC 2000, AFRICOVER, the MODIS continuous fields products, CORINE-land cover, and many national land inventory products. Various classification systems have been established by these projects. Collection of an exhaustive list of these projects and classification systems is impossible, but selected typical projects will be illustrated in this section. The criteria used to select these projects include: (1) periodical projects with historical products and that will be updated regularly, (2) projects covering a large area of land (greater than millions of square kilometres), and (3) projects that represent national, continental, and global levels. According to these criteria, the land inventories implemented in China, the United States, the European Union, and the United Nations are selected and introduced.

2.4.1 NLS (China)

The National Land Survey (NLS) database of China consists of two products completed by the first NLS (the 1984 land survey) and the second NLS (the 2007 land survey). The first NLS began in 1984 and required the intensive work of over 2 million geographical surveyors at a cost of more than RMB 1 billion ($129 million) to gather systematic county-level data on the types, area, location and ownerships of land; this project was completed in 1996 (Lin and Ho, 2003). Aerial photographs, topographical maps and Landsat imageries were collected, printed out, and classified using a two-level classification system, i.e., 8 level-I classes and 46 level-II classes. The validation work for the size, location and changes of land was conducted by field surveys. The final land cover products cover 2,843 counties, 43,000 towns, 740,000 villages, 25,000 farms, and 400,000 administrative units. Although the first NLS products are updated every year, the first NLS products are still out of date for the reason that significant changes in land use and land cover have taken place in the following 11 years.

The second NLS was launched in 2007 to obtain up-to-date, reliable, and accurate land data. Many advanced technologies were applied in the second NLS: (1) high resolution digital aerial photographs and remote sensing imagery, e.g., data acquired by DMC, ADS, IKNOS, QuickBird, RapidEye, WorldView, etc.

are the main sources for computer assisted visual interpretation; (2) advanced survey equipment, e.g., GPS, Total Station, and Electronic Level, that improved the survey and validation efficiency and products accuracy; (3) computer and network technologies employed throughout the project; (4) a new two-level multidiscipline land classification system with 12 level-I classes and 57 level-II classes released in the form of a national standard to guarantee a consistent product; and (5) different quality control software developed and implemented in every product submission process, among the contractor, county government, municipal government, provincial government, and central government. The survey was completed on the reference year of 2009. The product covers 2800 counties in 331 cities of 31 provinces, autonomous regions, and municipalities.

2.4.2 NLCD (United States)

The National Land Cover Database (NLCD) has become a regularly updated product of the United States and is produced every five years. Currently, the NLCD product series has included the NLCD1992 (Vogelmann et al., 2001), NLCD2001 (Homer et al., 2007), NLCD2006 (Fry et al., 2011), and NLCD2011 (http://www.mrlc.gov/nlcd2011.php accessed on 27 August, 2014). The NLCD1992 is the first national land-cover mapping project of the United States on the reference year 1992. NLCD1992 was completed in late 2000. The NLCD series is primarily based on TM and ETM+ data to generate a consistent and seamless 30-m product.

Two different classification systems are used in the NLCD series. The NLCD1992 is based on the NLCD1992 classification system (NLCD1992-CS), and the other three are based on the NLCD2001 classification system (NLCD2001-CS). The NLCD1992-CS is modified from the Anderson Land Cover Classification System (Anderson et al., 1976). The similarities and differences between Anderson and NLCD1992 systems are described in http://landcover.usgs.gov/classes.php (accessed on 27 August, 2014). The NLCD2001-CS, with minor modification from the NLCD1992-CS, is applied to the NLCD2006 and NLCD2011 without any modification. Both of these classification systems are two-level classification systems. There are 21 level-II classes and 9 level-I classes in the NLCD1992-CS and 16 Level-II classes (not including 9 classes in coastal areas and another 4 classes in Alaska only) and 8 Level-I classes in the NLCD2001-CS.

Compared with the NLCD1992-CS definitions, the water, forest, shrub, herbaceous, and wetland classes are nearly identical, and the agriculture, urban, and barren classes are slightly more adjusted in NLCD2001-CS (Homer et al., 2004). Hence, land cover identified with same terms from NLCD2001-CS and NLCD1992-CS might describe different meanings, e.g., 'Deciduous Forest', which has appended additional constraints in the NLCD2001-CS definition of 'generally greater than 5 meters tall, and greater than 20% of total vegetation cover'.

2.4.3 CORINE (European Union)

The CORINE Programme (Coordination of Information on the Environment), which is aimed at gathering information for many different environmental issues, was initiated in 1985 by the European Commission and is maintained and updated by the EEA (the European Environment Agency) based on interpretation of satellite images. The CORINE inventory uses a 3-level hierarchical classification system that includes 5 level-I classes, 15 level-II classes, and 44 level-III classes. The CORINE land cover (CLC) inventory has been implemented in most of the European countries, and until 2014, four CORINE databases have been produced, i.e., CLC1990, CLC2000, CLC2006, and CLC2012 on the reference years 1990, 2000, 2006, and 2012, respectively.

The CLC1990 is the first Europe-wide land cover inventory (Kleeschulte and Büttner, 2008). The CLC1990 inventory is implemented using the photointerpretation method on hardcopies which is discarded in subsequent projects for development of techniques and reduction of potential errors and costs. In the CLC2000 inventory, a computer assisted visual interpretation method is applied (Büttner et al., 2002). In the CLC2006 inventory, a semi-automatic method is used that combines the CLC2000 and the photo interpreted

CLC changes, whereas the CLC2012 uses the semi-automatic methodology of object-orientated technology.

The CLC products are primarily used in areas of environment, research and education, and agriculture, which account for nearly 70 percent in total (Kleeschulte and Büttner, 2008). The investments in CLC2000 and CLC2006 are close to 13 Million Euro and 18 Million Euro, respectively (Kleeschulte and Büttner, 2008). Details on CORINE can be found in http://land.copernicus.eu/pan-european/corine-land-cover (accessed on 27 August, 2014).

2.4.4 LCCS (United Nations)

The Land Cover Classification System (LCCS) produced by the Food and Agricultural Organization (FAO) of the United Nations is constructed using a set of predefined independent diagnostic attributes (classifiers) instead of predefined names of classes on the assumption that any land classes can be identified by a combination of these attributes. The LCCS is an a priori classification system (Herold et al., 2006b). Potentially more than 200,000 classes can be derived from a combination of these classifiers (Herold et al., 2006a; Herold et al., 2006b). The number of attributes determines the detail with which a land class can be defined. A large number of attributes are required for a specific class, and a small number of attributes are necessary for a general class.

The classification, which is driven by pragmatic and operational considerations, greatly reduces the number of attributes required for definitions and significantly simplifies the classification procedure but is implemented in two phases: an initial (dichotomous) phase with eight major land cover types and a second (modular-hierarchical) phase that allows for specification of greater detail. Further classification can be achieved by adding environmental attributes (e.g., climate and land form) and specific attributes (e.g., floristic composition and crop type) (Herold et al., 2006b).

A land class is usually described by a mixture of covering materials rather than a single covering material. Four types of mixtures can be represented by the LCCS, as illustrated in Table 2.2 (Herold et al., 2006b). The spatial mixture A/B, which means that both A and B exist, is due to the Minimum Mapping Unit. The thematic mixture A//B means that the parcel could contain A or B. The temporal mixture A///B, which is usually applied to define agricultural classes, means that the parcel is A in one year and B in another. Temporal information can be obtained from auxiliary data, e.g., consulting records in the field or time series images. In practice, only A or B exists at a certain time in the imagery. The layering mixture A+B is used if different layers exist, e.g., agro-forestry. In practice, A or B, or A and B are present in the imagery in this situation.

Table 2.2 Mixture Class Represented by LCCS

Type	Notation	Example	Description
Spatial Mixture	/	A/B	Cartographic generalization: A and B
Thematic Mixture	//	A//B	Thematic generalization: A or B
Temporal Mixture	///	A///B	A in one year, B in another (normally for agriculture)
Laying Mixture	+	A+B	A in one layer, B in another (normally for agriculture and natural vegetation)

The LCCS is intended to be independent of map scale, data source, geographical location, and application, and allows for correlation of existing land classes through the use of predefined attributes, which makes it an optimal classification system to act as a standard reference classification system (Latham, 2008). The LCCS has been accepted and released by ISO as an international land cover classification standard and includes two separate sections with ISO numbers 19144-1 and 19144-2. The 19144-1 is a generic standard for classification systems, and the 19144-2 is a specific standard for LCCS.

A number of large-area land inventory projects have adopted the LCCS directly, and certain existing classification systems have been translated to align with the LCCS, e.g., IGBP (International Geosphere-Biosphere Programme), CORINE 2000, IPCC (Intergovernmental Panel on Climate Change), GLC2000 (Global Land Cover 2000), EOSD (Earth Observation for Sustainable Development of Forests), and Anderson Level I and II (Herold and Schmullius, 2004). However, specific translation inconsistencies remain, e.g., tree density of 60% (IGBP) versus 65% (LCCS) and minimum tree height of 2 m (IGBP) versus 3 m (LCCS) (Herold et al., 2006b). Detailed experience in mapping with LCCS can be found in http://www.glcn.org/dat_6_en.jsp (accessed on 27 August, 2014) and http://unstats.un.org/unsd/envaccounting/seeaLES/egm/Issue3_EEA_FAO.pdf (accessed on 27 August, 2014).

2.5 Problems with the Current Classification Systems

2.5.1 Principles of Designing a Land Classification System

Ideally, a land classification system should partition land in a manner that is mutually exclusive and completely exhaustive. It is obvious that a successful classification system should identify classes that are useful in predefined applications. Certain criteria for establishing such a classification system have been proposed by researchers (DRDLR, 2013, p.13; Di Gregorio, 2005, p.11). These technical criteria include:

- Consistent, unique, objective, quantitative and systematic classification principles should be used;
- The inherent character of land is a mixture of different covering types;
- A complete range of classifiers is defined using clear boundaries;
- Different classifiers should be used at different levels of the class hierarchy;
- The main or predominant covering types should be defined;
- The classification system should be easily applied and repeated.

- The classification system should be independent of scale, data collection tools and time factors.

- The classification system must be available to obtain data at different scales from different sources now and in the future.

However, these criteria are only guidelines that are proposed qualitatively. It is difficult to establish a system that completely conforms to these criteria. Furthermore, problems always exist in existing classification systems at least.

2.5.2 Semantic Problems

Semantic problems (e.g., overlaps and gaps in class definitions) have been qualitatively identified by many researchers (Herold et al., 2006b; Di Gregorio, 2005, p.8; Wyatt et al., 1997, p.46). Certain definitions are defined with rigorous quantitative boundary conditions that provide a basis for objective and repeatable classification, but others are defined with insufficient details, e.g., no other information than their names, which create the potential for misinterpretation (Wyatt et al., 1997, p.43).

Commonly, greater or fewer overlaps exist between classes in the existing LCSs (Wyatt et al., 1997, p.44). Based on the LCCS, inconsistencies within and among Anderson CS (ACS), CORINE CS, IGBP DISCover CS and UMd CS have been analyzed (Herold et al., 2009). The results illustrate that overlaps exist throughout the ACS, especially in the class 'Tundra'. The consistencies among the CORINE CS classes are better compared with those of ACS. The results are quite good for IGBP DISCover CS and UMd CS because the classes are more general and the number of classes is only one-half to one-third compared with those of ACS and CORINE CS. The more classes that exist, the larger the number of attributes that are employed to distinguish them and the more likely it is that inconsistencies and overlaps will occur between classes.

Semantic gaps also have been found in the LCS. For example, in the EarthSat GeoCover Global Land Cover system, a forest is defined as woody vegetation canopy >35% and height ⩾3 m, and shrubs are defined as height <3m. However, a classification for areas with woody vegetation canopy ⩽35% and height ⩾3 m is not defined (Herold and Schmullius, 2004, p.6). In a hierarchical structure, every level should be a partition of the whole. However, semantic problems exist between classes. It is reasonable to infer that classes of one level cannot be perfectly covered by another. In practice, this problem has concealed the observation that data of a lower level are not collected independently but are aggregated from higher levels that make a concrete match among levels.

Furthermore, semantic similarities between classes with a common nominal parent class may be smaller than those between classes with different parent classes. This situation betrays the common assumption in classification that inner-similarity should be larger than the inter-similarity. For example, in remote sensing, it is assumed that 'pixels within classes are spectrally more similar to one another than they are to pixels in other classes' (http://www.learnremotesensing.org/modules/image_classification/index.php?target=image_class accessed on 27 August, 2014). Examples of such problems will be demonstrated in Chapter 8.

2.5.3 Interoperability Problems

Many causes hinder interoperability between land data from different resources, e.g., different classification methods (e.g. remote sensing vs. field survey) and different storage formats (e.g. vector vs. raster). Among these factors, the greatest hindrance originates from differences between classification systems (Wyatt et al., 1997, p.31). Many land classification systems were defined by different information communities in different application areas for different purposes. However, most of these systems are incompatible which makes it impossible (at least directly) to interact between different classification products. Most of the common problems of semantic interoperability between classification systems are (Di Gregorio, 2011): Common classes that are named using different terms (synonyms) with the direct result that common entities are assigned to different class names in different systems, multiple understandings of homonymous classes (polysemy), and common classes that are arranged in different hierarchical levels. Semantic overlaps also exist between classes with different names. Semantic problems between different classification systems affect the process of interoperability and decrease the accuracy of outputs. It is likely that classification systems cannot be perfectly interchanged without losing information and generating inconsistencies.

A direct problem that should be emphasized occurs in sequential classification systems, e.g., the NLCD1992 CS and NLCD2001 CS. The land should be perfectly covered by the classes in a system, and the deletion of certain classes should be balanced by the addition of new classes. Therefore, it is an obvious issue to just delete the class ‘Non-Natural Woody’ from NLCD1992 CS to create the level-I classes of NLCD2001 CS. No proper answer can be offered to explain where these ‘Non-Natural Woody’ lands go except that mistakes exist in at least one of these two systems.

2.6 Conclusions

Additional land inventory projects are (or are planned to be) proposed, implemented, and completed in anticipation of various benefits resulting from land data, maturation of technology, and accumulated valuable experiences in history projects. Land mapping data that are spread around the world in different institutions are similar to a deep-buried gold mine waiting to be exploited. However, inconsistent classification systems significantly hinder this process. Semantic overlaps and gaps exist even within a classification system. Classification data produced by such problematic classification systems are certain to be of poor quality, which may be unworthy of the efforts of thousands of participants and billions of dollars spent over many years.

Research studies on the semantic aspects of classification systems are rare, although it has been demonstrated that progress on the quality of classification systems will make a huge contribution to (at least) practical applications. It is gratifying that certain efforts have been carried out to establish a global reference for land classes. Two of the most famous reference systems are LCCS and SRS (Semantic Reference System) (Kuhn, 2003). Both of these systems engage in building a top-level ontology and a standardized classification system for all international, regional, national, and local land classification systems using a top-down method. Although projects of top-level ontologies have (thus far at least) failed (Smith and Mark, 2001) because no single reference system will serve all possible application needs (ISO 19144, 2009). Achievements have been produced by LCCS despite the problems that remain (see Section 2.4.4). Indeed, LCCS is not yet a top-level ontology yet because not all existing land classes can be precisely represented.

Potential research topics on land classification systems include: (1) design scheme or how to build a classification system when definitions of classes change over time; (2) classification criteria that should especially be suited to current technology and methods, e.g., remote sensing; (3) presentation methods and whether hierarchical trees and other forms should be added in addition to definitions; and (4) evaluation methods or how to systematically measure the quality of a classification system.

A systematic study on the semantic uncertainties of land classification systems has been conducted in this thesis, and the results will be illustrated in the following chapters.

CHAPTER 3 PROBING CLASS DEFINITIONS: A NEW DISTINCTION

3.1 Introduction

Land classification data are fundamental to researches and applications in natural resources, environmental protection, food security, and successful humanitarian programmes (Di Gregorio, 2005). With the emergence of remote sensing and the maturation of technology, fast mapping of global, continental, national and regional land classifications have become possible and popular. A LCS (land classification system) that defines the meaning of every class plays a dominant role in a land classification project. The most popular LCSs include the Land Cover Classification System (LCCS) of FAO, the National Land Cover Dataset/Database (NLCD) Classification Schemes of the United States, the CORINE Land Cover (CLC) Class of Europe, and the Current Land Use Classification System (CLUC) of China. However, few works in the literatures address these definitions in the geospatial domain. Scholars are engaged in tackling interoperability problems from different engineering aspects but also must have insight into whether a definition is exactly described and what a definition exactly includes.

Definitions in natural language have been used for thousands of years and have been studied and discussed for long in such areas as philosophy, linguistics, and psychology from their respective perspectives. This study proposes a new distinction of definitions from an engineering perspective. The remainder of this chapter is structured as follows. First, a simulated example is constructed to help understanding and comparison of the contents introduced throughout this chapter. A classical differentiation between intension and extension is introduced in Section 3.2. Second, a new distinction is proposed and definitions are divided into concepts and categories in Section 3.3. Third, definitions are classified using new criteria, and details of what is included in a land class definition are discussed. Next, a case study of the NLCD 92 level-I class definitions is examined to demonstrate the new distinction, followed by a discussion. Finally, the conclusions end this chapter.

3.2 Intension and Extension

3.2.1 An Example

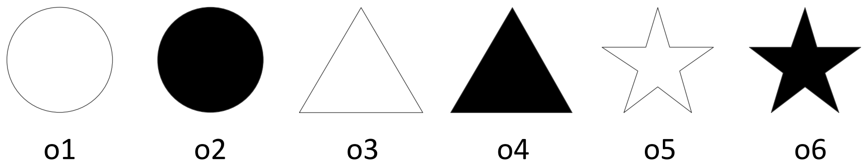


Figure 3.1 Example — A Universe with Six Graphs

Assume that the six graphs illustrated in Figure 3.1 (labelled from ‘o1’ to ‘o6’) constitute a universe. It is obvious that six objects exist (at least) that possess two ***properties***. One property (labelled as ‘p1’) can separate ‘o1’, ‘o3’, and ‘o5’ from ‘o2’, ‘o4’, and ‘o6’ but cannot make distinctions among ‘o1’, ‘o3’, and ‘o5’ or among ‘o2’, ‘o4’ and ‘o6’. Another property (labelled as ‘p2’) can separate ‘o1’ and ‘o2’, ’o3’ and ‘o4’, and ‘o5’ and ‘o6’ from each other, but cannot separate ‘o1’ from ‘o2’, ‘o3’ from ‘o4’, and ‘o5’ from’o6’. With a combination of ‘p1’ and ‘p2’, these six objects can be separated from each other.

The ***term*** ‘colour’ is allocated to the former property and ‘shape’ to the latter property for the purpose of storing and exchanging knowledge economically and conveniently, which simultaneously demonstrates that a term is a type of ***symbol*** for the reason that it can be replaced with other forms, e.g., the label ‘p1’ instead of the term ‘colour’. The property ‘colour’ has two ***attributes***: ‘a1’ (white) and ‘a2’ (black). Objects ‘o1’, ‘o3’, and ‘o5’ possess white and ‘o2’, ‘o4’, and ‘o6’ possess black. And another property ‘shape’ has three attributes: ‘a3’ (circle), ‘a4’ (triangle), and ‘a5’ (star). Objects ‘o1’ and ‘o2’ possess circle, ‘o3’ and ‘o4’ possess triangle, and ‘o5’ and ‘o6’ possess star.

By applying properties to the universe, it is partitioned into different ***classes*** that have been illustrated in Figure 3.2. In this example, two classes (with three objects in each class) are obtained by applying ‘p1’, three classes (with two objects in each class) are obtained by applying ‘p2’, and six classes (with only one object in each class) are obtained by applying ‘p1’ and ‘p2’ together.

class	attribute	object	graph	property
c1	a1	o1,o3,o5	○ △ ☆	p1
c2	a2	o2,o4,o6	● ▲ ★	p1
c3	a3	o1,o2	○ ●	p2
c4	a4	o3,o4	△ ▲	p2
c5	a4	o5,o6	☆ ★	p2
c6	a1,a3	o1	○	p1,p2
c7	a2,a3	o2	●	p1,p2
c8	a1,a4	o3	△	p1,p2
c9	a2,a4	o4	▲	p1,p2
c10	a1,a5	o5	☆	p1,p2
c11	a2,a5	o6	★	p1,p2

Figure 3.2 Classifying the Example

The use of the frontal example is twofold. First, although bold italic words (i.e., attribute and property, and term and symbol) are used interchangeably in most literatures, the result is that creating a rigorous definition for them is difficult or even impossible. This work applies a distinct usage of these words (instead of giving rigorous definitions) for the purpose of transferring information clearly and exactly. ***Term***, such as ‘land cover’, is a special case of ***symbol***, which can also include other forms, e.g., the legend of a map. ***Attribute*** is the value of a ***property***. And ‘common’ and ‘distinct’ are set to describe an attribute, whereas ‘same’ and ‘different’ are use to describe a property to create a distinguishing expression. Second, this example illustrates a simple classification process by applying two properties with six total attributes to create divisions for a six-object universe, which will simplify the description in following sections.

3.2.2 Intension, Extension and Symbol

The intension of a class refers to the set of attributes common to objects to which the class applies, and the extension of a class refers to all objects of that class which existed in the past, exist in the present, or will exist in the future. The intension and the extension of a class can be formalized based on formal concept analysis theory (Wille, 1992; Ganter and Wille, 1999). In a formal context, it holds that:

$$A' = B \text{ and } B' = A \tag{3.1}$$

where A and B are the intension and the extension of a class, respectively; A' is a set of objects possessing all attributes in A, and B' is a set of attributes shared by all objects in B. It indicates that the intension and the extension of a class determine each other.

In the geospatial domain, a class is usually a term applied from an aspatial perspective (e.g., in a land classification system) or a legend from a spatial perspective (e.g., in a land cover/land use map), both of which are a type of symbol for communicating conveniently and economically.

The relationships among intension, extension, and symbol are illustrated in Figure 3.3, and form a triangle. A solid line between intension (also extension) and symbol indicates a direct relationship, and a dashed line between intension and extension indicates there is no direct relationship. Symbol acts as a bridge to link intension and extension together. For example, when ‘Forested Upland’ of the NLCD 1992 project is mentioned, it refers to all attributes of its intension and all objects of its extension.

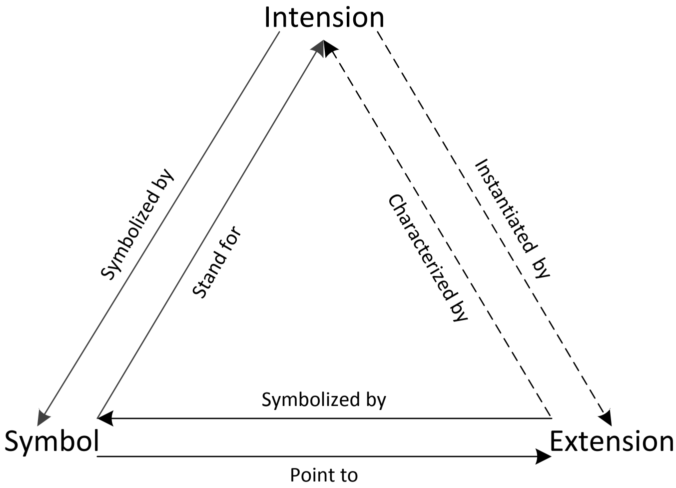


Figure 3.3 Relationships between Intension, Extension, and Symbol

3.2.3 From Extension to Intension

Attributes will not exist without objects. All attributes originate from objects of a class and will pass through a long period of testing of its ability to characterize this class. For example, radiometric characterization (Lin et al., 1999), spectral signature (Lam and Remmel, 2010), texture (Wood et al., 2012), and vegetation indices (Wagle et al., 2014) are all the results of characterizing objects in physical reality and are subsequently applied to remote sensing processes. Attributes that are available and sufficient to characterize the intended objects collection will be retained; otherwise, they will be modified or replaced according to the intended object collection until successful distinguishing ability is produced.

Figure 3.4 illustrates the procedure used to form a classification system, which describes the intended classes' information and acts as a standard and ontology to execute a land classification project.

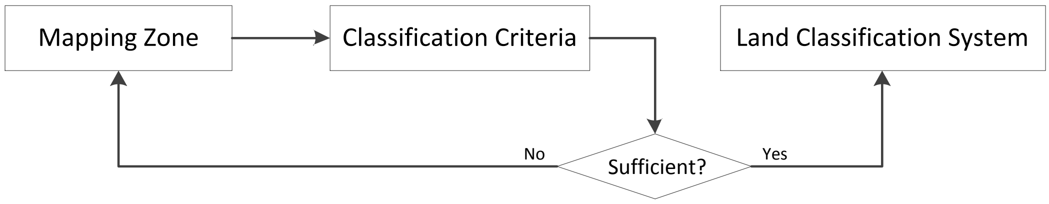


Figure 3.4 Procedure to Form a Classification System

First, the region in which the land classification will be executed should be defined. The mapping zone, which is a spatial region, limits the involved objects to a certain extent (global, continental, national, or regional) and makes the universe of objects fixed. Next, by characterizing the mapping zone, candidate classification criteria will be extracted according to the application purposes. If the land cover classified by the criteria fulfils the intended purposes, the classification criteria will be selected and preserved; otherwise, a modification will be applied. A pilot project is a mechanism used to test the sufficiency of the classification criteria.

In Figure 3.1, suppose that ‘o2’ and ‘o3’ constitute a mapping zone, which results in a universe that consists only of two objects. We want to separate ‘o2’ from ‘o3’. The candidate attributes, which are shape ‘circle’ and shape ‘triangle’, are selected and demonstrated to successfully classify this universe. Next, a possible example classification system described in natural language can be formed by these two attributes, as shown in Table 3.1.

Table 3.1 The Classification System of an Example

Class	Description
C3	Objects with circle shape.
C4	Objects with triangle shape.

3.2.4 From Intension to Extension

This process is an inversion of the former process (i.e., from extension to intension). Intension is not extracted for curiosity but to react to objects for different reasons, e.g., sensor design (Joseph, 1996), target recognition (Goel and Hsu, 1992), vegetation extraction (Liu and Yang, 2013), and land classification

(Hu and Wang, 2013). Intension is applied to objects for the intended purposes, and the results are always validated on the original objects. Generally, not all objects of an intended purpose can be perfectly grouped in the process, which leads to revisions of the intension for improvement.

Figure 3.5 illustrates the procedure for applying a land classification system to perform a land inventory, which is periodically or occasionally conducted by different international and national authorities.

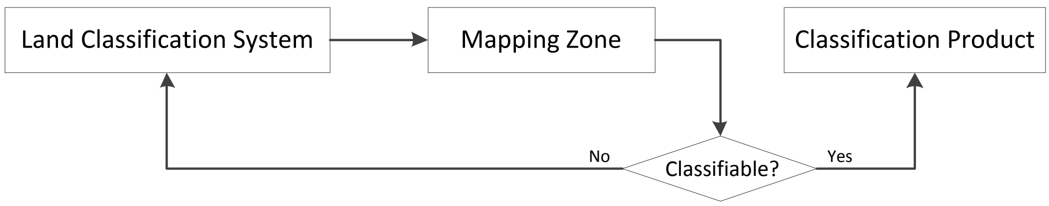


Figure 3.5 Procedure to Classify Land

A formal land classification system is released to guarantee the data produced by different organizations and participants consistent and act as ontology for information exchanges. However, it is obvious that it is difficult or even impossible to acquire a consistent result for certain types of land because of the complexity of land composition and the weakness of certain criteria in the classification system. In this case, the land classification system must be updated. Although both Figure 3.4 and Figure 3.5 offer options for improving the land classification system, the triggers are different. Both are based on the same judgment of the consistency between attributes and objects. The former is motivated intensionally by purposes, but the latter is motivated extensionally by objects.

Assuming that the former four objects in Figure 3.1 are intended to constitute the mapping zone, we wish to separate them from each other. The existing classification system described in Table 3.1 is used to conduct this work. However, the classification result, ‘o1’ and ‘o2’ in ‘C1’ and ‘o3’ and ‘o4’ in ‘C2’, are not sufficient. An adjustment to the classification system (which is a possible solution, as described in Table 3.2) is definitely required to satisfy the requirement. Using this adjustment, the intended classification products are obtained.

Table 3.2 The Modified Classification System of an Example

Class	Description
C6	Objects with circle shape and white colour.
C7	Objects with circle shape and black colour.
C8	Objects with triangle shape and white colour
C9	Objects with triangle shape and black colour.

3.2.5 A Loop between Intension and Extension

The intension and the extension of a symbol are not static and stable, but are changeable and dynamic for many reasons. For example, the intension and the extension of the term ‘road’ will be different among a planning department, a construction department, and a management department. A loop is formed between intension and extension by the connection of symbols, as illustrated in Figure 3.6. Note that symbols are differentiated by a superscript because the same symbol might stand for the intension or the extension.

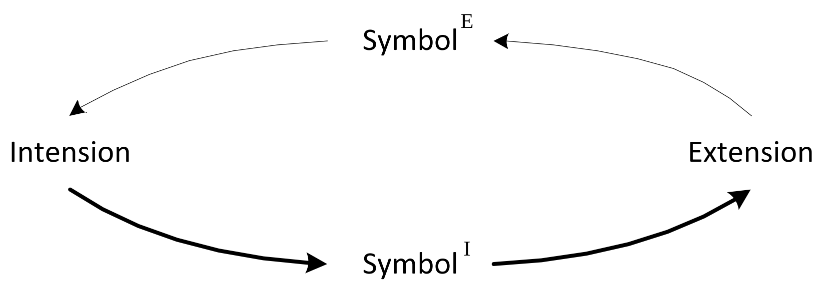


Figure 3.6 A Loop between Intension and Extension

Currently, this problem presents a significant challenge due to the accumulation of data and the requirement of exchanging information across disciplines, e.g., smart city construction. Practices and researches on harmonization and standardization (which are not topics of this research but require attention to the definitions involved) are engagements needed to reach a solution.

3.3 Concept and Category

3.3.1 Complete and Incomplete

Because the intension and the extension of a class are symbolized by one common symbol, it is a challenging or even impossible task to make an exact judgement of the meaning of the symbol. However, a practical refinement of symbols will benefit both applications and researches, e.g., evaluation of the semantic similarity between two sentences. Instead of separating the attribute symbol from the object symbol, for the purpose of quantification and comparison and as an analogue to constants and variables in mathematics, symbols are classified into two types, namely, complete and incomplete.

In a context, a symbol is incomplete if it requires other symbols for the complement but it is not certain to what extent they contribute; otherwise, a symbol is a complete symbol. Indeed, incomplete symbols will only emerge in a common property and normally the number of symbols of this property must be more than two.

For example, in the context of Figure 3.1, three class definitions are specified as in Table 3.3, and use three properties: colour, shape, and graph number.

Table 3.3 Class Definitions to Distinct Complete and Incomplete

Class	Description
C12	Containing two graphs and one is black and another is triangle.
C13	Containing two graphs and only one is black.
C14	Containing two graphs and one is triangle.

In ‘C12’, complementarity exists between ‘one’ and ‘another’, which should constitute ‘two’, but their contribution is concrete; therefore, they are all complete symbols. In ‘C13’, although complementarity exists between ‘one’ (black) and another colour in the context that only has two attributes, i.e., black and white, the complementarity is concrete as ‘one (white)’, which demonstrates that all are complete symbols. In ‘C14’, complementarity exists between ‘one (triangle)' and another shape, which is a choice among circle, star, and even triangle, which makes this symbol an incomplete symbol.

3.3.2 Relationship between Concept and Category

A concept is defined as a set of complete elements, and a category is defined as a set of elements in which at least one incomplete element exists.

A concept is always purely constructed of complete symbols, whereas a category is only a pure category in situations in which all of its symbols are incomplete symbols, which is quite rare in reality. Usually, a subset of a category is comprised of complete symbols that can be used to build other concepts, which makes a concept a component of a category. Together with changes of context, a complete symbol can change into an incomplete symbol and vice versa, which represents an interchange between concept and category. For example, presuming that the shape of all six objects is triangle in Figure 3.1, 'C14' will degrade to a concept from a category.

3.3.3 Complement to Intention and Extension

The characteristics of two distinctions related to symbol are listed in Table 3.4.

Intension and extension are essential properties of a symbol, and therefore they are stable. But it is rather difficult to make a separation between them and clearly express the reason for this separation. Intension and extension can be used to recognize synonyms and remove duplicated expressions based on the observation that intension and extension determine each other, as formulated in equation (3.1). Concept and category are easily differentiated from each other but may be interchangeable in the alteration of context. Concept and category can be used to determine whether the contents of an expression are fixed. In a comparison of the meanings of different symbols that are usually defined by definitions in natural language, a certain composition of contents is important but a redundant expression should be abandoned to ease the burden of unnecessary comparison. A combination is a solution for this case.

Table 3.4 Characteristics of Two Distinctions

	Intension and Extension	Concept and Category
Differentiable	Low	High
Context Influence	Low	High
Recognizing Synonym	Able	Unable
Recognizing Certainty	Unable	Able

3.4 Definition: A Mixture

A definition, which retains the meaning of a symbol, is essential for recording and exchanging knowledge in natural language. However, definitions do not receive sufficient attention and induce many semantic conflicts and misunderstandings. Land classification results may be criticized because classes labelled using a common term in different projects cannot be compared directly. This work does not systematically classify definitions into different types (interested readers can reference Kavouras and Kokla, 2008). The focus of this research is the proposed new distinction. Hence, only two types are discussed.

3.4.1 Intensional and Extensional Definition

A definition can be classified intensionally or extensionally (Kavouras and Kokla, 2008, p.114). Intensional definition specifies the properties of the objects to which a symbol pertains. The genus (common property) is inherited from a more general type of object that indicates the is-a relationship, whereas the differentiae (differentiated property) is applied to differentiate members of this type from other objects of the same genus. Extensional definition specifies the objects to which a symbol points by listing all (or selected) specific (or general) objects of a symbol. A general object is a group of objects of the same type. An example of listing all specific objects is defining the ‘permanent members of the United Nations Security Council’ by listing all five governments (specific objects) ‘China, France, Russia, the United Kingdom, and the United States’. An example of listing all general objects is defining ‘telephone’ by listing all of its general objects as ‘landline telephone and mobile phone’.

Land class definitions are prone to intensional definitions because they usually contain an unlimited number of objects (Kavouras and Kokla, 2008, p.174). Indeed, the number of attributes for any object is infinite. For example, in Figure 3.1, such properties as graph area, graph perimeter, angles of graph, and solid or hollow can be listed to characterize these graphs. Therefore, different attributes will be discovered from different perspectives, but they should be fixed in a certain context.

A popular example is a list of objects that possess clear properties of a class and is referred to as a prototype instead of all of this type. For example, robin, not penguin, is a prototype of bird. In the geospatial domain, certain prototypes are also used to reduce the burden of understanding in formal definitions.

3.4.2 Definition Based on Concept and Category

A definition also can be classified into concept definition and category definition. A concept definition is a set of concepts, and a category definition contains categories. A category usually contains a set of concepts which are composed of complete symbols of the category. Most of the rigorous definitions, e.g., definitions in a dictionary, are concept definitions. A category definition is popular in engineering applications, e.g., land classification projects, and is often used for description of its composition.

3.4.3 Land Class Definition

A land class in a classification system is normally defined using three components: code, term, and definition. A code is a number sequence of the same pattern in a level. Codes in the lower level are inherited by its subordinate classes or zeros are applied to fully fill the positions of its subordinate level. The function of the code is fourfold. First, a code acts as a symbol with the same meaning as the term. Second, a code is an index in a database that eases machine processing. Third, a hierarchical relationship is explicitly expressed. And four, because a code is not a common communication method, it will force users to figure out its meaning from definitions, which will reduce misunderstanding. Term is a word or phrase primarily used for human communication. Definition is the essential meaning of code and term, which contains all of the information needed.

Land class definition is not purely defined as influences of scale, context, and minimum mapping unit effects. One of the dominant characteristics is that land classes are composed of materials (artificial or natural) on the land (soil, tree, water, etc.). An area summary of these materials is equal to the area of the land. Another characteristic is that only dominant materials will be described in the definition for the vast variation of its composition. Therefore, a land class definition is a mixture of different types, which will be illustrated in the case study.

3.5 Case Study: NLCD1992 Level-I Class Definition

3.5.1 Class Definition

The NLCD1992 (National Land Cover Dataset 1992) classification system (classification scheme/legend), which is modified from the Anderson Land Cover Classification System (Anderson et al., 1976), was applied to the first national land-cover mapping project of the United States in the reference year 1992 and was completed in late 2000 (Vogelmann et al., 2001). The NLCD1992-CS (NLCD1992 Classification System) is a two-level hierarchical classification system with 9 level-I classes (Table 3.5) and 21 level-II classes designed to acquire information from satellite images (primarily Landsat TM).

3.5.2 Synonyms

Synonyms in these class definitions are listed in

Table 3.6. Although it is quite difficult to decide whether they are intension symbols or extension symbols, they obtain the same attributes and objects.

In land classification definitions, intensional synonyms are usually identified with a bracket, and extensional synonyms are usually enumerated as examples and begin with words, such as e.g., for example, etc.

Table 3.5 NLCD 1992 Level-I Class Definitions

Definition
10. **Water**—All areas of open water or permanent ice/snow cover.
20. **Developed**—Areas characterized by a high percentage (30 percent or greater) of constructed materials (e.g. asphalt, concrete, buildings, etc.).
30. **Barren**—Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with little or no "green" vegetation present regardless of its inherent ability to support life. Vegetation, if present, is more widely spaced and scrubby than that in the "green" vegetated categories; lichen cover may be extensive.
40. **Forested Upland**—Areas characterized by tree cover (natural or semi-natural woody vegetation, generally greater than 6 meters tall); tree canopy accounts for 25-100 percent of the cover.
50. **Shrubland**—Areas characterized by natural or semi-natural woody vegetation with aerial stems, generally less than 6 meters tall, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.
60. **Non-Natural Woody**—Areas dominated by non-natural woody vegetation; non-natural woody vegetative canopy accounts for 25-100 percent of the cover. The non-natural woody classification is subject to the availability of sufficient ancillary data to differentiate non-natural woody vegetation from natural woody vegetation.
70. **Herbaceous Upland**—Upland areas characterized by natural or semi-natural herbaceous vegetation; herbaceous vegetation accounts for 75-100 percent of the cover.
80. **Planted/Cultivated**—Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation accounts for 75-100 percent of the cover.
9. **Wetlands**—Areas where the soil or substrate is periodically saturated with or covered with water.

(Source: http://www.epa.gov/mrlc/definitions.html#1992 accessed on 29 July, 2014)

Table 3.6 Synonyms in this Case Study

20. **Developed**	a high percentage	30 percent or greater
	constructed materials	asphalt, concrete, buildings, etc.
40. **Forested Upland**	Tree	natural or semi-natural woody vegetation

3.5.3 Incomplete Symbols

An obvious incomplete symbol of a land class definition is the cover percentage of its components for which their sum always equals 1. Incomplete symbols in class definitions are listed in Table 3.7. Hence, there are six categories out of all nine definitions in contrast with the other three concepts.

Table 3.7 Incomplete Symbols in this Case Study

	Incomplete Symbol
20. **Developed**	a high percentage (30 percent or greater)
30. **Barren**	is more widely spaced and scrubby than
40. **Forested Upland**	25-100 percent
60. **Natural Woody**	25-100 percent
70.**Herbaceous Upland**	75-100 percent
80.**Planted/Cultivated**	75-100 percent

3.5.4 Concepts in Category

After removal of the incomplete symbols and synonyms, concepts in the category definitions are listed in Table 3.8.

Table 3.8 Concepts in the Category

20. **Developed**—Areas characterized by constructed materials.
30. **Barren**—Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with vegetation present regardless of its inherent ability to support life. Vegetation; lichen.
40. **Forested Upland**—Areas characterized by tree cover (natural or semi-natural woody vegetation, generally greater than 6 meters tall); tree canopy.
60. **Non-Natural Woody**—Areas dominated by non-natural woody vegetation.
70. **Herbaceous Upland**—Upland areas characterized by natural or semi-natural herbaceous vegetation. Herbaceous vegetation.
80. **Planted/Cultivated**—Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation

3.5.5 Concepts in NLCD 1992 Level-I Class Definitions

All concepts in NLCD 1992 Level-I class definitions are listed in Table 3.9.

Table 3.9 Concepts in NLCD 1992 Level-I Class Definitions

10.	**Water**—All areas of open water or permanent ice/snow cover.
20.	**Developed**—Areas characterized by constructed materials.
30.	**Barren**—Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with vegetation present regardless of its inherent ability to support life. Vegetation; lichen.
40.	**Forested Upland**—Areas characterized by tree cover (natural or semi-natural woody vegetation, generally greater than 6 meters tall); tree canopy.
50.	**Shrubland**—Areas characterized by natural or semi-natural woody vegetation with aerial stems, generally less than 6 meters tall, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.
60.	**Non-Natural Woody**—Areas dominated by non-natural woody vegetation.
70.	**Herbaceous Upland**—Upland areas characterized by natural or semi-natural herbaceous vegetation. Herbaceous vegetation.
80.	**Planted/Cultivated**—Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation
9.	**Wetlands**—Areas where the soil or substrate is periodically saturated with or covered with water.

3.6 Discussion

3.6.1 Ontology

A land classification system is a domain ontology, and every level should be a partition of land that makes it different from other ontologies. Concepts in these definitions are materials that cover the land. For example, one of the ‘Forested Upland’ materials is ‘greater than 6 meters tall natural or semi-natural woody vegetation’, which at least involves the concepts listed in Table 3.10.

All concepts listed all have the opportunity to act as a component of land, which means that most of the land classes are categories. However, it is impossible to be certain of the components, i.e., from one perspective, only characteristic components are of interest, but from another perspective, we cannot know what will be contained in that class unless it is applied in practice. Decomposition of definitions into concepts is a possible solution because any concept is a possible component, at least to the extent of what we know. Other components that are not involved in definitions can only be determined in practice and subsequently judged on whether they should be contained. Therefore, concepts and categories can be used to find the gaps in the ontology.

Table 3.10 An Example of Concepts Involved

(1) vegetation
(2) woody vegetation
(3) natural or semi-natural vegetation
(4) greater than 6 meters tall vegetation
(5) natural or semi-natural woody vegetation
(6) greater than 6 meters tall woody vegetation
(7) greater than 6 meters tall natural or semi-natural vegetation
(8) greater than 6 meters tall natural or semi-natural woody vegetation

3.6.2 Semantic overlap

Another issue that should be noted is an intersection between different classes. The occurrence of this problem must be forbidden from a definition for the reason that objects of intersected concepts can be classified into any of those classes. A firm answer (and not a choice) is needed to the question of a partition of land. Overlaps among concepts can be tackled by decomposing them into details, but the questions that remain are how to decompose and to what extent. Finding the overlaps among categories and figuring out how to convert incomplete symbols into complete symbols are the essential issues. After conversion, the problem becomes a concept issue.

3.6.3 Semantic similarity

A third issue of concern is semantic similarity, which plays a key role in the classification process. Normally, similarities between subclasses of a common superordinate class should be larger than the similarities between one class and another not inherited with the superordinate class. However, classes are usually hierarchically structured using expert knowledge and experience, and no scientific calculations are employed in evaluation. The same situation also applies to land classes in a classification system. This situation creates a large problem if the intra-class similarity (the similarity between subclasses of a common superordinate class) is smaller than an inter-class similarity (the similarity between a subclass of a superordinate class and a class not inherited from the superordinate class) because nearly all classification algorithms in the geospatial domain are based on the assumption that intra-class similarity is larger than inter-class similarity. The question can only be answered after a scientific calculation of the concept classes and category classes in a classification system.

3.7 Conclusion

Although concept and category are commonly used synonymously, rigorously defining their meanings is difficult and even impossible (Smith, 2004; Frumkina and Mikhejev, 1996). Therefore, this work proposes to distinguish these terms using new criteria: complete symbol and incomplete symbol. Compared with the distinction via intension and extension, this distinction is better for computation, comparison, and operation. Using the example of the NLCD 1992 level-I class definitions, a real distinction is analysed and reveals that the distinction of concept definition and category definition is a possible solution that can be used to tackle such problems as semantic gaps in a classification system, semantic overlaps between classes, and semantic similarities between classes. However, methods for scientific evaluation of these problems are not proposed and are topics of future studies that will follow this chapter.

CHAPTER 4 FORMALIZATION OF CONCEPT

4.1 Introduction

A concept is normally defined in a dictionary, a standard, and even a project. Concepts are defined, interchanged, and understood between participants, institutions and disciplines. The meaning of a concept does not rely on what it is named but what is defined in its definition. The concept discussed in this chapter is constructed entirely by complete symbols in contrast to a category that includes at least one incomplete symbol (see Chapter 3). Although concepts are necessary in any discussion, a concept is treated as common sense in a domain or is simply defined by natural language. Certain efforts have been made to formalize a concept as a tube. However, definitions are complex in practice and not constrained to a fixed style, which makes the tube model adaptable to a limited situation. We propose a novel formalization of concepts in this chapter. This model is designed to adapt to all types of concepts.

The objectives of this chapter are:

- To classify concepts into four types.
- To propose a new method for modelling concepts using product and union operations.
- To reveal and quantify characteristics of concepts in the novel model.
- To demonstrate that the meaning of a concept is usually unchanged, although the order or combination of its constitute symbols may be altered.

The remainder of this chapter is structured as follows. Section 4.2 first presents a new classification for concepts and subsequently proposes a method that is able to model all types of concepts. Section 4.3 describes the characteristics of concepts in the proposed model and especially in a classification system. Two examples, i.e., a simulated classification system for triangles and a practical classification system for land inventory, are applied as case studies in Section 4.4.

Section 4.5 presents a discussion on the proposed model based on the results of the case studies. Finally, conclusions are outlined in Section 4.6.

4.2 Modelling Concept

4.2.1 Types of Concept

Based on whether complete symbols are possessed mutually or separately, concepts can be grouped into four types: Single-concept (S-concept), Union-concept (U-concept), Join-concept (J-concept) and Complex-concept (C-concept). Because all symbols in a concept are complete symbols rather than incomplete symbols, symbol(s) is used to represent a complete symbol(s) for simplification in the following description if it is not re-declared. An S-concept is a concept for which the meaning is defined by only one symbol. A U-concept is a concept for which the meaning is a union of every single symbol; in other words, any symbol in a U-concept is this type of concept. A J-concept is a concept for which the meaning is defined mutually by all symbols; in other words, any partial of a J-concept is not this type of concept. A C-concept is a concept for which a partial of symbols in this concept is a type of this concept.

The relationships among these concept types are illustrated in Figure 4.1. A C-concept is reduced to a U-concept if every symbol in this C-concept is a type of the C-concept, whereas it is reduced to a J-concept if only one partial of symbols (indeed it is the entirety of symbols) in this C-concept is a type of the C-concept. A U-concept is reduced to an S-concept if there is only one symbol in this concept, and the same applies to a J-concept.

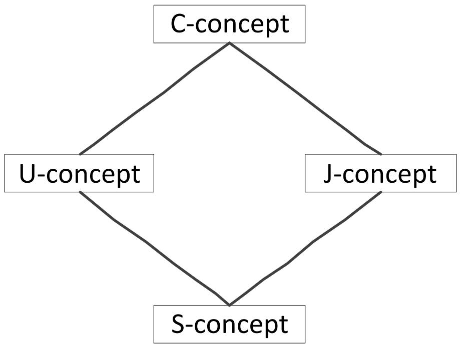


Figure 4.1 Relationships among Different Concept Types

4.2.2 Generic and Unique Representation of Concept

Two operations, i.e., the product and the union, are used to mutually or separately characterize symbols in a concept. The product operation is used if the meaning of a concept is mutually defined by symbols, and the union operation is used if the symbols on both sides of the operation are a type of the concept. The order of symbols in a concept may be important in situations for emphasis, but in most cases, the order is insignificant, or in other words, the shift of symbol order does not change the meaning of a concept. Therefore, the product discussed in this research is an unordered product. However, the method can be applied to an ordered situation, if a Cartesian product is considered.

Until now, assuming a and b are two symbols, the product of a and b is formalized by $a \,\&\, \mathrm{b}$, which results in $\left(a, \mathrm{b}\right)$, where $\&$ is the unordered product operation, and the union of a and b is formalized by $a \cup \mathrm{b}$, which results in $\left\{a, \mathrm{b}\right\}$, where $\cup$ is the union operation. Hence, a parenthesis $(\;)$ indicates that elements within it mutually define the concept, and a brace $\{\;\}$ indicates that every element within it is a type of the concept. The priority of the product operation is higher than that of the union operation. In other words, $a \cup \mathrm{b} \,\&\, c = a \cup \left(\mathrm{b} \,\&\, c\right)$, where c is a symbol as well.

Each type of concept can be denoted as:

$$C^S = a \tag{4.1}$$

$$C^U = \bigcup a_i \tag{4.2}$$

$$C^J = \& a_i \tag{4.3}$$

$$C^C = \bigcup \& \ a_{ij} \tag{4.4}$$

where C represents a concept that is refined by its superscript: C^S is an S-concept, C^U is a U-concept, C^J is a J-concept, and C^C is a C-concept, and a, a_i, a_{ij} are symbols used to define a concept.

For the reasons that every single symbol is a S-concept, formula (4.1) to (4.4) can be rewritten into:

$$C^S = C^S \tag{4.5}$$

$$C^U = \bigcup C_i^S \tag{4.6}$$

$$C^J = \& C_i^S \tag{4.7}$$

$$C^C = \bigcup \& \ C_i^S \tag{4.8}$$

Therefore, a concept always can be decomposed into a collection of S-concepts and rebuilt by product and union operations on these S-concepts. The S-concepts in a concept are unique. A unique representation of a concept can be denoted as:

$$C = \bigcup \& \ C_i^S \tag{4.9}$$

Any concept can be used to define another concept and hence a general representation of a concept can be denoted as:

$$C = \bigcup \& C_i \qquad (4.10)$$

In a broad sense, a concept is a S-concept if it is built without applying either a product operation or a union operation. A concept is a U-concept if it is built only using union operations. A concept is a J-concept if it is built only by product operations. A concept is a C-concept if it is built by applying both product operations and union operations. Hence, broadly speaking, a S-concept can also consist of many symbols. For example, the level-III classes in Figure 4.3 are broad-sense S-concepts, although many symbols are involved in every concept.

4.3 Characteristics of the Novel Representation

4.3.1 Fundamental characteristics

For all concepts A, B, and C, the following formula applies.

$$A \bigcup A = A \textit{ and } \quad A \& A = A \qquad (4.11)$$

The union of a concept and itself is always equal to the concept. The product of a concept and itself is always equal to the concept as well. In other words, the meaning of a concept will be not altered no matter what duplication of a symbol is added into a concept or which synonym is removed from the concept.

$$A \bigcup \Phi = A \textit{ and } \quad A \& \Phi = \Phi \qquad (4.12)$$

In formula(4.12), Φ is an empty concept, which means inexistence. One possible method for obtaining an empty concept is to define a concept self-contradictorily. For example, a concept M defined as ‘a type of vegetation belonging both to woody vegetation and herbaceous vegetation’ is an empty concept. Suppose a concept A is defined as ‘a type of vegetation taller than 6 meters’. Therefore, $A \bigcup M = A \bigcup \Phi$ is the concept defined by ‘a type of vegetation taller than 6 meters or a type of vegetation belonging both to woody vegetation and herbaceous vegetation’, which is identical to A, and $A \& M = A \& \Phi$ is the concept defined by ‘a type of vegetation taller than 6 meters and belonging to both woody vegetation and herbaceous vegetation’ which is also an empty concept.

$$A \cup B = B \cup A \text{ and } \quad A \,\&\, B = B \,\&\, A \tag{4.13}$$

$$\left(A \cup B\right) \cup C = A \cup \left(B \cup C\right) \text{ and } \quad \left(A \,\&\, B\right) \,\&\, C = A \,\&\, \left(B \,\&\, C\right) \tag{4.14}$$

Formula (4.13) and formula (4.14) indicate that the union operation and the product operation follow the commutative law and associative law. The meaning of a U-concept (or J-concept) remains unchanged no matter how the order is changed or how symbols are combined together. If the order of symbols is considered, formula (4.13) is broken up, but formula (4.14) is still retained.

$$\left(A \cup B\right) \,\&\, C = \left(A \cup C\right) \,\&\, \left(\mathrm{B} \cup C\right) \tag{4.15}$$

Formula (4.15) indicates the distributive law of product. As mentioned in Section 4.2.2, the priority of a product is higher than a union. Hence, formula (4.9) can be rewritten as:

$$C = \cup C_i^J \tag{4.16}$$

which means that a concept can always be decomposed into a union of J-concepts.

These characteristics are fundamental to prove that although the definition of a concept may be decomposed into a different number of components with the order of symbols changed or unchanged, the meaning is persistent and identical. Hence, it is feasible to decompose a concept from its definition and gain insight into the concept without changing its meaning.

4.3.2 Characteristics in a Classification System

4.3.2.1 Superordinate Concept and Subordinate Concept

A superordinate concept is a more general concept of the base concept, whereas a subordinate concept is a more specific concept of the base concept. From an extensional perspective, all objects that belong to a subordinate concept must be objects of its superordinate concept. From an intentional perspective, all attributes of a superordinate concept must be inherited by its subordinate concept. A more general concept is produced by sequentially applying union operations. A more specific concept is produced by sequentially applying product operations. The symbol $\subseteq$ is applied to denote A as the subordinate concept of B if it holds that $A \subseteq B$. Thus, B is the superordinate concept of A which holds that $B \supseteq A$. If $A \subseteq B$ and $B \subseteq A$, we say that A equals B which is denoted by $A = B$. In other words, the meaning of A is identical to the meaning of B if $A = B$.

For U-concepts, if $A \subseteq B$, then it holds that

$$B = A \cup C \tag{4.17}$$

For J-concepts, if $A \subseteq B$, then it holds that

$$A = B \,\&\, C \tag{4.18}$$

For C-concepts, if $A \subseteq B$, then it holds that

$$A_i^J \subseteq B_j^J \tag{4.19}$$

where C is a concept, A_i^J and B_j^J are one of J-concepts of A and B, respectively, and A_i^J and B_j^J can be obtained based on formula (4.16).

4.3.2.2 Hierarchical Classification System

A hierarchical classification system is an ordered set $(P, \subseteq)$, where P is a finite set of all classes in a classification system that holds reflectivity, anti-symmetry, and transitivity.

（1） $\forall x \in P, x \subseteq x$ (reflectivity)

（2） $\forall x, y \in P$, if $x \subseteq y$ and $y \subseteq x$, then $x = y$ (anti-symmetry)

（3） $\forall x, y, z \in P$, if $x \subseteq y$ and $y \subseteq z$, then $x \subseteq z$ (transitivity)

In a classification system $(P, \subseteq)$, there exists one and only one most general class that is the superordinate class of all classes in P but is usually declared implicitly as common sense in an application domain, e.g., in land classification. This most general class is known as the root class and is denoted as R. Every level of a hierarchical classification system should be a partition of R, which means that all classes in the same level should be non-overlapped and non-empty classes and the union of these classes should be identical to R.

（1） $\forall x \in P^L,\ x \neq \Phi$

（2） $\forall x, y \in P^L$, if $x \neq y$, then $x \cap y = \Phi$

（3） $\forall x \in P^L, \bigcup x = R$

where P^L is a set of all classes in a same level of $(P, \subseteq)$.

The classes in a lower level are inherited by the higher level classes. The number of classes in a lower level is smaller than that in a higher level. A class in any level can be constructed by sequential application of product operations to the root class (entirely or partially) or by application of union operations to all of its subordinate classes in the highest level. We propose two indicators to quantify the position of a class in a classification system: the depth and the width.

The depth of a class is the number of product operations applied to the root class which is denoted by $\left|\ \right|^R$. A larger number for depth indicates that additional insights are offered into the class. The width of a class is the number of subordinate classes in the highest level of the class which is denoted by $\left|\ \right|^L$. A larger number for width indicates that additional refinements are divided from the class.

The depth of a superordinate class is no larger than that of the base class, whereas the width of a superordinate class is no smaller than that of the base class, and vice versa for a subordinate class.

$\forall x, y \in P$, if $x \subseteq y$, then $\left|x\right|^R \geq \left|y\right|^R$ and $\left|x\right|^L \leq \left|y\right|^L$

Moreover, the sum of the depths of all classes in a lower level is smaller than that in a higher level, and the sum of widths of all classes in any level are identical to each other.

$\forall x \in P^{L1}, y \in P^{L2}$, if $L1 < L2$, then $\sum \left|x\right|^R \leq \sum \left|y\right|^R$

$\forall x \in P^{L1}, y \in P^{L2}$, $\sum \left|x\right|^L = \sum \left|y\right|^L$

4.4 Case Study

We use two examples to illustrate the novel formalization of concepts. One example is a simulated classification system used to classify triangles, and the other is a practical classification system for land inventory. The reason for constructing a simulated classification system is twofold: concepts in a practical classification system in the geospatial domain (a) are usually complex, and (b) are prone to involving more or less semantic problems, e.g., semantic overlap. The simulated example is used to illustrate the formalization of a concept using different types and to demonstrate the characteristics in a classification system.

The practical example is modelled to test the novel formalization in reality and to reveal issues that should be examined further.

4.4.1 Case I: A Simulated Classification System

A three-level hierarchical classification system used to classify triangles is presented in Figure 4.2. There are two classes in level-I, four classes in level-II, and eight classes in level-III. Assuming the colour of the triangles can be only black or white, every level is a partition of triangles.

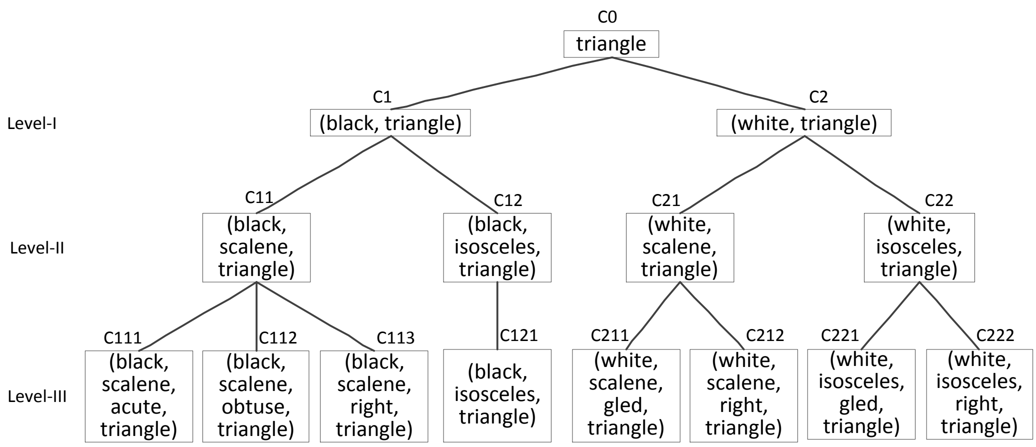


Figure 4.2 A Simulated Hierarchical Classification System

4.4.1.1 Types of Concepts

Concepts in Figure 4.2 are presented as J-concepts because every concept is defined by refinement of its superordinate concept. Every concept also can be defined by the union of all of its subordinate concepts, as illustrated in Figure 4.3, the results of which is that every concept is a U-concept. The corresponding concepts in Figure 4.2 and Figure 4.3 are identical to each other. Therefore, these concepts can replace each other without changing their meanings. More generally, a concept can be defined jointly by the concepts in Figure 4.2 and Figure 4.3, which is the result of product and union operations. For example, $C1$ can be defined by {(black, scalene, triangle), (black, isosceles, triangle)}, which is a C-concept.

This example demonstrates that the type of concept is not fixed but is determined by how it is defined. Meanings of different types of concepts can be identical.

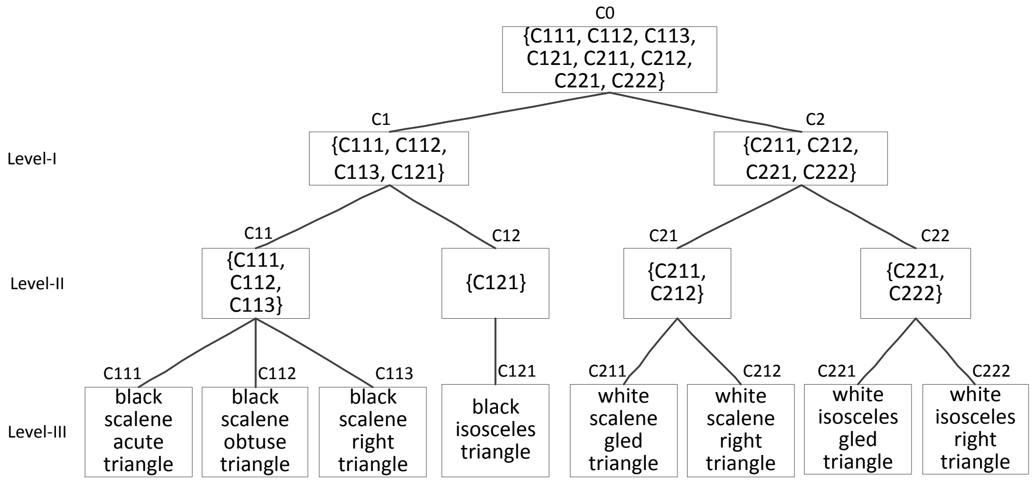


Figure 4.3 The Same Classification System defined differently

4.4.1.2 Superordinate and Subordinate Concepts

Formula (4.18) is illustrated in Figure 4.2, which shows that every subordinate concept is built by applying only product operations on its superordinate concept. Formula (4.17) is illustrated by Figure 4.3, which shows that every superordinate concept is constructed by applying only union operations on its subordinate concepts. For instance, $C1$ is defined as the union of two J-concepts in Section 4.4.1.1, and $C1 = \left\{C_1^J, C_2^J\right\}$, where C_1^J = (black, scalene, triangle), and C_2^J = (black, isosceles, triangle). We rebuild $C11$ into a C-concept as $C11 = \left\{C_3^J, C_4^J, C_5^J\right\}$, where C_3^J = (black, scalene, acute, triangle), C_4^J = (black, scalene, obtuse, triangle), and C_5^J = (black, scalene, right, triangle). We state that $C_3^J \subseteq C_1^J$, $C_4^J \subseteq C_1^J$, and $C_5^J \subseteq C_1^J$, which follow formula (4.19).

4.4.1.3 Depth and Width in the Classification System

The depths and widths of the classes in Figure 4.2 are listed in Table 4.1. It is obvious that the depth of a superordinate concept is no larger than any of its subordinate concepts, and the width of a superordinate concept is no smaller than any of its subordinate concepts. The sums of the depths in level-I, level-II, and level-III are equal to 4, 12, and 31, respectively, and are larger at a higher level.

The sums of the widths in level-I, level-II, and level-III are equal to 8, 8, and 8, respectively, which are identical at different levels.

Table 4.1 Depth and Width of Classes in the Simulated Classification System

<table>
<tr><td colspan="8">C1</td><td colspan="8">C2</td></tr>
<tr><td colspan="4">2</td><td colspan="4">4</td><td colspan="4">2</td><td colspan="4">4</td></tr>
<tr><td colspan="6">C11</td><td colspan="2">C12</td><td colspan="4">C21</td><td colspan="4">C22</td></tr>
<tr><td colspan="3">3</td><td colspan="3">3</td><td>3</td><td>1</td><td colspan="2">3</td><td colspan="2">2</td><td colspan="2">3</td><td colspan="2">2</td></tr>
<tr><td colspan="2">C111</td><td colspan="2">C112</td><td colspan="2">C113</td><td colspan="2">C121</td><td colspan="2">C211</td><td colspan="2">C212</td><td colspan="2">C221</td><td colspan="2">C222</td></tr>
<tr><td>4</td><td>1</td><td>4</td><td>1</td><td>4</td><td>1</td><td>3</td><td>1</td><td>4</td><td>1</td><td>4</td><td>1</td><td>4</td><td>1</td><td>4</td><td>1</td></tr>
</table>

Notes: Depths are in greyed cells while widths are in blank cells.

4.4.2 Case II: Concepts in NLCD2001 Level-I Class Definition

The NLCD2001 classification system (classification scheme/legend) is a two-level classification system with 8 level-I classes (Table 4.2) and 16 level-II classes. Compared with the NLCD1992 classification system, the Level-II classes of the NLCD2001 classification system have been slightly modified, and the Level-I classes are inherited consistently except for abandonment of the class ‘Non-Natural Woody’ from the NLCD1992 classification system. The NLCD2001 classification system has been inherited without any modification by the NLCD2006 and NLCD2011.

4.4.2.1 Concepts in Classes

Concepts are constructed by complete symbols, and categories consist of incomplete symbols. Selected symbols in NLCD2001 Level-I classes are incomplete symbols, e.g., area percentage. After eliminating the incomplete symbols, the concepts in level-I classes are listed in Table 4.3.

4.4.2.2 Representation of Concepts

An S-concept is a symbol in a narrow sense or a concept in a broad sense. In natural language, the former is a word, and the latter is a meaningful word, phrase, or even a sentence that is essential to the definition.

It is obvious that a narrow representation is a straight formalization of a concept. In Table 4.3, the concept of 'water' can be formulated by:

{All}&{areas}&{of}&({open}&{water}⋃({permanent}&({ice}⋃ {snow})))&{cover}

Table 4.2 NLCD 2001 Level-I Class Definitions

Definition
10. **Water**—All areas of open water or permanent ice/snow cover.
20. **Developed**—Areas characterized by a high percentage (30 percent or greater) of constructed materials (e.g. asphalt, concrete, buildings, etc.).
30. **Barren**—Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with little or no "green" vegetation present regardless of its inherent ability to support life. Vegetation, if present, is more widely spaced and scrubby than that in the "green" vegetated categories; lichen cover may be extensive.
40. **Forested Upland**—Areas characterized by tree cover (natural or semi-natural woody vegetation, generally greater than 6 meters tall); tree canopy accounts for 25-100 percent of the cover.
50. **Shrubland**—Areas characterized by natural or semi-natural woody vegetation with aerial stems, generally less than 6 meters tall, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.
70. **Herbaceous Upland**—Upland areas characterized by natural or semi-natural herbaceous vegetation; herbaceous vegetation accounts for 75-100 percent of the cover.
80. **Planted/Cultivated**—Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation accounts for 75-100 percent of the cover.
9. **Wetlands**—Areas where the soil or substrate is periodically saturated with or covered with water.

(Source: http://www.epa.gov/mrlc/definitions.html#2001), and 'Wetland' is inherited from NLCD 1992 Classification Scheme.

Furthermore, this concept can be decomposed into a union of three J-concepts.

J-concept 1: {All}&{areas}&{of}&{open}&{water}&{cover}

J-concept 2: {All}&{areas}&{of}&{permanent}&{ice}&{cover}

J-concept 3: {All}&{areas}&{of}&{permanent}&{snow}&{cover}

Table 4.3 Concepts in NLCD 2001 Level-I Classes

	Concept
10. **Water**	All areas of open water or permanent ice/snow cover.
20. **Developed**	Areas characterized by constructed materials (e.g. asphalt, concrete, buildings, etc.).
30. **Barren**	Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with "green" vegetation present regardless of its inherent ability to support life. Vegetation, "green" vegetated categories, lichen
40. **Forested Upland**	Areas characterized by tree cover (natural or semi-natural woody vegetation, generally greater than 6 meters tall). tree canopy
50. **Shrubland**	Areas characterized by natural or semi-natural woody vegetation with aerial stems, generally less than 6 meters tall, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.
70. **Herbaceous Upland**	Upland areas characterized by natural or semi-natural herbaceous vegetation. herbaceous vegetation
80. **Planted/Cultivated**	Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation
9. **Wetlands**	Areas where the soil or substrate is periodically saturated with or covered with water.

This type of representation is definitely unique, but certain of the S-concepts are only structural elements (e.g., ‘of’) that will not alter the meaning if they are deleted. Hence, ‘water’ can be represented by:

$\{\text{open water}\} \cup \{\text{permanent ice}\} \cup \{\text{permanent snow}\}$

which is a U-concept built on three broad-sense S-concepts. In land classification systems, land classes are defined by covering materials on the land. However, it is obvious that it is impossible to enumerate all coverings of a class. Of all covering materials, certain ones play characteristics of a class and are normally defined explicitly to separate them from other classes. Characteristic covering materials are usually similar to the literal meaning of the class name. For example, ‘open water’, ‘permanent ice’, and ‘permanent snow’ are characteristic covering materials of ‘water’.

Characteristic covering material concepts in the NLCD2001 level-I classes are listed in Table 4.4.

Table 4.4 Representation of Characteristic Covering Concepts in NLCD2001 Level-I Classes

	Characteristic Covering
10. **Water**	{open water} ⋃ {permanent ice} ⋃ {permanent snow}
20. **Developed**	{constructed materials}
30. **Barren**	{bare rock} ⋃ {gravel} ⋃ {sand} ⋃ {silt} ⋃ {clay} ⋃ {other earthen material}
40. **Forested Upland**	{greater than 6 meters tall}&{natural or semi-natural woody vegetation}
50. **Shrubland**	{less than 6 meters tall}&{ natural or semi-natural woody vegetation }
70. **Herbaceous Upland**	{natural or semi-natural herbaceous vegetation}
80. **Planted/Cultivated**	{planted}&{herbaceous vegetation} ⋃ {intensively managed}&{herbaceous vegetation}
9. **Wetlands**	{periodically saturated with or covered with water}

Examples and detailed explanations of definitions are not included in representations. For example, ‘asphalt, concrete, buildings, etc.’ are examples of ‘constructed materials’ and ‘for the production of food, feed, or fibre; or is maintained in developed settings for specific purposes.’ is an explanation of the purpose of ‘Planted/Cultivated’.

4.5 Discussion

4.5.1 Is-a and Part-of Relationship

Two most important relationships associated with concepts in the geospatial domain are the is-a (i.e., inclusion, type of, subtype/supertype, hyponymy/hypernymy) relationship and the part-of (i.e., part-whole, partonomy, meronymy/holonymy) relationship. Suppose A is a more specific concept compared with B. If A and B retain the is-a relationship, then A can be named the same as B, whereas A cannot possess the same name as B if the part-of relationships exists between them. For example, ‘black triangle’ and ‘triangle’ retain the is-a relationship. If A is a ‘black triangle’, then A is a ‘triangle’ but not vice versa. Similarly, ‘edge’ and ‘triangle’ retain the part-of relation. If A is an ‘edge’, we cannot say that A is a ‘triangle’ and vice versa. The is-a relationship is normally based on properties, but the part-of relationship is normally based on objects.

Hence, a concept can be defined based on an is-a relationship by adding more attributes to its superordinate concept, or a concept can be defined based on a part-of relationship by listing all parts of the concept. For example, a ‘red sports car’ can be defined by adding the attributes ‘red’ and ‘sports’ to ‘car’, which is defined based on the is-a relationship. A ‘car’ consists of ‘engine’, ‘chassis’, ‘wheel’, ‘body’, and ‘electrical system’, which are defined based on the part-of relationship. The first concept can be represented by (red, sports, car), whereas the second concept can be represented by (engine, chassis, wheel, body, electrical system). Although both of these concepts are built using product operations, the difference is that there must be at least one superordinate concept implicitly contained in an is-a based definition, whereas both superordinate and subordinate cannot appear in a part-of based definition. For example, (red, car), (sports, car) and (car) are superordinate concepts of (red, sports, car).

The parts in a part-of based definition that employs product operations should be differentiated with symbols of the applied union operations in the U-concepts and C-concepts. All parts connected by union operations are a type of the concept. For example, (engine, chassis, wheel, body, electrical system) is a concept of ‘car’, and {engine, chassis, wheel, body, electrical system} could be a concept of ‘car accessories’. Additional details are depicted if additional product and union operations are employed. In a classification system, the union operation increases the number of classes horizontally, and the product operation increases the number of classes vertically. Concepts are usually defined based only on is-a relationships, and categories usually contains part-of relationships.

4.5.2 Concept Types and Quantification

It has been demonstrated that the meaning of a concept can persist no matter what types are contained in a concept. However, the form of the representation reveals the perspective of the definer. For example, the car concept represented as (red, sports, car) is different from (red sports, car). The former treats ‘red’ and ‘sports’ as two attributes and the latter takes ‘red sports’ as one attribute. In other words, (red sports, car) is suitable in situations in which there is no ‘red car’ or ‘sports car’ other than ‘red sports car’. Moreover, quantifications of concepts vary among different concept types. For example, the depth of (red, sports, car) is two, but it is one for (red sports, car).

Hence, to quantify concepts, the form of representation should be fixed to acquire a unique value. One possible solution is to release a standard to guide the decomposition of definitions. The advantage of this method is that an absolute unique value can be obtained, which is beneficial to global interoperability and comparison. However, this type of standard is only theoretically feasible. It is actually a type of top ontology, which all has failed, at least until recently. Another possible solution is altering the representation of concepts when needed to address that component of the concept, e.g., if a comparison is required or special interest is focused on that part. This method is feasible both theoretically and practically, however, this method is context aware. Quantification values may change if new considerations are included. For example, if only $C1$ and $C111$ in Figure 4.2 are considered, $C1$ and $C111$ can be represented as (black triangle) and (scalene acute, black triangle), respectively. If $C11$ is also considered, then $C1$, $C11$, and $C111$ can be represented as (black triangle), (scalene, black triangle) and (acute, scalene, black triangle), respectively.

4.5.3 Universe of Property

In practice, symbols, which describe the same types of characteristics that belong to an identical property, are special values (or attributes) of the property. For example, ‘black’ and ‘white’ describe the colours of triangles in Figure 4.2.

Hence, ‘colour’ is a property of triangles, and ‘black’ and ‘white’ are attributes of ‘colour’. A collection of all possible attributes makes up the universe of a property, which is denoted as P^U . In a predefined context, we always suppose that we know the universe of the property in the given context. In Figure 4.2, the universe of colour is ‘black’ and ‘white’, and is represented as $P\left(colour\right)^U$ = {black, white}. It holds that

$$A \;\&\; P^U = A \tag{4.20}$$

where A is a concept, and P^U is the universe of a property that can be added to A.

For example, suppose A is $C0$ in Figure 4.2, then $A \;\&\; P\left(colour\right)^U$ = triangle& (black$\cup$white)=(triangle&black)$\cup$(triangle&white)={ $C1$, $C2$ }= A.

4.6 Conclusion

The novel model proposed in this chapter can be used in extensive forms of definitions, e.g., concepts and categories, objects and attributes, is-a based and part-of based, etc. Different types of concepts can be built by applying product and union operations on other concepts. The intents of the definer and the details of concepts can be reflected by the representation of the concepts. Two case studies demonstrate the feasibility and characteristics of the model. To acquire a unique quantification of a concept, two solutions have been proposed, namely, a global standard and a dynamic refinement process, and the latter is feasible from a practical point of view. Further studies should focus on applications of the model, e.g., semantic comparability and semantic interoperability.

CHAPTER 5 DYNAMIC SEMANTIC REFERENCE SYSTEM

5.1 Introduction

To analyse the definitions using a common basic foundation, a dynamic semantic reference system is set up in this chapter.

The objectives of this chapter are:

- To establish a Dynamic Semantic Reference System instead of a Semantic Reference System to present semantic information.
- To present a method used to establish the Dynamic Semantic Reference System.
- To build the Dynamic Semantic Reference System for the NLCD classification systems of the United States.

The remainder of this chapter is structured as follows. Section 5.2 describes a Dynamic Semantic Reference System and how such a system is built. Section 5.3 uses the NLCD classification systems as examples to construct a DSRS in a step-by-step manner. Section 5.4 presents a discussion on how to represent classes using the DSRS and illustrates that spatial reference systems and temporal reference systems are special cases of semantic reference systems. Finally, conclusions are summarized in Section 5.5.

5.2 Dynamic Semantic Reference System

5.2.1 Criteria for a Reference System

To develop a new reference system, it is necessary to identify the criteria to which a reference system should adhere. Generally, the following criteria should be applied for a reference system.

- A reference system is a theoretically and practically feasible system constructed based on scientific methods.
- A reference system is an organized structure applicable for representing all potential observations.
- A unique representation exists for each observation in the reference system.
- Different observations can be distinguished in the reference system.
- A reference system offers the ability to define measurements and to quantify observations.

The spatial reference system and temporal reference system are two most common reference systems. A spatial reference system is a local, regional, or global coordinate system used to represent the positions of geographical entities. A temporal reference system is defined for time measurements based on the rotation of the Earth. Two types of spatial reference system exist in the geospatial domain: geographic coordinate systems and projected coordinate systems. Many temporal reference systems are available, e.g., Universal Time (UT), Greenwich Sidereal Time (GST), Terrestrial Dynamical Time (TDT), International Atomic Time (TAI), Coordinated Universal Time (UTC), and GPS Time (GPST). Spatial and temporal reference systems allow dynamic representation of geographical entities and have the ability to transform these representations from one system to another without changing their meanings (Kuhn and Raubal, 2003).

5.2.2 Semantic Reference System

In addition to spatial and time information, thematic information is a third type of interesting information in the geospatial domain. Analogous to the spatial reference system, a semantic reference system is proposed to explain the meaning of thematic data (Kuhn, 2003). The proposed semantic reference system includes three components: semantic datum, semantic projection, and semantic transformation. A semantic datum contains the most basic concepts (Bian and Hu, 2007) that ground the meaning. A semantic projection reduces the representational complexity of concepts within a semantic reference system. A semantic transformation transforms concepts from one system to another system.

The semantic reference system is actually a top-level ontology (Agarwal, 2005). However, projects with top-level ontologies have failed (thus far, at least) (Smith and Mark, 2001) because no single reference system can serve all possible application needs (ISO, 2009).

5.2.3 Dynamic Semantic Reference System

"If all the universe were blue, there would be no blueness, since there would be nothing to contrast with blue. The same is true for the meanings of words. They have meaning only in terms of systematic contrasts with other words which share certain features with them but contrast with them in respect to other features." (Nida, 1975, p.31)

It is unnecessary to establish a universe semantic reference system (Kuhn, 2003), if the system can represent any concept in a predefined context and can be extended to describe other concepts if necessary. Symbols used to define a concept can be classified into two types: contrast and not-contrast. In this work, 'contrast' means the ability to be compared, and contrast components ought to describe characteristics of the same property. Hence, contrast components may be identical or different.

For example, 'black triangle' and 'white triangle' have two corresponding contrast components, i.e., 'black' and 'white', which are different attributes that describe the property of colour. 'Triangle' is an identical attribute in both concepts that describes the property of shape. For another concept pair of 'black triangle' and 'black scalene acute triangle', there is one contrast component and one not-contrast component. The 'black triangle' is an identical contrast component that describe properties of colour and shape, whereas 'scalene acute' in the latter concept is a not-contrast component. Although 'black' and 'triangle' can be separated as two identical contrast components, it is meaningless to add this refinement because it will increase the burdens of memory, storage, and processing. We state that ‘black triangle’ is a maximal contrast component for this concept pair.

A maximal contrast component of a concept is a part that all attributes are used to contrast with other concepts. It is meaningless to add a refinement to a maximal contrast component. An identical contrast component and a different contrast component are distinct for each other. Hence, ‘black’ and ‘triangle’ are two maximal contrast components of the concept ‘black triangle’ for the concept pair ‘black triangle’ and ‘white triangle’. Similarly, there exists a maximal not-contrast component of a concept for which any refinement is meaningless. Maximal contrast components and maximal not-contrast components are awareness of context. In a fixed context, there is only one maximal not-contrast component and one or more maximal contrast components in a concept. If new concepts are added into the context, refinement of contrast components and not-contrast components may occur. A refinement of contrast components results only in contrast components which own fewer attributes but offer additional insight into the concept. A refinement of not-contrast components results in contrast components that contrast with newly added concepts and not-contrast components that with the remainder of the attributes (other than the contrast components).

For example, the previously mentioned concept pair ‘black triangle’ and ‘black scalene acute triangle’ is decomposed into a contrast component (i.e., ‘black triangle’) and a not-contrast component (i.e., ‘scalene acute’) in the context of this concept pair. If the concept of ‘white triangle’ is added, the contrast component ‘black triangle’ is split into two contrast components, i.e., ‘black’ and ‘triangle’, in the context of these three concepts.

The amount of detail that should result from concept decomposition is determined by the context in which it exists. Although over refinements of concepts can produce satisfying results, it is useless to add these over refinements. The optimal decomposition reduces a concept into maximal contrast components and not-contrast components that can deliver satisfying results and a unique decomposition.

Based on the maximal contrast components and maximal not-contrast components, a Dynamic Semantic Reference System (DSRS) can be constructed. The representations of concepts in this system use the models introduced in Chapter 4 as a result of product and union operations. The DSRS fulfils all criteria listed in Section 5.2.1. Different from the Semantic Reference System, which engages in constructing a top-level ontology via a top-down method, a DSRS is constructed using a bottom-up method from the application-induced system to a top-level ontology if all concepts are added into the context.

A top-down method can be used to design a classification system initially. Predefined purposes are fully considered in the initial system. A corresponding specification can be released to be obeyed by all participants. In this way, a consistent result can be achieved. However, to analyse semantic uncertainty, a bottom-up method is better because the intended classes are not certain. All classes are potential classes which may be evaluated.

The main differences between the DSRS and SRS are listed in Table 5.1. Although a running example (Kuhn and Raubal, 2003) of a semantic reference system for navigation has been implemented based on the Haskell functional language standard (Hudak, 2000), it is not a top-level ontology.

The result of a DSRS is a collection of reference concepts built from maximal contrast components, maximal not-contrast components, and complement components. A complement component refers to the complement of all parts, i.e., contrast and not-contrast, that belong to the same property. The simplest way to produce a complement component is a negation of existing components. For example, ‘not black’ is the complement of ‘black’. The complement components are added because if an attribute is mentioned, this not only means that the attribute is interesting but also implies that the property to which it belongs is interesting. Complement components ensure a complete description of interesting concepts.

Table 5.1 A Comparison of DSRS and SRS

	DSRS	SRS
Unit	maximal contrast components and maximal not-contrast components proposed in this chapter	semantic primitives
Characteristic	dynamic, changes with concepts concerned	static, contains semantic primitives for all concepts
Construct Method	a bottom-up method	a top-down method
Level of Difficulty	Easy	hard
Feasibility	feasible theoretically and practically which will be demonstrated in this chapter	feasible theoretically but failed in practice (Smith and Mark, 2001)

For example, in the context of a concept 'triangle', a DSRS contains two reference concepts represented as (triangle) and (not triangle). In the context of the two concepts 'triangle' and 'black triangle', a DSRS contains four reference concepts: (black, triangle), (not black, triangle), (black, not triangle), and (not black, not triangle). In this manner, all concepts in the context can be represented as union of these reference concepts. To visualize the representation, reference concepts can be drawn using a tree structure in which the root is the universe of interesting concepts, and the leaves are the reference concepts. In this type of concept tree, leaf concepts (leaves in the tree) act as reference concepts to represent and distinguish all concepts in the context. Although internal nodes in the tree may be different, unique representations of leaf concepts can be acquired.

Figure 5.1 illustrates a concept tree in a context built from 'triangle' and 'black triangle'. The directed edge indicates a subordinate and superordinate relationship that stems from a subordinate concept to its superordinate concept. The label on an edge indicates the property used to produce a subordinate concept from its superordinate concept. In Figure 5.1, p1={triangle, not triangle} and p2={black, not black}.

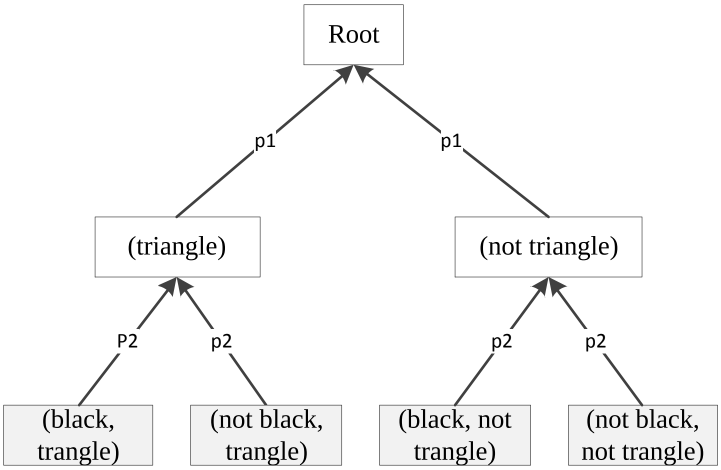


Figure 5.1 An Example of Concept Tree

A DSRS can be constructed with the following steps:

Table 5.2 An Algorithm to Construct a Dynamic Semantic Reference System

STEP 1	Add a concept and decompose its definition into maximal contrast components and maximal not-contrast components in the new context;
STEP 2	Replace a synonym by one consistent representation;
STEP 3	Refine the existing maximal contrast components and maximal not-contrast components in the new context;
STEP 4	Arrange the attributes in every property;
STEP 5	Analyse and split the overlaps among attributes in any property and rearrange attributes in the property;
STEP 6	Construct complement parts for every property if needed and rearrange attributes in the property;
STEP 7	Draw the concept tree;
STEP 8	The DSRS is represented by the collection of all leaf concepts.

5.3 Case Study: The NLCD Classification System as an Example

5.3.1 Experimental Data

The NLCD classification system (NLCD-CS) refers to both the NLCD1992 classification system (NLCD1992-CS) and the NLCD2001 classification system

(NLCD2001-CS), which have been employed in all national land inventory projects of the United States (i.e., NLCD1992, NLCD2001, NLCD2006, and NLCD2011) until now. The NLCD CS is modified from the Anderson Land Cover Classification System (Anderson et al., 1976). The NLCD2001 CS, with minor modification from the NLCD1992 CS, is applied to the NLCD2006 and NLCD2011 without any modification. Hence, two classification systems exist, i.e., the NLCD92 CS and the NLCD2001 CS, for four national land classification products. Both classification systems are two-level classification systems. There are 21 level-II classes and 9 level-I classes in the NLCD1992 CS, and there are 16 Level-II classes (not including 9 classes in coastal areas and another 4 classes in Alaska only) and 8 Level-I classes in the NLCD2001 CS. Compared with the NLCD1992 CS definitions, the classes water, forest, shrub, herbaceous, and wetland are nearly identical, and the classes agriculture, urban, and barren are slightly more adjusted in the NLCD2001 CS (Homer et al., 2004). Hence, land cover identified with the same terms from the NLCD2001 CS and NLCD1992 CS might describe different meanings, e.g., ‘Deciduous Forest’, which is appended as additional constraints in the NLCD2001 CS definition by ‘generally greater than 5 meters tall, and greater than 20% of total vegetation cover’. Complete definitions of the NLCD1992 CS and NLCD2001 CS are listed separately in Table 5.3 and Table 5.4.

Table 5.3 The NLCD1992 Classification System

L1	L2	Definition
10.		**Water**—All areas of open water or permanent ice/snow cover.
	11.	**Open Water**—Areas of open water, generally with less than 25% cover of vegetation/land cover.
	12.	**Perennial Ice/Snow**—All areas characterized by year-long surface cover of ice and/or snow.
20.		**Developed**—Areas characterized by a high percentage (30 percent or greater) of constructed materials (e.g. asphalt, concrete, buildings, etc.).
	21.	**Low Intensity Residential**—Includes areas with a mixture of constructed materials and vegetation. Constructed materials account for 30-80 percent of the cover. Vegetation may account for 20 to 70 percent of the cover. These areas most commonly include single-family housing units. Population densities will be lower than in high intensity residential areas.
	22.	**High Intensity Residential**—Includes highly developed areas where people reside in high numbers. Examples include apartment complexes and row houses. Vegetation accounts for less than 20 percent of the cover. Constructed materials account for 80 to100 percent of the cover.
	23.	**Commercial/Industrial/Transportation**—Includes infrastructure (e.g. roads,

railroads, etc.) and all highly developed areas not classified as High Intensity Residential.

30. **Barren**—Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with little or no "green" vegetation present regardless of its inherent ability to support life. Vegetation, if present, is more widely spaced and scrubby than that in the "green" vegetated categories; lichen cover may be extensive.
    31. **Bare Rock/Sand/Clay**—Perennially barren areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, beaches, and other accumulations of earthen material.
    32. **Quarries/Strip Mines/Gravel Pits**—Areas of extractive mining activities with significant surface expression.
    33. **Transitional**—Areas of sparse vegetative cover (less than 25 percent of cover) that are dynamically changing from one land cover to another, often because of land use activities. Examples include forest clearcuts, a transition phase between forest and agricultural land, the temporary clearing of vegetation, and changes due to natural causes (e.g. fire, flood, etc.)
40. **Forested Upland**—Areas characterized by tree cover (natural or semi-natural woody vegetation, generally greater than 6 meters tall); tree canopy accounts for 25-100 percent of the cover.
    41. **Deciduous Forest**—Areas dominated by trees where 75 percent or more of the tree species shed foliage simultaneously in response to seasonal change.
    42. **Evergreen Forest**—Areas dominated by trees where 75 percent or more of the tree species` maintain their leaves all year. Canopy is never without green foliage.
    43. **Mixed Forest**—Areas dominated by trees where neither deciduous nor evergreen species represent more than 75 percent of the cover present.
50. **Shrubland**—Areas characterized by natural or semi-natural woody vegetation with aerial stems, generally less than 6 meters tall, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.
    51. **Shrubland**—Areas dominated by shrubs; shrub canopy accounts for 25-100 percent of the cover. Shrub cover is generally greater than 25 percent when tree cover is less than 25 percent. Shrub cover may be less than 25 percent in cases when the cover of other life forms (e.g. herbaceous or tree) is less than 25 percent and shrubs cover exceeds the cover of the other life forms.
60. **Non-Natural Woody**—Areas dominated by non-natural woody vegetation; non-natural woody vegetative canopy accounts for 25-100 percent of the cover. The non-natural woody classification is subject to the availability of sufficient ancillary data to differentiate non-natural woody vegetation from natural woody vegetation.
    61. **Orchards/Vineyards/Other**—Orchards, vineyards, and other areas planted or maintained for the production of fruits, nuts, berries, or ornamentals.
70. **Herbaceous Upland**—Upland areas characterized by natural or semi-natural herbaceous vegetation; herbaceous vegetation accounts for 75-100 percent of the cover.
    71. **Grassland/Herbaceous**—Areas dominated by upland grasses and forbs. In rare cases, herbaceous cover is less than 25 percent, but exceeds the combined cover of the woody species present. These areas are not subject to intensive management, but they are often utilized for grazing.
80. **Planted/Cultivated**—Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation accounts for 75-100 percent of the cover.
    81. **Pasture/Hay**—Areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops.
    82. **Row Crops**—Areas used for the production of crops, such as corn, soybeans, vegetables, tobacco, and cotton.
    83. **Small Grains**—Areas used for the production of graminoid crops such as wheat,

barley, oats, and rice.

84. **Fallow**—Areas used for the production of crops that do not exhibit visable vegetation as a result of being tilled in a management practice that incorporates prescribed alternation between cropping and tillage.

85. **Urban/Recreational Grasses**—Vegetation (primarily grasses) planted in developed settings for recreation, erosion control, or aesthetic purposes. Examples include parks, lawns, golf courses, airport grasses, and industrial site grasses.

9. **Wetlands**—areas where the soil or substrate is periodically saturated with or covered with water

90. **Woody Wetlands**—Areas where forest or shrubland vegetation accounts for 25-100 percent of the cover and the soil or substrate is periodically saturated with or covered with water.

95. **Emergent Herbaceous Wetlands**—Areas where perennial herbaceous vegetation accounts for 75-100 percent of the cover and the soil or substrate is periodically saturated with or covered with water.

*cited from http://www.mrlc.gov/nlcd92_leg.php. (accessed on 3 July, 2014)

Table 5.4 The NLCD2001 Classification System* (The Conterminous United States)

L1	L2	Definition
10.		**Water**—All areas of open water or permanent ice/snow cover.
	11.	**Open Water**—All areas of open water, generally with less than 25 percent cover of vegetation or soil.
	12.	**Perennial Ice/Snow**—All areas characterized by a perennial cover of ice and/or snow, generally greater than 25 percent of total cover.
20.		**Developed**—Areas characterized by a high percentage (30 percent or greater) of constructed materials (e.g. asphalt, concrete, buildings, etc.).
	21.	**Developed, Open Space**—Includes areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20 percent of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes
	22.	**Developed, Low Intensity**—Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20–49 percent of total cover. These areas most commonly include single-family housing units.
	23.	**Developed, Medium Intensity**—Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50–79 percent of the total cover. These areas most commonly include single-family housing units.
	24.	**Developed, High Intensity**—Includes highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses, and commercial/industrial. Impervious surfaces account for 80 to 100 percent of the total cover.
30.		**Barren**—Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with little or no "green" vegetation present regardless of its inherent ability to support life. Vegetation, if present, is more widely spaced and scrubby than that in the "green" vegetated categories; lichen cover may be extensive.
	31.	**Barren Land(Rock/Sand/Clay)**—Barren areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits, and other accumulations of earthen material. Generally, vegetation accounts for less than 15 percent of total cover.
40.		**Forested Upland**—Areas characterized by tree cover (natural or semi-natural woody

vegetation, generally greater than 6 meters tall); tree canopy accounts for 25-100 percent of the cover.

41. **Deciduous Forest**—Areas dominated by trees generally greater than 5 meters tall, and greater than 20 percent of total vegetation cover. More than 75 percent of the tree species shed foliage simultaneously in response to seasonal change.
42. **Evergreen Forest**—Areas dominated by trees generally greater than 5 meters tall, and greater than 20 percent of total vegetation cover. More than 75 percent of the tree species maintain their leaves all year. Canopy is never without green foliage.
43. **Mixed Forest**—Areas dominated by trees generally greater than 5 meters tall, and greater than 20 percent of total vegetation cover. Neither deciduous nor evergreen species are greater than 75 percent of total tree cover.

50. **Shrubland**—Areas characterized by natural or semi-natural woody vegetation with aerial stems, generally less than 6 meters tall, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.
    52. **Shrub/Scrub**—Areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20 percent of total vegetation. This class includes true shrubs, young trees in an early successional stage, or trees stunted from environmental conditions.
70. **Herbaceous Upland**—Upland areas characterized by natural or semi-natural herbaceous vegetation; herbaceous vegetation accounts for 75-100 percent of the cover.
    71. **Grassland/Herbaceous**—Areas dominated by grammanoid or herbaceous vegetation, generally greater than 80 percent of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.
80. **Planted/Cultivated**—Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation accounts for 75-100 percent of the cover.
    81. **Pasture/Hay**—Areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20 percent of total vegetation.
    82. **Cultivated Crops**—Areas used for the production of annual crops, such as corn, soybeans, vegetables, tobacco, and cotton, and also perennial woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20 percent of total vegetation. This class also includes all land being actively tilled.
9. **Wetlands**—areas where the soil or substrate is periodically saturated with or covered with water
    90. **Woody Wetlands**—Areas where forest or shrubland vegetation accounts for greater than 20 percent of vegetative cover and the soil or substrate is periodically saturated with or covered with water.
    95. **Emergent Herbaceous Wetlands**—Areas where perennial herbaceous vegetation accounts for greater than 80 percent of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

*cited from http://www.epa.gov/mrlc/definitions.html with '95. Emergent Herbaceous Wetlands' replaced referencing Homer et al. (2004) and http://www.mrlc.gov/nlcd01_leg.php and Level I class 'Wetlands' is added according to http://www.mrlc.gov/nlcd92_leg.php. (websites accessed on 3 July, 2014)

The DSRS of the NLCD CS will be established in a step-by-step manner following the algorithm presented in Section 5.2.3 by adding the classes one by one. An explanation for how the categories are represented will be covered in

Section 5.4.2. For reasons of space and duplication, only the first three classes in Table 5.3 are used for illustration of the complete process.

5.3.2 The Initial Class: ‘Water’

The first class added into the context is ‘Water’ with code ‘10’. Because there is no contrast for only one class, only one maximal not-contrast component exists. The class ‘Water’ can be represented as (open water or permanent ice/snow). The complement of this not-contrast component can be built by negating it as (not open water or permanent ice/snow). A concept tree is drawn as shown in Figure 5.2.

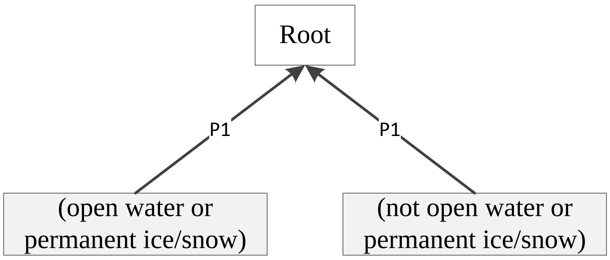


Figure 5.2 The Concept Tree in the Context of ‘Water’

The complete process following Table 5.2 is described as following:

STEP1.	(open water or permanent ice/snow)
STEP2.	not applicable
STEP3.	not applicable
STEP4.	P1={open water or permanent ice/snow}
STEP5.	not applicable
STEP6.	P1={open water or permanent ice/snow, not open water or permanent ice/snow}
STEP7.	as shown in Figure 5.2
STEP8.	two reference concepts: (open water or permanent ice/snow) and (not open water or permanent ice/snow)

Hence, the DSRS in the context of ‘water’ is constructed by the two reference concepts (open water or permanent ice/snow) and (not open water or permanent ice/snow).

5.3.3 The Second Class: ‘Open Water’

The ‘open water’ in the class of ‘Open Water’ is an identical contrast component with ‘open water’ in the class of ‘Water’. Therefore, ‘open water’ is decomposed as a maximal contrast component, and ‘vegetation/land’ is decomposed as a maximal not-contrast component in the first step. In the third step, the reference concepts in Figure 5.2 are split into four parts for the emergence of ‘open water’ in the first step. The concept tree is illustrated in Figure 5.3, and the process steps are described as follows.

STEP1.	(open water), (vegetation/land)
STEP2.	not applicable
STEP3.	(open water), (permanent ice/snow), (not open water), (not permanent ice/snow)
STEP4.	P1={open water or permanent ice/snow, not open water, not permanent ice/snow, vegetation/land} P2={open water, permanent ice/snow}
STEP5.	P1={open water or permanent ice/snow, vegetation/land, other}
STEP6.	not applicable
STEP7.	as shown in Figure 5.3
STEP8.	four reference concepts: (open water), (permanent ice/snow), (vegetation/land) and (other)

In step 4, overlap occurs among ‘not open water’, ‘not permanent ice/snow’ and ‘vegetation/land’; therefore, the property is split and rearranged in step 5. In this situation, the DSRS is constructed using four reference concepts: (open water), (permanent ice/snow), (vegetation/land) and (other).

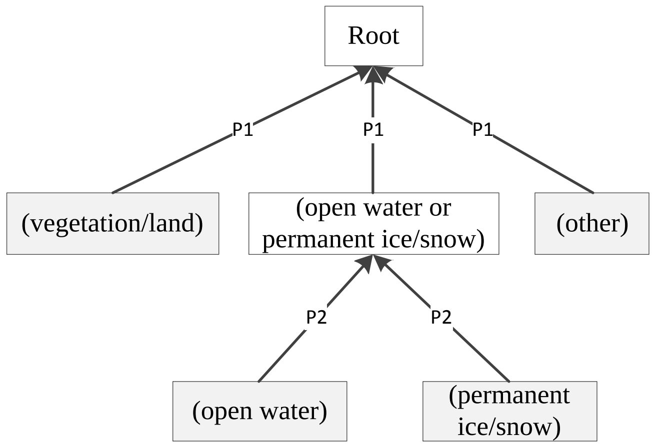


Figure 5.3 The Concept Tree in the Context of ‘Water’ and ‘Open Water’

5.3.4 The Third Class: ‘Perennial Ice/Snow’

The ‘year-long ice and/or snow’ in the class ‘Perennial Ice/Snow’ is replaced by its synonym of ‘permanent ice/snow’ in step 2. The complete process is described as follows.

STEP1. (year-long ice and/or snow)

STEP2. replace by (permanent ice/snow)

STEP3. not applicable

STEP4. not applicable

STEP5. not applicable

STEP6. not applicable

STEP7. as shown in Figure 5.3

STEP8. four reference concepts: (open water), (permanent ice/snow), (vegetation/land) and (other)

The complete process used to build a DSRS has been demonstrated in the first three classes. The DSRS may remain unchanged if a new concept is added into the context, e.g., adding the class ’Perennial Ice/Snow’ into the context of ‘Water’ and ‘Open Water’.

5.3.5 DSRS for the NLCD-CS

The final concept tree for the NLCD-CS is illustrated in Figure 5.4. In the DSRS for the NLCD-CS, there are 21 reference concepts, which are listed in Table 5.5. In other words, all concepts mentioned in the NLCD1992 CS and the NLCD2001 CS can be represented by these reference concepts. These reference concepts are produced by 13 properties (Table 5.6). Every reference concept is a J-concept (as defined in Chapter 4). The more product operations are applied, the more specific the reference concept will be. For land classification, every reference concept is a type of covering material. It can be observed from Figure 5.4 that vegetation is the most refined covering material compared with the constructed material, water and soil. There are 15 reference concepts under vegetation that are more than twice the sum of the remaining three. This observation reflects that in the NLCD project, vegetation is the most prominent and interesting type of coverage.

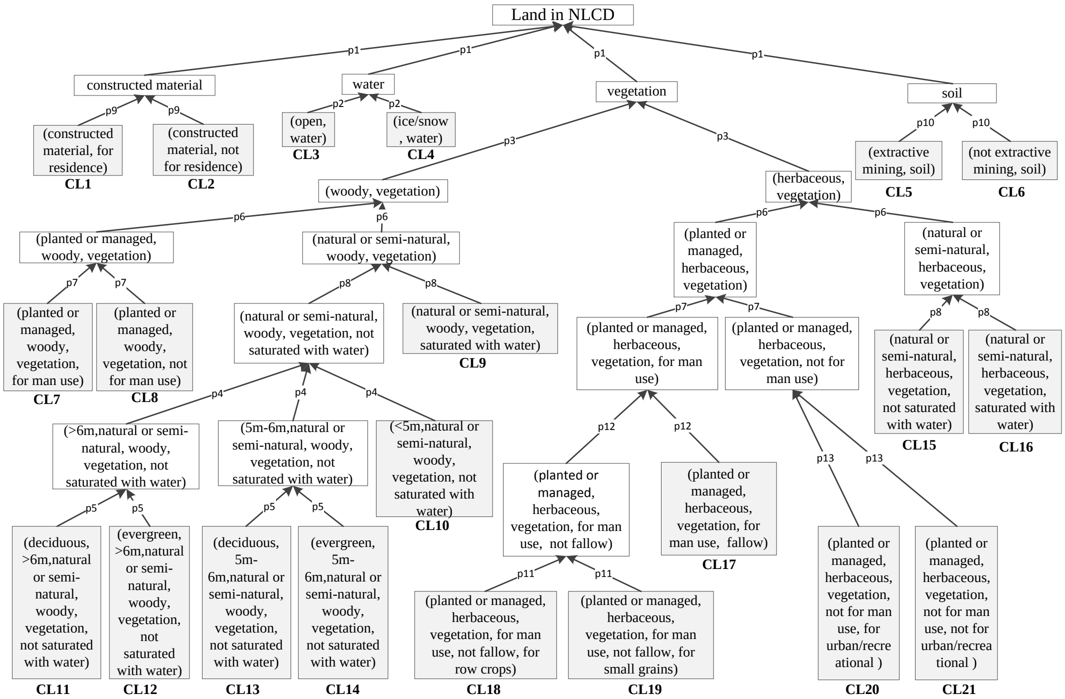


Figure 5.4 The Concept Tree in the Context of NLCD-CS

Table 5.5 Reference Concepts for the DSRS of the NLCD-CS

CL1	(constructed material, for residence)
CL2	(constructed material, not for residence)
CL3	(open, water)
CL4	(ice/snow, water)
CL5	(extractive mining, soil)
CL6	(not extractive mining, soil)
CL7	(planted or managed, woody, vegetation, for man use)
CL8	(planted or managed, woody, vegetation, not for man use)
CL9	(natural or semi-natural, woody, vegetation, saturated with water)
CL10	(<5m,natural or semi-natural, woody, vegetation, not saturated with water)
CL11	(deciduous, >6m,natural or semi-natural, woody, vegetation, not saturated with water)
CL12	(evergreen, >6m,natural or semi-natural, woody, vegetation, not saturated with water)
CL13	(deciduous, 5m-6m,natural or semi-natural, woody, vegetation, not

	saturated with water)
CL14	(evergreen, 5m-6m,natural or semi-natural, woody, vegetation, not saturated with water)
CL15	(natural or semi-natural, herbaceous, vegetation, not saturated with water)
CL16	(natural or semi-natural, herbaceous, vegetation, saturated with water)
CL17	(planted or managed, herbaceous, vegetation, for man use, fallow)
CL18	(planted or managed, herbaceous, vegetation, for man use, not fallow, for row crops)
CL19	(planted or managed, herbaceous, vegetation, for man use, not fallow, for small grains)
CL20	(planted or managed, herbaceous, vegetation, not for man use, for urban/recreational)
CL21	(planted or managed, herbaceous, vegetation, not for man use, not for urban/recreational)

Table 5.6 Properties Employed to Construct DSRS of NLCD-CS

P1.	{water, vegetation, soil, constructed material}
P2.	{open, ice/snow}
P3.	{woody, herbaceous}
P4.	{>6m, 5m-6m, <5m}
P5.	{deciduous, evergreen}
P6.	{natural or semi-natural, planted}
P7.	{not for man use, for man use}
P8.	{saturated with water, not saturated with water}
P9.	{for residence, not for residence}
P10.	{extractive mining, not extractive mining}
P11.	{for row crops, for small grains}
P12.	{fallow, not fallow}
P13.	{for urban/recreational, not for urban/recreational}

5.4 Discussion

Definitions can be divided into two types: concept and category. In land classification, a concept determines the covering on the land and a category determines how much is covered by a concept. In this section, we primarily discuss how to represent the concept and category using the DSRS.

5.4.1 Representation of Concepts by the DSRS

All concepts in the NLCD CS can be represented by applying union operations to reference concepts. For example, the class of 'Water' (code: 10) is the union of CL3 and CL4 in Table 5.5. Formally, we state that

$$C = \bigcup CL_i \tag{5.1}$$

where C is a concept mentioned in the NLCD CS, $\bigcup$ is the union operation, CL_i is a reference concept, and i is employed to indicate all reference concepts that belong to C.

Therefore, the 'Water' (code: 10) can be represented by $CL3 \cup CL4 =$ {CL3, CL4}. In the concept tree, any node can be represented as the union of its descendant leaf nodes. Although not all concepts are depicted as nodes in the concept tree, they also can be represented by the DSRS. For example, although 'natural or semi-natural woody vegetation, less than 6 meters tall' in the class of 'Shrubland' (code: 50) is not depicted in Figure 5.4, it can be represented as $CL10 \cup CL13 \cup CL14 =$ {CL10, CL13, CL14}.

5.4.2 Representation of Categories by the DSRS

The main incomplete symbol in a land classification system describes the percentage of area coverage of a covering material. Hence, for every parcel, it holds that

$$\sum AP\left(CL_i\right) = 1 \tag{5.2}$$

where $AP(\ )$ is the area percentage of a covering material in a parcel, and i indicates all covering materials of the parcel.

In a definition, we normally describe the main characteristics and neglect non-essential properties, which is also suitable for the land class definition. Land parcels are more or less a mixture of different covering materials instead of a single type of land coverage. It is not possible and also not necessary to list all coverings of the class. Hence, every class can be represented as an equation set that includes equation (5.2). For example, 'Perennial Ice/Snow' (code: 12) in the NLCD2001-CS can be represented by an equation set $\begin{cases} AP(\mathrm{CL}_4) \geq 25\% \\ \sum AP(\mathrm{CL}_i) = 1 \end{cases}$. Only parcels that fulfil this equation set belong to this class.

5.4.3 Artificial Zone and Natural Zone

The priority for every covering material in the mixture is not equal. The main divergence is caused by a geographical division between artificial zones and natural zones, which constitute a partition of the land. In this work, an artificial zone is different from an urban area or residential area. A natural area can exist in an urban area, e.g., hills in an urban area. In the context of the NLCD CS, 9 reference concepts can be considered as located in artificial zone, i.e., CL1, CL2, CL7, CL8, CL17, CL18, CL19, CL20, and CL21. The remainder of the 12 reference concepts belong to the natural zone. The main reason for making a division between the artificial zone and natural zone is to reduce meaningless mixtures of covering materials. For example, it is rare that a parcel is a mixture of CL1 and CL4.

5.4.4 Semantic, Spatial, and Temporal Reference Systems

The spatial reference system, which describes the position of an entity, is a special case of the semantic reference system (Kuhn, 2003; Kuhn and Raubal, 2003). Every axis is a property used to construct the DSRS. Together with more interesting positions added into the context, the DSRS is approximated with the

SRS. For example, assume we are interested in the positions of cars on a road, which can be represented in a one-dimensional spatial reference system but is to be established by the method proposed in this chapter.

We establish this one-dimensional system by recording a report from the driver who tells us how far s/he drives from a toll station named as the start point according to the reference of the odometer on the car. If the driver reports once every 10 miles, we can establish a DSRS represented by $\{x_i\}$, where x_i is a reference concept and $x_i = 10n, n \in N^+$. If the driver reports once every 5 miles, we can establish a DSRS represented by $\{x_i\}$, where $x_i = 5n, n \in N^+$. The more reports are recorded, the more precise the system will be. In addition to the equal spacing system, we can also establish an unequal spacing system by the DSRS, e.g., {2,5,6,8…}, although it is rare in practice. A two-dimensional system can be established in the same way and can be represented by $\{(x_i, y_i)\}$.

Similarly, the temporal system is a special case of the semantic reference system.

5.5 Conclusion

Instead of creating a top-level Semantic Reference System, this chapter proposed the establishment of a Dynamic Semantic Reference System, which has been demonstrated as theoretically and practically feasible by the case study on the NLCD classification systems. The DSRS method can also be used to establish spatial reference systems and temporal reference systems. The emergence of the DSRS represents vital progress in quantifying semantic information rather than qualifying it. The DSRS is the foundation for semantic applications, e.g., semantic overlap, semantic similarity, and semantic interoperability. The engagement of our studies will focus on these applications based on the DSRS.

CHAPTER 6 SEMANTIC OVERLAP

6.1 Introduction

Semantic overlaps exist between classes within a classification system and between different classification systems. Many examples have been pointed out (Herold et al., 2006b; Di Gregorio, 2005, p.8; Wyatt et al., 1997, p.46). For example, there are overlapping definitions of crown cover parameters and modifiers in LCCS (Jansen et al., 2008). Attribute overlaps are usually employed to calculate semantic similarity (Ahlqvist, 2004; Ahlqvist and Gahegan, 2005), especially in feature-based methods (Tversky, 1977). This chapter proposes a method to quantify semantic overlaps.

The objectives of this chapter are:

- To study semantic overlaps between classes other than attributes;
- To propose a quantitative measurement of semantic overlaps and to demonstrate the measurement using NLCD2001 CS classes;
- To calculate and compare semantic overlaps in different situations, including (1) every class is treated as an independent and complete definition; (2) level-II class definitions are constrained by level-I class definitions; and (3) priorities of classes are considered in classification.

The remainder of this chapter is structured as follows. Section 6.2 proposes a measurement to quantify semantic overlap. Section 6.3 takes NLCD2001 CS classes as examples to evaluate performance of the proposed measurement. Section 6.4 discusses the results, and Section 6.5 makes conclusions.

6.2 Modelling Semantic Overlap

6.2.1 Types of Semantic Overlap

There are semantic overlaps between two definitions if an entity can be defined by either of them. Formally, we say that definitions C_1 and C_2 are overlapped, if it holds that $\exists o, o \in C_1$ and $o \in C_2$, where o is an entity.

There are three types of semantic overlap: identical, contained, and intersected. If $\forall o \in C_1, o \in C_2$ *and* $\forall o \in C_2, o \in C_1$, then C_1 and C_2 are identical and denoted as $C_1 = C_2$. If two definitions are identical, any entity of C_1 is an entity of C_2, and vice versa. If $\forall o \in C_1, o \in C_2$ *and* $\exists o \in C_2, o \notin C_1$, then C_1 is contained by C_2 and denoted as $C_1 \subset C_2$ or $C_2 \supset C_1$. If definition C_1 is contained by definition C_2, any entity of C_1 is an entity of C_2, but there exist entities in C_2 that do not belong to C_1. If $\exists o \in C_1, o \in C_2$ and $\exists o \in C_1, \mathrm{o} \notin \mathrm{C}_2$ and $\exists o \in C_2, o \notin C_1$, we say that C_1 and C_2 are intersected. The type of semantic overlap is drawn in Figure 6.1.

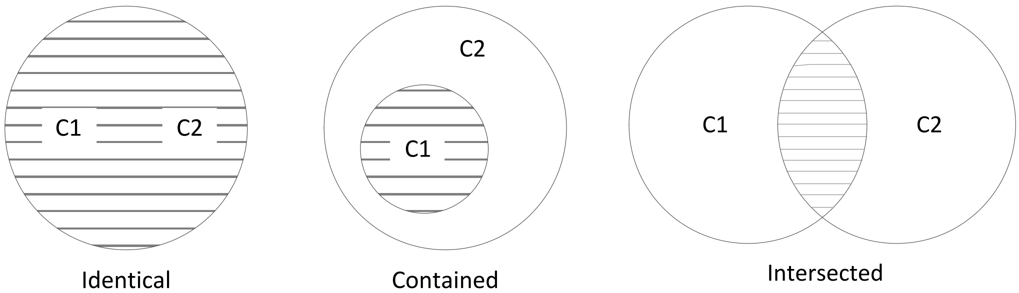


Figure 6.1 Three Semantic Overlap Types

The type of semantic overlap between synonyms is identical, such as the name of a class and the code of a class. The type of semantic overlap between a subordinate class and its superordinate class should be contained. In a classification system, every subordinate class should be contained by its superordinate class, and classes in a hierarchical level should not be overlapped with each other. However, there is no indicator to measure the semantic overlap yet. In this chapter, a measurement is proposed to fill this gap.

6.2.2 Measurement of Semantic Overlap

The proposed measurement of semantic overlap is based on the Dynamic Semantic Reference System (DSRS, Chapter 5). The DSRS is a partition of the semantic domain of interest. Definitions can be decomposed into reference concepts in the DSRS. The semantic overlap between two definitions can be measured by equation (6.1).

$$O\left(C_1, C_2\right) = \frac{\left|C_1 \cap C_2\right|^L}{\left|C_1\right|^L} \qquad (6.1)$$

where C_1 and C_2 are definitions which can be concepts or categories, $O\left(C_1, C_2\right)$ indicates the semantic overlap between definition C_1 and C_2, $\cap$ is the intersection operation, $C_1 \cap C_2$ results in common reference concepts or common combinations between C_1 and C_2, and $\left| \; \right|^L$ returns the number of reference concepts or the number of combinations.

For example, based on the DSRS of the triangle illustrated in Figure 4.2, the semantic overlap between $C21$ ('white scalene triangle') and $C2$ ('white triangle') can be measured as follows. $C21$ is decomposed into the union of $C211$ and $C212$, and $C2$ is decomposed into the union of $C211$, $C212$, $C221$, and $C222$. The intersection of $C21$ and $C2$ results in $C211$ and $C212$. Hence, $\left|C21 \cap C2\right|^L = 2$ and $\left|C21\right|^L = 2$. The semantic overlap between $C21$ and $C2$ is 1.

6.2.3 Characteristics of the Measurement

The range of the measurement is $\left[0, 1\right]$. 0 indicates that there is no semantic overlap between C_1 and C_2. In other words, entities belonging to these two definitions can be clearly separated from each other. 1 indicates that definition C_1 is identical or subordinated to definition C_2. The larger the number is, the more C_1 is semantically overlapped with the latter definition. An interpretation of the measurement is the percentage that C_1 is overlapped by C_2. For example, a measurement of 0.5 means that half of C_1 is overlapped by C_2.

Another main characteristic of the measurement is asymmetry. The overlap $O\left(C_1, C_2\right)$ may be different from $O\left(C_2, C_1\right)$. The measurement of definitions without overlap is always equal to 0. The measurement for identical overlap is always equal to 1. The measurement of overlap between a subordinate definition and a superordinate definition is always equal to 1, while the opposite is always smaller than 1. The measurement of an intersected overlap $O\left(C_1, C_2\right)$ is equal to $O\left(C_2, C_1\right)$, if and only if the number of reference concepts contained by both definitions is identical; otherwise, they are different.

6.3 Case Study and Analysis

6.3.1 Experimental Data

We consider the NLCD2001 classification system (NLCD2001-CS) as an example to evaluate the performance of the proposed semantic overlap measurement. There are many reasons to choose this system as an example. Firstly, this hierarchical classification system is the most popular structure in land classification systems. Secondly, it has been employed by three national land inventory projects in the United States: NLCD2001, NLCD2006, and NLCD2011. Complete definitions of the classes in NLCD2001-CS are described in Table 5.4. The DSRS of NLCD2001-CS is employed from Chapter 5.

Because the covering materials (reference concepts) of the artificial zone and the natural zone are rarely adjacent with each other, the classes in the artificial zone and natural zone are separately represented. There are 2 level-I classes and 6 level-II classes located in the artificial zone, including classes coded with 20, 80, 21, 22, 23, 24, 81, and 82. The other 6 level-I classes and 10 level-II classes are located in the natural zone. Classes in the artificial zone are represented by 9 covering materials: CL1, CL2, CL7, CL8, CL17, CL18, CL19, CL20, and CL21, as shown in Table 5.5. The other 12 covering materials are applied to represent classes in the natural zone.

A class is defined as a concept or category and is always characterized by some covering materials. The measurement of the semantic overlap is only evaluated among concepts or categories. The semantic overlap between a concept and a category is not assessed because it is meaningless.

Covering materials and category representation for corresponding classes are listed in Table 6.1 for classes in the artificial zone and Table 6.2 for classes in the natural zone. Aside from the equations listed in the two tables, a common equation, $\sum AP(\mathrm{CL}_i) = 1$, should be fulfilled by each category. For example, the class 'Developed' with code '20' is a category that can be represented by the following equation set: $\begin{cases} AP\left(CL_1 + CL_2\right) \geq 30\% \\ \sum AP(\mathrm{CL}_i) = 1 \end{cases}$.

Table 6.1 NLCD 2001 Classes in Artificial Zone

Code	Characteristic Covering	Category
20	CL1, CL2	AP(CL1+CL2)>=30%
21	CL1, CL2	0<AP(CL1+CL2)<20%
22	CL1, CL2	20%<=AP(CL1+CL2)<50%
23	CL1, CL2	50%<=AP(CL1+CL2)<80%
24	CL1, CL2	AP(CL1+CL2)>=80%
80	CL17, CL18, CL19, CL20, CL21	AP(CL17+CL18+CL19+CL20+CL21)>= 75%
81	CL21	AP(CL21)/AP(CL7+CL8+CL17+CL18+ CL19+ CL20+CL21)>20%

82	CL7, CL17, CL18, CL19	AP(CL7+CL17+CL18+CL19)/AP(CL7+CL8+CL17+CL18+CL19+ CL20+CL21)>20%

Table 6.2 NLCD 2001 Classes in Natural Zone

Code	Characteristic Covering	Category
10	CL3, CL4	
11	CL3	AP(CL3)>=75%
12	CL4	AP(CL4)>25%
30	CL5,CL6	
31	CL5,CL6	AP(CL5+CL6)>=85%
40	CL11,CL12	AP(CL11+CL12)>=25%
41	CL11, CL13	(1)AP(CL11+CL12+CL13+CL14)/AP(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL16)>20%, and (2)AP(CL11+CL13)/AP(CL11+CL12+CL13+CL14)>75%
42	CL12, CL14	(1)AP(CL11+CL12+CL13+CL14)/AP(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL16)>20%, and (2)AP(CL12+CL14)/AP(CL11+CL12+CL13+CL14)>75%
43	CL11, CL12,CL13, CL14	(1)AP(CL11+CL12+CL13+CL14)/AP(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL16)>20% ,and (2)AP(CL11+CL13)/AP(CL11+CL12+CL13+CL14)<=75%, and (3)AP(CL12+CL14)/AP(CL11+CL12+CL13+CL14)<=75%
50	CL10, CL13, CL14	
52	CL10	AP(CL10)/AP(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL16)>20%
70	CL15	AP(CL15)>=75%
71	CL15	AP(CL15)/AP(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL16)>80%
9	CL9, CL16	
91	CL9	AP(CL9)/AP(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL16)>20%
92	CL16	AP(CL16)/AP(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL16)>80%

Concepts with the same name are represented differently in the artificial zone and in the natural zone. For example, vegetation is represented as {CL7, CL8, CL17, CL18, CL19, CL20, CL21} in the artificial zone and {CL9, CL10, CL11, CL12, CL13, CL14, CL15, CL16} in the natural zone. For rare classes, categories are not represented by the area percentage of characteristic covering materials, such as the class ‘Open Water’ with code ‘11’ in NLCD2001 CS. In this case, the area percentage of characteristic covering materials is inferred from $\sum AP(\mathrm{CL}_i) = 1$.

The measurement of semantic overlap is performed based on three different points of view. Firstly, characteristic covering materials are vital to a class. To some extent, a class is defined by the existence of characteristic covering materials. The semantic overlap between characteristic covering materials, which is deemed to be a kind of semantic overlap between classes, is calculated in section 6.3.2. Secondly, every definition should define a class completely and independently. Hence, a class should be represented by its definition and semantic overlaps are calculated in section 6.3.3. Thirdly, all information of a superordinate class should be inherited by its descendants. Semantic overlaps between classes constrained by superordinate classes are calculated in section 6.3.4.

6.3.2 Characteristic Covering Concept Overlap

6.3.2.1 Artificial Zone

The results of the characteristic covering concept overlap of classes in the artificial zone are presented in Table 6.3. Semantic overlaps between class ‘20’ and class ‘80’, including their subordinate classes, are all equal to 0, which means that class ‘20’ can be clearly separated from class ‘80’. Semantic overlaps between classes including ‘20’ and its subordinate classes are all equal to 1, because they are characterized by the same reference concepts: CL1 and CL2. Class ‘81’ is contained by its superordinate class ‘80’, while only 75% of ‘82’ is intersected by ‘80’. The reason is that CL7 is a kind of characteristic covering material of ‘82’, but not of ‘80’. There is no overlap between ‘81’ and ‘82’.

Measurements along the diagonal are self-overlaps, which should always be identical to 1.

Table 6.3 Characteristic Covering Concept Overlap in Artificial Zone

	20	21	22	23	24	80	81	82
20	1	1	1	1	1	0	0	0
21	1	1	1	1	1	0	0	0
22	1	1	1	1	1	0	0	0
23	1	1	1	1	1	0	0	0
24	1	1	1	1	1	0	0	0
80	0	0	0	0	0	1	0.2	0.6
81	0	0	0	0	0	1	1	0
82	0	0	0	0	0	0.75	0	1

6.3.2.2 Natural Zone

Table 6.4 Characteristic Covering Concept Overlap in Natural Zone

	10	11	12	30	31	40	41	42	43	50	52	70	71	9	91	92
10	1	0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	1	0.5	0.5	1	0	0	0	0	0	0	0
41	0	0	0	0	0	0.5	1	0	1	0.5	0	0	0	0	0	0
42	0	0	0	0	0	0.5	0	1	1	0.5	0	0	0	0	0	0
43	0	0	0	0	0	0.5	0.5	0.5	1	0.5	0	0	0	0	0	0
50	0	0	0	0	0	0	0.33	0.33	0.67	1	0.33	0	0	0	0	0
52	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.5	0.5
91	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

The results of characteristic covering concept overlap of classes in the natural zone are presented in Table 6.4. There is no overlap between one level-I class and another level-I class (including their subordinate classes) except for between class ‘40’ (including its subordinate classes) and class ‘50’. There is no overlap between ‘40’ and ‘50’. Overlaps exist between ‘41’, ‘42’, ‘43’ and ‘50’ because ‘CL13’ and ‘CL14’ are characteristic covering materials for them.

All characteristic covering concepts of subordinate classes are contained by that of superordinate classes except for subordinate classes of ‘40’. Only half of ‘41’, ‘42’ and ‘43’ are contained by ‘40’, because ‘CL13’ and ‘CL14’ are not characteristic covering materials of ‘40’. Moreover, it is strange that ‘40’, ‘41’, and ‘42’ are contained by ‘43’.

The results of characteristic covering concept overlap reveal that most classes are clearly described by different characteristic coverings. This is because characteristic coverings are mainly judgement to separate the class from others. And what emerges in mind when a class is mentioned is also its characteristic covering materials.

6.3.3 Original Category Overlap

6.3.3.1 Class Distribution

An equation set representing a category is solved by computer-assisted numerical simulation in this research. The covering percentage is discretized into 20 pieces. The interval for every piece is 5% as illustrated in Figure 6.2. A category is then represented as a mixture of different covering materials. For example, ((30% CL2), (70% CL21)) belongs to class ‘20’.

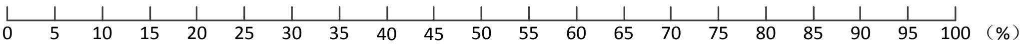


Figure 6.2 Discretization of Area Percentage

In total, there are 3,108,105 combinations of reference concepts in the artificial zone and 84,672,315 combinations in the natural zone. The number of combinations of a class divided by the total number of combinations produces the percentage of the class out of the total. The distribution of each class is described in Table 6.5 for classes in the artificial zone and in Table 6.6 for classes in the natural zone. Because categories are represented by definitions independently, the percentage of a subordinate class can exceed the percentage of its superordinate class.

Table 6.5 Class Distribution in Artificial Zone

Class Code	20	21	22	23	24	80	81	82
Percentage (%)	32.74	37.38	47.55	7.48	0.19	20.13	27.19	95.00

Table 6.5 indicates that the most dominated class is ‘82’ while the least dominated class is ‘24’. Based on their definitions, 95.00% of land belongs to ‘82’ while only 0.19% belongs to ‘24’ from a semantic point of view. A semantic point of view means that the property of any combination is equally treated without considering weights affected by practical situations.

The class ‘20’ and its subordinate classes own the same characteristic covering materials and are defined by explicitly declaring the percentage of these characteristic coverings. The percentage is updated by its subordinate classes. However, the range of covering percentage of subordinate classes is not entirely covered by ‘20’. For example, ‘less than 20 percent’ in the definition of ‘21’ is beyond the scope of ‘20’. This is the reason that covering percentages of ‘21’ and ‘22’ are larger than that of ‘20’. For ‘80’ and its subordinate classes ‘81’ and ‘82’, the percentage of covering materials is defined in contrast to different references. The former is relative to the total, while the latter is relative to ‘vegetation’. Hence, the percentage of ‘81’ and ‘82’ can exceed the percentage of ‘80’.

Table 6.6 Class Distribution in Natural Zone

Class Code	10	11	12	30	31	40	41	42
Percentage (%)	-	0.005	5.264	-	0.006	28.427	18.427	18.427
Class Code	43	50	52	70	71	9	91	92
Percentage (%)	53.748	-	22.875	0.005	0.066	-	22.875	0.066

In Table 6.6, '-' indicates that the class is defined as a concept rather than a category. The most dominated class is '43' while the least dominated class is '11' and '70' in the natural zone. Based on their definitions, 53.748% of land is dominated by '43', while only 0.005% of land is dominated by '11' and '70' from a semantic point of view.

Among these classes, only '40' and '70' in level-I are defined as categories, and all level-II classes are categories. The reason that percentages of subordinate classes exceed superordinate classes is that different references are applied in their definitions, which is the same as class '80' and its subordinate classes.

6.3.3.2 Artificial Zone

Measurements of semantic overlap in Table 6.7 are different from those in Table 6.3. Firstly, more overlaps are reflected in Table 6.7. Secondly, subtle differences are depicted in Table 6.7.

There is no overlap between level-II classes of '20', because the same covering materials and areal reference are employed in their definitions. The range of area percentage of the same covering materials is partitioned by these four level-II classes. Classes '20' and '21' do not overlap with each other because the area percentages are not intersected. There is no overlap between '80' and '20', '23', and '24' because of the conflict of area percentages which cannot fulfil $\sum AP(\mathrm{CL}_i) = 1$.

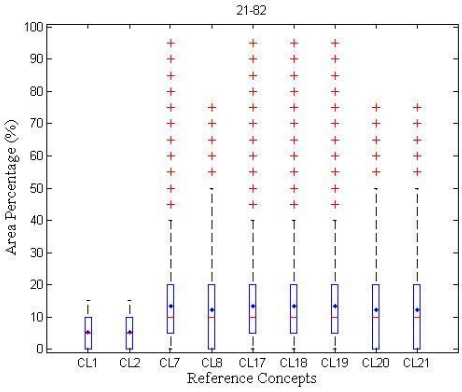
Overlaps more or less exist between other class pairs. Surprisingly, there is a high percentage of overlap (around 90%) between each class and '82' (measurements underlined in Table 6.7). Elements of overlaps between level-II classes are analysed by boxplots as shown in Figure 6.3. The dot in a box is the mean area percentage of a reference concept. The sum of means in every sub-figure is equal to 1.

The distribution can be classified into three types according to the area percentage of each reference concept as closed (more than 60-70 percent), open (70-60 percent to 20-10 percent), and sparse (20-10 percent to 1 percent) (Di Gregorio 2005, p.104). For analysis, a hard division is made: closed (more than 70 percent), open (70 percent to 20 percent), and sparse (less than 20 percent) in this research.

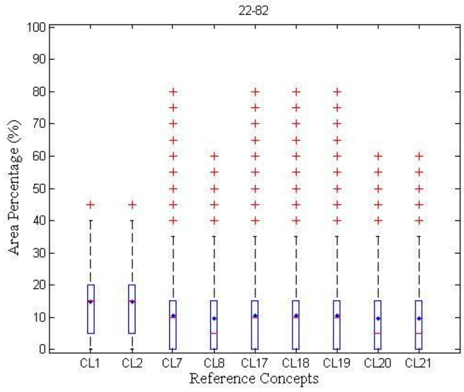
Table 6.7 Category Overlap in Artificial Zone

	20	21	22	23	24	80	81	82
20	1	0	0.7658	0.2285	0.0057	0	0.2853	0.9433
21	0	1	0	0	0	0.3594	0.2869	*0.9612*
22	0.5273	0	1	0	0	0.0584	0.2658	*0.9462*
23	1	0	0	1	0	0	0.2686	*0.9208*
24	1	0	0	0	1	0	0.3610	*0.8909*
80	0	0.6676	0.1379	0	0	1	0.3635	0.9580
81	0.3436	0.3945	0.4649	0.0739	0.0025	0.2691	1	*0.8979*
82	0.3250	0.3782	*0.4736*	0.0725	0.0017	0.2029	0.2569	1

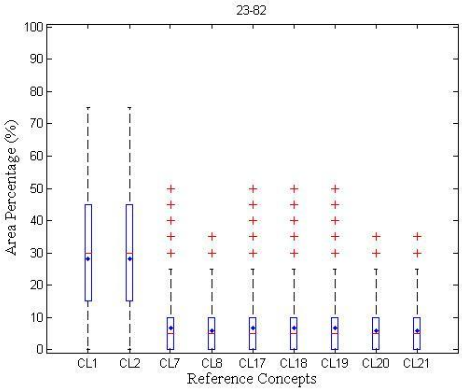
The overlap between ‘21’ and ‘82’ is illustrated in Figure 6.3 (a). All medians and means are smaller than 15%, which shows that most overlaps are sparse. The mean is larger than the median for every reference concept, which indicates that more values are smaller than the mean. The medians of CL1 and CL2 are smaller than the medians of other reference concepts. The largest interquartile range occurs on CL8, CL20, and CL21, while the smallest interquartile range occurs on CL1 and CL2, which indicates that values of the former are disperse than the latter. Outliers exist except for CL1 and CL2. The maximum percentages of CL7, CL17, CL18, and CL19 approach 95%. The maximum percentages of CL8, CL20, and CL21 approach 75%. The maximum percentages of CL1 and CL2 approach 15%.



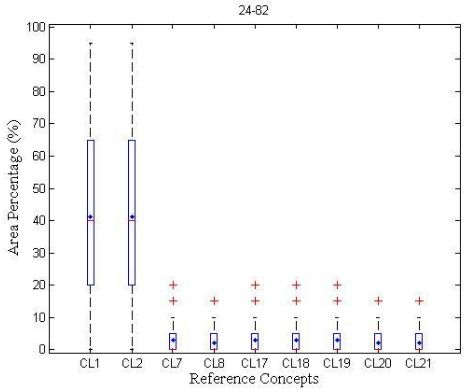
(a)



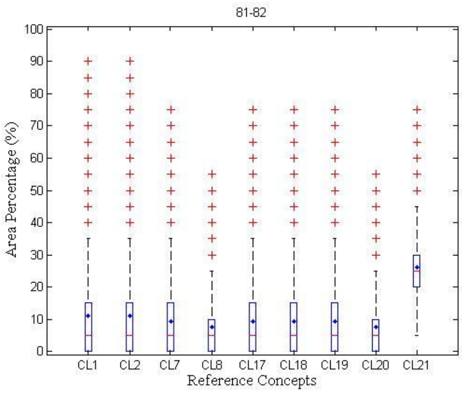
(b)



(c)



(d)



(e)

Figure 6.3 Analysis of the Most Significant Overlap for Each Level-II Class in Artificial Zone

The overlap between ‘22’ and ‘82’ is illustrated in Figure 6.3 (b). All medians and means are smaller than 15%, which shows that most overlaps are sparse. The mean is much larger than the median for CL8, CL20, and CL21, which indicates a significant skew toward the right with more values being smaller than the mean. The medians of CL1 and CL2 are larger than those of other reference concepts. The interquartile range for every reference concept is equal. Outliers exist for every reference concept. The maximum percentages of CL7, CL17, CL18, and

CL19 approach 80%. The maximum percentages of CL8, CL20, and CL21 approach 60%. The maximum percentages of CL1 and CL2 approach 45%.

The overlap between ‘23’ and ‘82’ is illustrated in Figure 6.3 (c). The medians and means of CL1 and CL2 are larger than 20%, which shows that most of these two are open. The medians and means of other reference concepts are smaller than 10%, which shows that most of them are sparse. The means of CL1 and CL2 are smaller than the medians, which indicate that most values are larger than the means. The means are larger than the medians for other reference concepts, which indicate that most values are smaller than the means. The medians of CL1 and CL2 are larger than the medians of other reference concepts. The largest interquartile range occurs on CL1 and CL2, which indicates that their values are more disperse. Outliers exist except for CL1 and CL2. The maximum percentages of CL7, CL17, CL18, and CL19 approach 50%. The maximum percentages of CL8, CL20, and CL21 approach 35%. The maximum percentages of CL1 and CL2 approach 75%.

The overlap between ‘24’ and ‘82’ is illustrated in Figure 6.3 (d). The medians and means of CL1 and CL2 are larger than 20%, which show that most of these two are open. The medians and means of other reference concepts are smaller than 10%, which show that most of those are sparse. The mean is larger than the median for every reference concept, which indicates that more values are smaller than the mean. The medians of CL1 and CL2 are much larger than the medians of other reference concepts. The largest interquartile range occurs on CL1 and CL2, which indicates that their values are more disperse. Outliers exist except for CL1 and CL2. The maximum percentages of CL7, CL17, CL18, and CL19 approach 20%. The maximum percentages of CL8, CL20, and CL21 approach 15%. The maximum percentages of CL1 and CL2 approach 95%.

The overlap between ‘81’ and ‘82’ is illustrated in Figure 6.3 (e). The median and mean of CL21 is larger than 20%, which shows that most of it is open. The medians and means of other reference concepts are smaller than 20%, which show that most of those are sparse. The mean is larger than the median for every reference concept, which indicates that more values are smaller than the mean.

The median of CL21 is much larger than that of other reference concepts. Outliers exist for every reference concept. The maximum percentages of CL7, CL17, CL18, CL19, and CL21 approach 75%. The maximum percentages of CL8 and CL20 approach 55%. The maximum percentages of CL1 and CL2 approach 90%.

6.3.3.3 Natural Zone

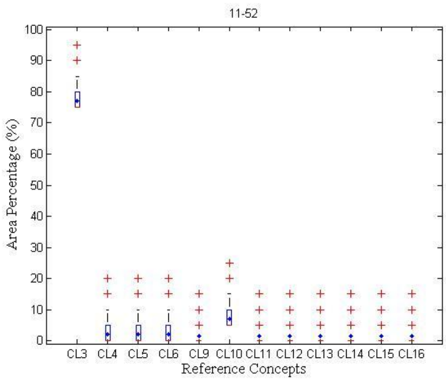
Compared with Table 6.4, more overlaps are depicted in Table 6.8. It shows that there is no overlap between class ‘70’ and other classes except for its subordinate class ‘71’, because the characteristic covering material belongs to vegetation and its area percentage is very high. The largest value of overlap (0.5001) occurs between ‘12’ and ‘43’, which is much lower than that (0.9612) in artificial zone.

The largest overlaps for every level-II class are underlined in Table 6.8 and analysed by boxplots in Figure 6.4.

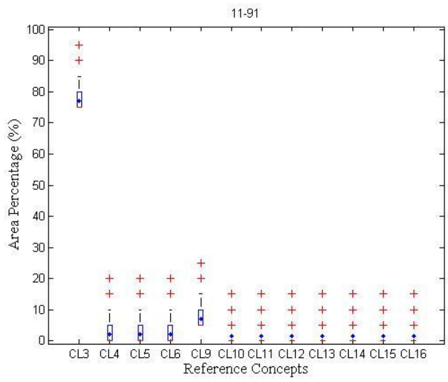
The largest overlap of ‘11’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.4 (a) and (b), respectively. These two figures express the same characteristics. The analysis for ‘11’ and ‘91’ can be acquired from that of ‘11’ and ‘52’ by interchangeably substituting ‘CL10’ with ‘CL9’. The median and mean of CL3 are larger than 70%, which show that most of this covering is closed, while the median and mean of other reference concepts are smaller than 10%, which show that most of those are sparse. The medians except for CL3 and CL10 are 0, and the means are much larger than medians, which indicate these concepts are rare in overlap areas, especially for CL9, CL11, CL12, CL13, CL14, CL15, and CL16 whose interquartile ranges are zero. Outliers exist for every class. The maximum percentage of CL3 approaches 95%. The maximum percentages of CL4, CL5, and CL6 approach 20%. The maximum percentages of CL9, CL11, CL12, CL13, CL14, CL15, and CL16 approach 15%. The maximum percentage of CL10 approaches 25%.

Table 6.8 Category Overlap in Natural Zone

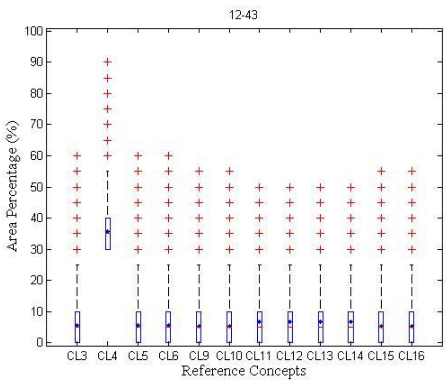
	11	12	31	40	41	42	43	52	70	71	91	92
11	1	0	0	0.0014	0.2633	0.2633	0.2601	*0.2644*	0	0.0160	*0.2644*	0.0160
12	0	1	0	0.1413	0.1906	0.1906	*0.5001*	0.2326	0	0.0015	0.2326	0.0015
31	0	0	1	0	*0.2826*	*0.2826*	0.1385	0.2296	0	0.0354	0.2296	0.0354
40	0	0.0262	0	1	0.2010	0.2010	0.5979	0.1352	0	0	0.1352	0
41	0.0001	0.0545	0.0001	0.3101	1	0	0	*0.2139*	0	0	*0.2139*	0
42	0.0001	0.0545	0.0001	0.3101	0	1	0	*0.2139*	0	0	*0.2139*	0
43	0	0.0490	0	0.3163	0	0	1	*0.2044*	0	0	*0.2044*	0
52	0.0001	0.0535	0.0001	0.1680	0.1723	0.1723	*0.4803*	1	0	0	0.1587	0
70	0	0	0	0.0014	0.0082	0.0082	0.0284	0.0011	1	0.6538	0.0011	0
71	0.0012	*0.1215*	0.0033	0	0	0	0	0	0.0508	1	0	0
91	0.0001	0.0535	0.0001	0.1680	0.1723	0.1723	*0.4803*	0.1587	0	0	1	0
92	0.0012	*0.1215*	0.0033	0	0	0	0	0	0	0	0	1



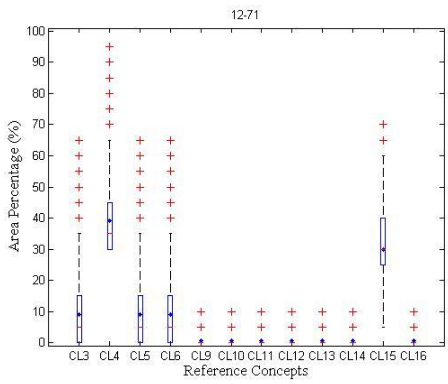
(a)



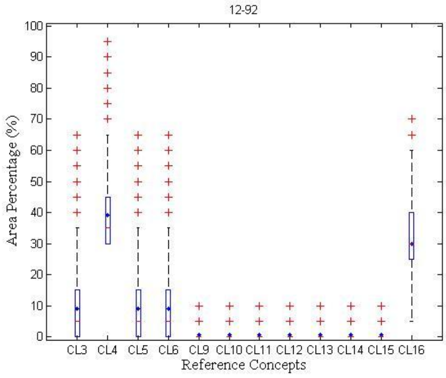
(b)



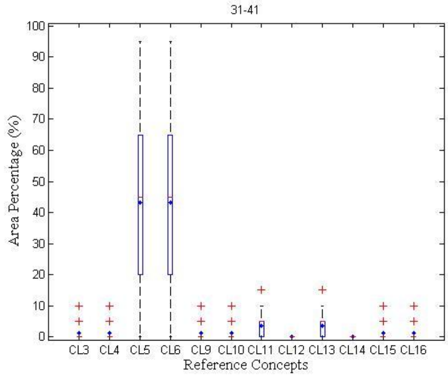
(c)



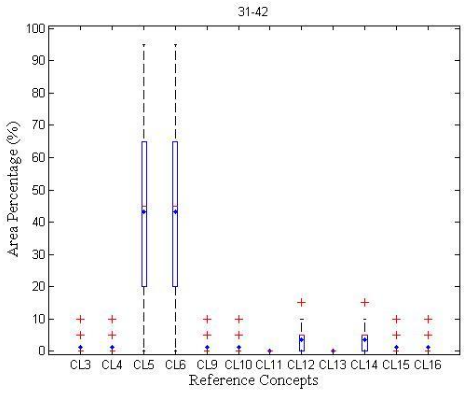
(d)


12-92
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

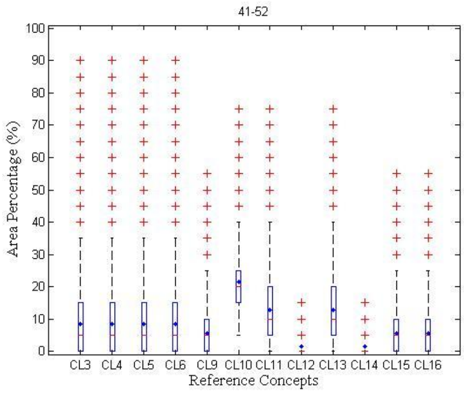
(e)


31-41
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

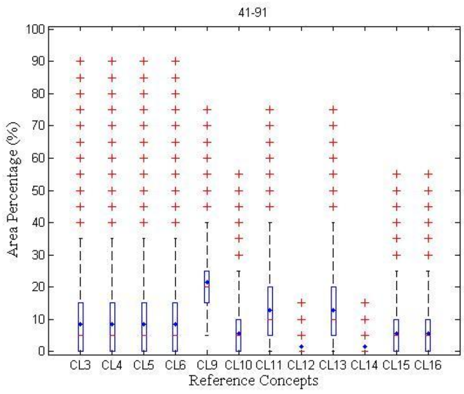
(f)


31-42
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

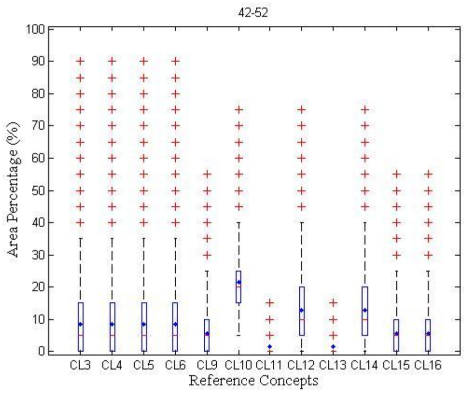
(g)


41-52
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

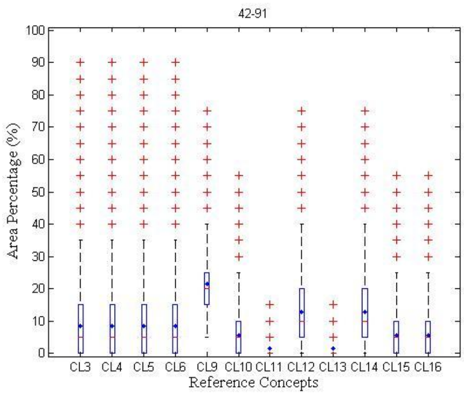
(h)

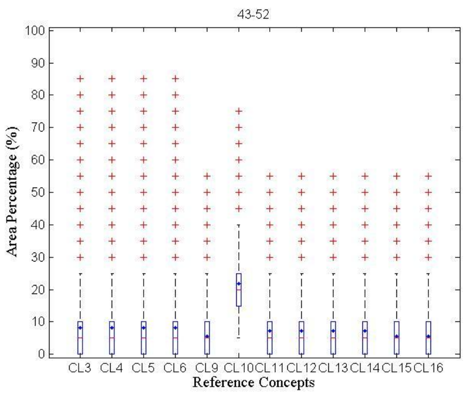

41-91
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

(i)

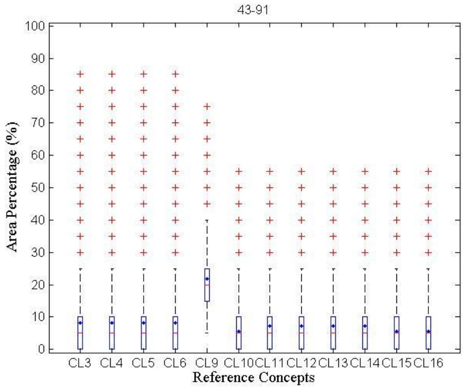

42-52
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

(j)


42-91
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts


43-52
Area Percentage (%)
0 10 20 30 40 50 60 70 80 90 100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

(k) (l)



(m)

Figure 6.4 Analysis of the Most Significant Overlap for each Level-II Class in Natural Zone

The overlap between '12' with '43' is illustrated in Figure 6.4 (c). The median and mean of CL4 are larger than 30%, which shows that most of this covering is open, while the median and mean of other reference concepts are smaller than 10%, which show that most of those are sparse. All means are larger than medians, which indicate more values are smaller than means. Outliers exist for every class. The maximum percentages of CL3, CL5, and CL6 approach 60%. The maximum percentage of CL4 approaches 90%. The maximum percentages of CL9, CL10, CL15, and CL16 approach 55%. The maximum percentages of CL11, CL12, CL13, and CL14 approach 50%.

The largest overlap of '31' occurs with '41' and '42', which are illustrated in Figure 6.4 (f) and (g), respectively. These two figures express the same characteristics. The analysis for '31' and '42' can be acquired from that of '31' and '41' by interchangeably substituting 'CL11' and 'CL13' with 'CL12' and 'CL14'. The means and medians of CL5 and CL6 are larger than 40%, which indicate they are open, while others are smaller than 10%, which indicate they are sparse. The means of CL5, CL6, CL11 and CL13 are smaller than the medians, which indicate most values are larger than means. The largest interquartile range occurs on CL5 and CL6, and the second occurs on CL11 and CL13. Interquartile ranges are 0 for other reference concepts. In this type of overlap, most of the compositions are CL5, CL6, CL11 and CL13. Outliers exist except for CL5, CL6, CL12, and CL14. The maximum percentages of CL3, CL4, CL9, CL10, CL15, and CL16 approach 10%. The maximum percentages of CL5 and CL6 approach 95%. The maximum percentages of CL11 and CL13 approach 15%. CL12 and CL14 do not exist.

The largest overlap of ‘41’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.4 (h) and (i), respectively. These two figures express the same characteristics. The analysis for ‘41’ and ‘91’ can be acquired from that of ‘41’ and ‘52’ by interchangeably substituting ‘CL10’ with ‘CL9’. The median and mean of CL10 are larger than 20%, which show that most of this covering is open, while the median and mean of other reference concepts are smaller than 20%, which show that most of those are sparse. All means are much larger than medians, which indicate more values are smaller than the means. The interquartile ranges of CL12 and CL14 are 0. Outliers exist in every concept. The maximum percentages of CL3, CL4, CL5, and CL6 approach 90%. The maximum percentages of CL9, CL15, and CL16 approach 55%. The maximum percentages of CL10, CL11, and CL13 approach 75%. The maximum percentages of CL12 and CL14 approach 15%.

The largest overlap of ‘42’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.4 (j) and (k), respectively. These two figures are similar to Figure 6.4 (h) and (i), respectively, which can be interpreted by interchangeably substituting ‘CL11’ and ‘Cl13’ with ‘CL12’ and ‘CL14’.

The largest overlap of ‘43’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.4 (l) and (m), respectively. These two figures express the same characteristics. The analysis for ‘43’ and ‘91’ can be acquired from that of ‘43’ and ‘52’ by interchangeably substituting ‘CL10’ with ‘CL9’. The median and mean of CL10 are larger than 20%, which show that most of this covering is open. The median and mean of other reference concepts are smaller than 10%, which show that most of those are sparse. All means are much larger than medians, which indicate more values are smaller than the means. Outliers exist in every concept. The maximum percentages of CL3, CL4, CL5, and CL6 approach 85%. The maximum percentages of CL9, CL11, CL12, CL13, CL14, CL15, and CL16 approach 55%. The maximum percentage of CL10 approaches 75%.

The overlap between ‘71’ with ‘12’ is illustrated in Figure 6.4 (d). The medians and means are larger than 30% for CL4 and CL15, which indicate they are open, while others are smaller than 10%, which indicate they are sparse. The means are much larger than medians for CL3, CL4, CL5, and CL6. The interquartile ranges for CL9, CL10, CL11, CL12, CL13, CL14, and CL16 are 0, which indicate they are rare in overlap areas. This type of overlap is mainly constituted by CL3, CL4, CL5, CL6, and CL15. Outliers exist for every class. The maximum percentages of CL3, CL5, and CL6 approach 65%. The maximum percentage of CL4 approaches 95%. The maximum percentages of CL9, CL10, CL11, CL12, CL13, CL14, and CL16 approach 10%. The maximum percentage of CL15 approaches 70%.

The overlap between ‘92’ with ‘12’ is illustrated in Figure 6.4 (e) and can be interpreted by interchangeably substituting ‘CL15’ with ‘CL16’ in Figure 6.4 (d).

The results of original category overlap reveal that it is hard to reach a strict subordinate-superordinate relationship from a category perspective if classes in different levels are defined separately. Overlaps are prone to occur for classes which are defined by characteristic covering with small area percentage, because it is more likely to occur in a land parcel in practice.

6.3.4 Category Overlap Constrained by Level-I Definition

6.3.4.1 Constrained Level-II Definition

The attributes of level-I classes should be inherited by their subordinate level-II classes. Considering level-I class definitions, 6 level-II classes are constrained: 2 in the artificial zone and 4 in the natural zone. Characteristic coverings and their categories are represented in Table 6.9 and Table 6.10 respectively. Inherited information is highlighted in grey.

Table 6.9 Constrained Level-II Classes in Artificial Zone

Code	Characteristic Covering	Condition
81	CL21	(1)AP(CL21)/AP(CL7+CL8+CL17+CL18+CL19+ CL20+CL21)>20%; (2) AP(CL17+CL18+CL19+CL20+CL21)>=75%
82	CL7, CL17, CL18, CL19	(1)AP(CL7+CL17+CL18+CL19)/AP(CL7+CL8+CL 17+CL18+CL19+ CL20+CL21)>20%; (2) AP(CL17+CL18+CL19+CL20+CL21)>=75%

Table 6.10 Constrained Level-II Classes in Natural Zone

Code	Characteristic Covering	Condition
41	CL11, CL13	(1)AP(CL11+CL12+CL13+CL14)/AP(CL9+CL10+C L11+CL12+CL13+CL14+CL15+CL16)>20%, and (2)AP(CL11+CL13)/AP(CL11+CL12+CL13+CL14)> 75%; (3) AP(CL11+CL12)>=25%
42	CL12, CL14	(1)AP(CL11+CL12+CL13+CL14)/AP(CL9+CL10+C L11+CL12+CL13+CL14+CL15+CL16)>20%, and (2)AP(CL12+CL14)/AP(CL11+CL12+CL13+CL14)> 75%; (3) AP(CL11+CL12)>=25%
43	CL11, CL12,CL13, CL14	(1)AP(CL11+CL12+CL13+CL14)/AP(CL9+CL10+C L11+CL12+CL13+CL14+CL15+CL16)>20% ,and (2)AP(CL11+CL13)/AP(CL11+CL12+CL13+CL14)< =75%, and (3)AP(CL12+CL14)/AP(CL11+CL12+CL13+CL14)< =75%; (4) AP(CL11+CL12)>=25%
71	CL15	(1)AP(CL15)/AP(CL9+CL10+CL11+CL12+CL13+C L14+CL15+CL16)>80%; (2) AP(CL15)>=75%

6.3.4.2 Changes of Class Distribution

The dominated percentages of all constrained classes decrease dramatically. The results are expressed and compared in Table 6.11 and Table 6.12.

Table 6.11 Changes of Class Distribution in Artificial Zone

	Constrained		Not Constrained	
	81	82	81	82
Percentage (%)	7.31	19.28	27.19	95.00

Table 6.12 Changes of Class Distribution in Natural Zone

	Constrained				Not Constrained			
	41	42	43	71	41	42	43	71
Percentage (%)	5.715	5.715	16.998	0.003	18.427	18.427	53.748	0.066

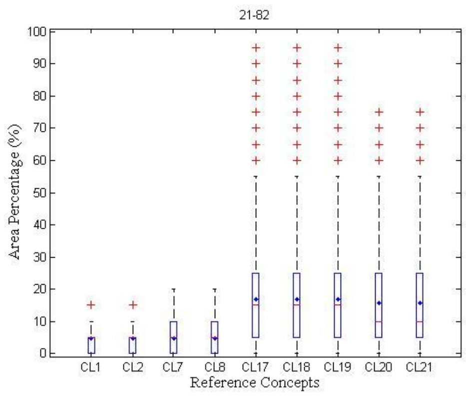
6.3.4.3 Artificial Zone

Comparing Table 6.13 with Table 6.7, overlaps between ‘23’ and ‘24’ with other level-II classes have been removed. Some overlap measurements grow because the number of possible mixtures is decreased in the situation considering the constraint of level-I definitions. The maximum overlaps still appear between ‘21’ and ‘82’, ‘22’ and ‘82’, and ‘81’ and ‘82’, except that it changes from between ‘82’ and ‘22’ to between ‘82’ and ‘21’ for class 82.

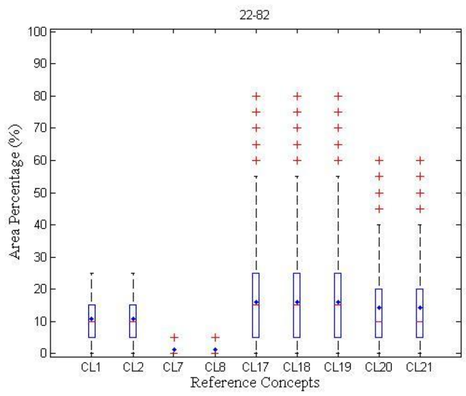
Table 6.13 Category Overlap in Artificial Zone

	20	21	22	23	24	80	81	82
20	1	0	0.7658	0.2285	0.0057	0	0	0
21	0	1	0	0	0	0.3594	0.1370	*0.3462*
22	0.5273	0	1	0	0	0.0584	0.0209	*0.0548*
23	1	0	0	1	0	0	0	0
24	1	0	0	0	1	0	0	0
80	0	0.6676	0.1379	0	0	1	0.3635	0.9580
81	0	*0.7001*	0.1361	0	0	1	1	*0.9166*
82	0	*0.6712*	0.1352	0	0	1	0.3477	1

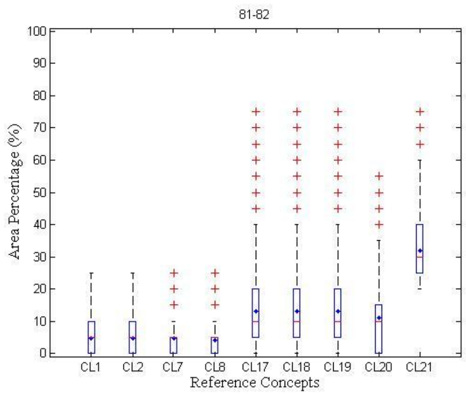
Elements of overlaps between level-II classes constrained by level-I definitions are analysed by boxplots as shown in Figure 6.5.



(a)



(b)



(c)

Figure 6.5 Analysis of the Most Significant Overlap for Each Constrained Level-II Class in Artificial Zone

The overlap between ‘21’ and ‘82’ is illustrated in Figure 6.5 (a). All medians and means are smaller than 15%, which show that most overlaps are sparse. The mean is approximately the same as the median for CL1, CL2, CL7 and CL8, while the mean is much larger than the median for the rest. The interquartile ranges of CL17, CL18, CL19, CL20, and CL21 are larger than the rest. Outliers exist except for CL7 and CL8. The maximum percentages of CL17, CL18, and CL19 approach 95%. The maximum percentages of CL20 and CL21 approach 75%. The maximum percentages of CL1 and CL2 approach 15%. The maximum percentages of CL7 and CL8 approach 20%.

Compared with Figure 6.3 (a), the means and medians decrease for CL7 and CL8, while they increase for CL17, CL18, and CL19. The interquartile ranges of CL1 and CL2 become narrower. The maximum percentages of CL7 and CL8 are reduced.

The overlap between '22' and '82' is illustrated in Figure 6.5 (b). All medians and means are smaller than 15%, which show that most overlaps are sparse. The mean is larger than the median for every reference concept, which indicates that more values are smaller than the mean. The values of CL7 and CL8 are concentrated on 0%, which indicates they seldom appear in the overlap areas. Outliers exist except for CL1 and CL2. The maximum percentages of CL17, CL18, and CL19 approach 80%. The maximum percentages of CL20 and CL21 approach 60%. The maximum percentages of CL1 and CL2 approach 25%. The maximum percentages of CL7 and CL8 approach 5%.

Compared with Figure 6.3 (b), the means and medians decrease for CL1, CL2, CL7 and CL8 while they increase for CL17, CL18, CL19, CL20 and CL21. They are more concentrated for values of CL7 and CL8, while they are more disperse for CL17, CL18, and CL19. The maximum percentages of CL1, CL2, CL7 and CL8 are reduced.

The overlap between '81' and '82' is illustrated in Figure 6.5 (c). The median and mean of CL21 are larger than 30%, which show that most of this covering is open, while the median and mean of other reference concepts are smaller than 20%, which show that most of those are sparse. The mean is approximately the same as the median for CL1, CL2, CL7, and CL8, while the mean is much larger than the median for the rest. Outliers exist except for CL1 and CL2. The maximum percentages of CL17, CL18, CL19, and CL21 approach 75%. The maximum percentage of CL20 approaches 60%. The maximum percentages of CL1, CL2, CL7, and CL8 approach 25%.

Compared with Figure 6.3 (e), the means and medians increase for CL17, CL18, CL19, CL20, and CL21. They are more concentrated for values of CL1, CL2, CL7, and CL8, while they are more disperse for CL20 and CL21. The maximum percentages of CL1, CL2, CL7, and CL8 are reduced.

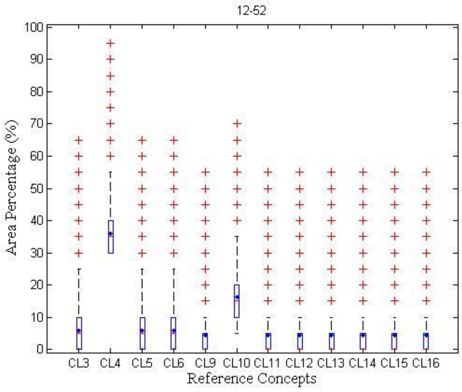
6.3.4.4 Natural Zone

Comparing Table 6.14 with Table 6.8, some overlaps have been removed, such as overlaps between ‘71’ and other level-II classes. As shown in Table 6.12, only the four classes ‘41’, ‘42’, ‘43’, and ‘71’ are changed. Overlaps not involving these four classes are unchanged. However, the largest overlap for every class is changed except for ‘11’ and ‘92’. The largest overlap for ‘12’ occurs with ‘52’ and ‘91’ instead of ‘43’. The largest overlap for ‘31’ occurs with ‘52’ and ‘91’ instead of ‘41’ and ‘42’. The largest overlap for ‘41’, ‘42’ and ‘43’ still occur with ‘52’ and ‘91’ but the value decreases. The largest overlap for ‘52’ occurs with ‘91’ instead of ‘43’. The largest overlap for ‘91’ occurs with ‘52’ instead of ‘43’.

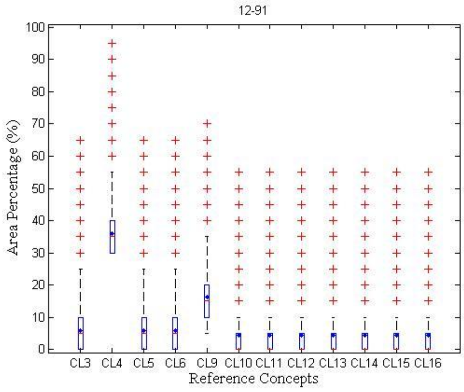
Table 6.14 Category Overlap in Natural Zone

	11	12	31	40	41	42	43	52	70	71	91	92
11	1	0	0	0.0014	0.0005	0.0005	0.0005	*0.2644*	0	0	*0.2644*	0.0160
12	0	1	0	0.1413	0.0313	0.0313	0.0788	*0.2326*	0	0	*0.2326*	0.0015
31	0	0	1	0	0	0	0	*0.2296*	0	0	*0.2296*	0.0354
40	0	0.0262	0	1	0.2010	0.2010	0.5979	0.1352	0	0	0.1352	0
41	0	0.0288	0	1	1	0	0	*0.1456*	0	0	*0.1456*	0
42	0	0.0288	0	1	0	1	0	*0.1456*	0	0	*0.1456*	0
43	0	0.0244	0	1	0	0	1	*0.1282*	0	0	*0.1282*	0
52	0.0001	0.0535	0.0001	0.1680	0.0364	0.0364	0.0953	1	0	0	*0.1587*	0
70	0	0	0	0.0014	0.0005	0.0005	0.0005	0.0011	1	0.6538	0.0011	0
71	0	0	0	0	0	0	0	0	1	1	0	0
91	0.0001	0.0535	0.0001	0.1680	0.0364	0.0364	0.0953	*0.1587*	0	0	1	0
92	0.0012	*0.1215*	0.0033	0	0	0	0	0	0	0	0	1

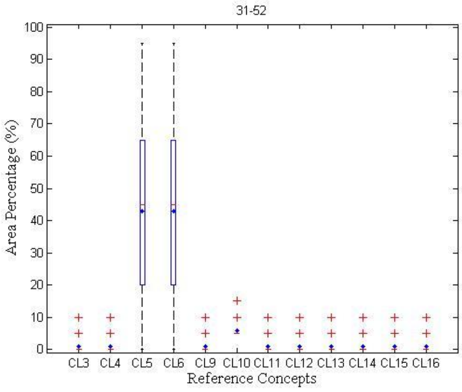
Elements of overlaps between level-II classes constrained by level-I definitions are analysed by boxplots as shown in Figure 6.6.


12-52
Area Percentage (%)
0
10
20
30
40
50
60
70
80
90
100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

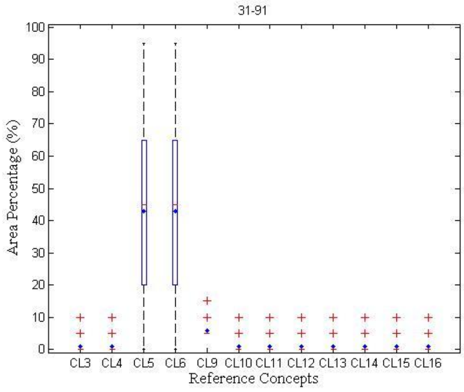
(a)


12-91
Area Percentage (%)
0
10
20
30
40
50
60
70
80
90
100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

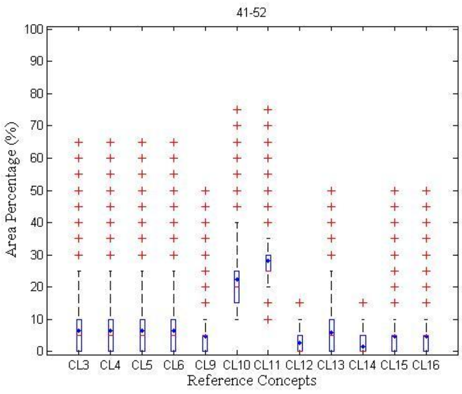
(b)


31-52
Area Percentage (%)
0
10
20
30
40
50
60
70
80
90
100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

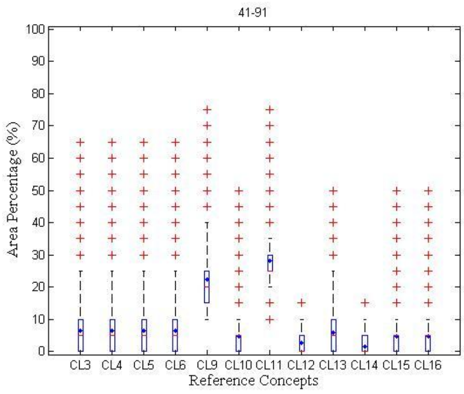
(c)


31-91
Area Percentage (%)
0
10
20
30
40
50
60
70
80
90
100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

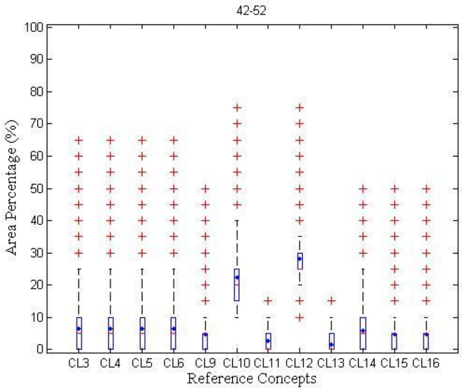
(d)


41-52
Area Percentage (%)
0
10
20
30
40
50
60
70
80
90
100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

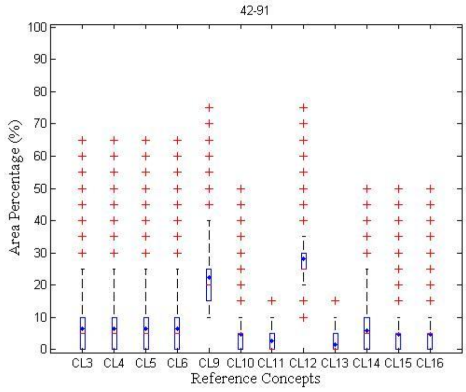
(e)


41-91
Area Percentage (%)
0
10
20
30
40
50
60
70
80
90
100
CL3 CL4 CL5 CL6 CL9 CL10 CL11 CL12 CL13 CL14 CL15 CL16
Reference Concepts

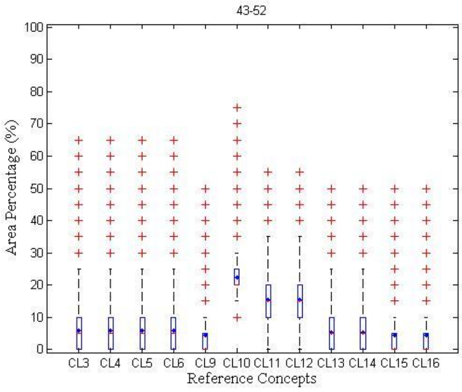
(f)



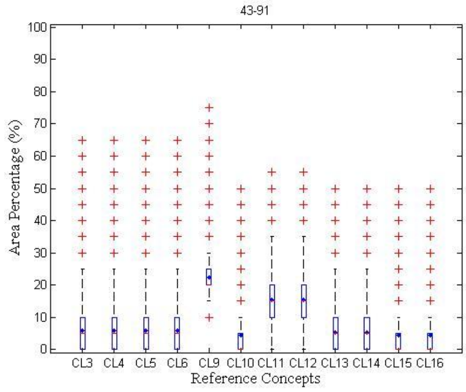
(g)



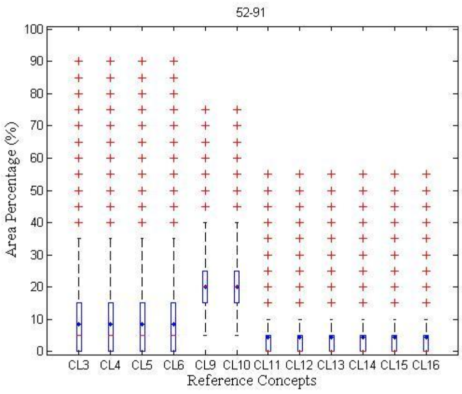
(h)



(i)



(j)



(k)

Figure 6.6 Analysis of the Most Significant Overlap for Each Constrained Level-II Class in Natural Zone

The largest overlap of '12' occurs with '52' and '91', which are illustrated in Figure 6.6(a) and (b), respectively. These two figures express the same characteristics. The analysis for '12' and '91' can be acquired from that of '12' and '52' by interchangeably substituting 'CL10' with 'CL9'. The median and mean of CL4 are larger than 30%, which show that most of this covering is open, while the median and mean of other reference concepts are smaller than 20%, which show that most of those are sparse. All means are larger than medians, which indicate more values are smaller than the means. This type of overlap is dominated by CL4 and CL10. Outliers exist for every concept. The maximum percentages of CL3, CL5, and CL6 approach 65%. The maximum percentage of CL4 approaches 95%. The maximum percentages of CL9 CL11, CL12, CL13, CL14, CL15, and CL16 approach 55%. The maximum percentage of CL10 approaches 70%.

The largest overlap of ‘31’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.6(c) and (d), respectively. These two figures express the same characteristics. The analysis for ‘31’ and ‘91’ can be acquired from that of ‘31’ and ‘52’ by interchangeably substituting ‘CL10’ with ‘CL9’. The means and medians of CL5 and CL6 are larger than 40%, which indicate they are open, while others are smaller than 10%, which indicate they are sparse. The largest interquartile range occurs in CL5 and CL6, and others are 0. This type of overlap is dominated by CL5, CL6, and CL10. Outliers exist for every concept. The maximum percentages of CL3, CL4, CL9, CL11, CL12, CL13, CL14, CL15, and CL16 approach 10%. The maximum percentages of CL5 and CL6 approach 95%. The maximum percentage of CL10 approaches 15%.

The largest overlap of ‘41’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.6(e) and (f), respectively. These two figures express the same characteristics. The analysis for ‘41’ and ‘91’ can be acquired from that of ‘41’ and ‘52’ by interchangeably substituting ‘CL10’ with ‘CL9’. The median and mean of CL10 and CL11 are larger than 20%, which show that most of these coverings are open, while the median and mean of other reference concepts are smaller than 10%, which show that most of those are sparse. The means are smaller than medians for CL9, CL15, and CL16, while they are larger for others. Outliers exist for every concept. The maximum percentages of CL3, CL4, CL5, and CL6 approach 65%. The maximum percentages of CL9, CL13, CL15, and CL16 approach 50%. The maximum percentages of CL10 and CL11 approach 75%. The maximum percentages of CL12 and CL14 approach 15%.

Compared with Figure 6.4 (h), the means and medians of CL11 are larger than CL10. The maximum percentage is reduced for every concept except for CL10, CL11, CL12, and CL14.

The largest overlap of ‘42’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.6(g) and (h), respectively. These two figures are similar to Figure 6.6(e) and (f), respectively, which can be interpreted by interchangeably substituting ‘CL11’ and ‘Cl13’ with ‘CL12’ and ‘CL14’.

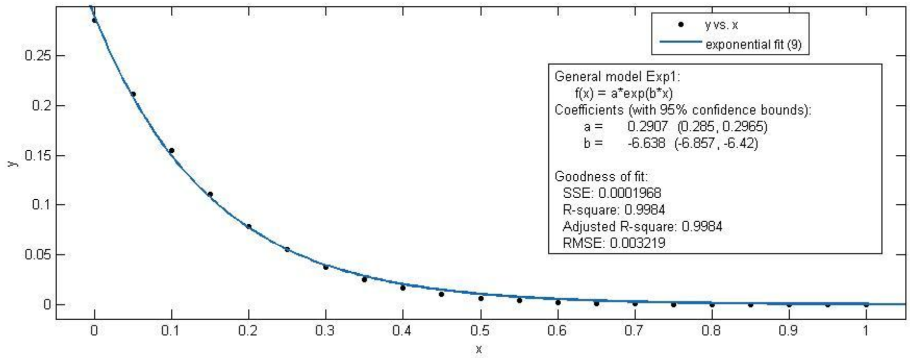
The largest overlap of ‘43’ occurs with ‘52’ and ‘91’, which are illustrated in Figure 6.6 (i) and (j), respectively. These two figures express the same characteristics. The analysis for ‘43’ and ‘91’ can be acquired from that of ‘43’ and ‘52’ by interchangeably substituting ‘CL10’ with ‘CL9’. The median and mean of CL10 are larger than 20%, which show that most of this covering is open, while the median and mean of other reference concepts are smaller than 20%, which show that most of those are sparse. All means are larger than medians, which indicate more values are smaller than the means. Outliers exist for every class. The maximum percentages of CL3, CL4, CL5, and CL6 approach 65%. The maximum percentages of CL9, CL13, CL14, CL15, and CL16 approach 50%. The maximum percentage of CL10 approaches 75%. The maximum percentages of CL11 and CL12 approach 55%.

The overlap between ‘52’ and ‘91’ is illustrated in Figure 6.6 (k). The medians and means of all concepts are smaller than 20%, which show that most of those are sparse. All means are larger than medians, which indicate more values are smaller than the means. Outliers exist for every class. The maximum percentages of CL3, CL4, CL5, and CL6 approach to 90%. Maximum percentages of CL9 and CL10 approach 75%. The maximum percentages of CL11, CL12, CL13, CL14, CL15, and CL16 approach 55%.

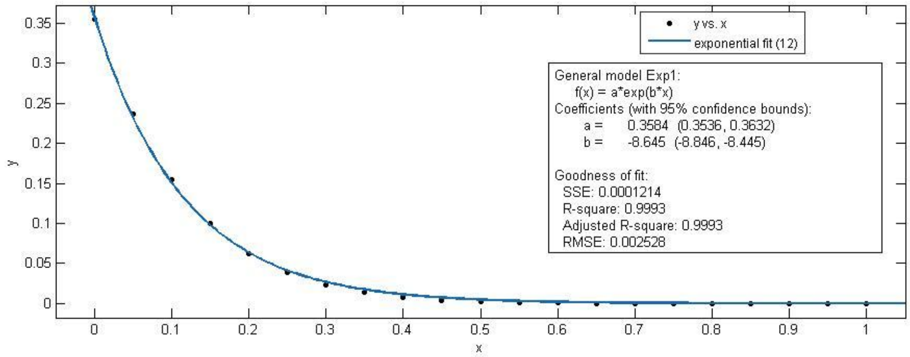
6.4 Discussion

6.4.1 Distribution of Area Percentage

The combinations of reference concepts are not evenly distributed. For example, the number of combinations of 'less than 10%' is dramatically different from that of 'greater than 90%'. Combinations concentrate in low area percentage and decrease exponentially. The distribution of combinations in the artificial zone and natural zone are illustrated in Figure 6.7 (a) and (b), respectively. The greater the number of reference concepts, the sparser the combinations.



(a)



(b)

Figure 6.7 Distribution of Area Percentage

6.4.2 Category Overlap Considering Priority

Overlaps still exist between level-II classes constrained by superordinate class definitions. In practice, such conflicts are solved based on the judgement of interpreters or feedback from authorities. Usually, the priority of classes plays a key role. Although a parcel fulfils definitions of more than one class, it is more suitable to assign it to some class which is of more interest to the project. Normally, the more the class relates to human beings, the more important the class is. A possible priority for classes in the case study is listed in Table 6.15 and Table 6.16. The priority of level-I classes and level-II classes is ordered separately and the priority of level-II classes is higher than that of level-I classes. For example, the priority is 1 for both '24' and '20'. The parcel is assigned '24' when it belongs to both of them.

Table 6.15 Priority Table in Artificial Zone

Level-I	20				80	
Priority	1				2	
Level-II	21	22	23	24	81	82
Priority	4	3	2	1	6	5

Table 6.16 Priority Table in Natural Zone

Level-I	10		30	40			50	70	9	
Priority	1		6	3			4	5	2	
Level-II	11	12	31	41	42	43	52	71	91	92
Priority	2	1	10	6	5	7	8	9	3	4

The dominance of every class dramatically decreases when considering priorities. The distribution of classes is illustrated in Table 6.17 and Table 6.18.

Table 6.17 Class Distribution Considering Priority in Artificial Zone

	20	21	22	23	24	80	81	82
Percentage (%)	32.74	37.38	47.55	7.48	0.19	3.92	0.12	3.73

Table 6.18 Class Distribution Considering Priority in Natural Zone

	10	11	12	30	31	40	41	42
Percentage (%)	-	0.005	5.264	-	0.003	23.915	4.736	4.736
	43	50	52	70	71	9	91	92
Percentage (%)	14.442	-	14.680	0.005	0.003	-	21.649	0.058

Semantic overlaps are described in Table 6.19 and Table 6.20. There is no overlap between classes in a level.

Table 6.19 Category Overlap in Artificial Zone

	20	21	22	23	24	80	81	82
20	1	0	0.7658	0.2285	0.0057	0	0	0
21	0	1	0	0	0	0	0	0
22	0.5273	0	1	0	0	0	0	0
23	1	0	0	1	0	0	0	0
24	1	0	0	0	1	0	0	0
80	0	0	0	0	0	1	0.0319	0.9532
81	0	0	0	0	0	1	1	0
82	0	0	0	0	0	1	0	1

Table 6.20 Category Overlap in Natural Zone

	11	12	31	40	41	42	43	52	70	71	91	92
11	1	0	0	0	0	0	0	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0	0	0	0
31	0	0	1	0	0	0	0	0	0	0	0	0
40	0	0	0	1	0.1981	0.1981	0.6039	0	0	0	0	0
41	0	0	0	1	1	0	0	0	0	0	0	0
42	0	0	0	1	0	1	0	0	0	0	0	0
43	0	0	0	1	0	0	1	0	0	0	0	0
52	0	0	0	0	0	0	0	1	0	0	0	0
70	0	0	0	0	0	0	0	0	1	0.6563	0	0
71	0	0	0	0	0	0	0	0	1	1	0	0
91	0	0	0	0	0	0	0	0	0	0	1	0
92	0	0	0	0	0	0	0	0	0	0	0	1

6.5 Conclusion

This chapter proposes a method to quantify semantic overlaps based on DSRS. The range of the measurement is $\left[0,1\right]$. The larger the value is, the more overlaps occur. By employing NLCD2001 CS classes as examples, semantic overlaps were calculated and analysed. The results revealed that many overlaps exist between classes. Overlaps are more likely to occur if the covering materials are sparse. NLCD2001 CS is not a rigorous tree structure. Subordinate classes are not entirely covered by their superordinate class and even exceed the scope of their superordinate class. This problem is called semantic gap and overflow and will be examined in Chapter 7.

CHAPTER 7 SEMANTIC GAP AND SEMANTIC OVERFLOW

7.1 Introduction

Theoretically, every land parcel in the real world should be assigned to a class. However, some parcels are not defined in the classification system. Semantic gaps are left in current land classification systems because the vast number of combinations belonging to a class cannot be handled by current methods of definitions (Di Gregorio, 2005, P.8). For example, there is a gap in the EarthSat GeoCover Global Land Cover legend where no classes are defined to categorize land with tree height >3 meters and tree cover <35% (Herold and Schmullius, 2004).

The objectives of this chapter are:

- To quantify semantic gaps in a land classification system;
- To propose and quantify semantic overflows in a land classification system;
- To measure semantic gap and overflow demonstrated on NLCD2001 CS.

The remainder of this chapter is structured as follows. Measurements for evaluating semantic gap and overflow are proposed in section 7.2. Section 7.3 demonstrates the performance of the measurements based on the Dynamic Semantic Reference System (DSRS). Discussions on semantic gap and overflow are made in section 7.4. Finally, conclusions are presented in section 7.5.

7.2 Modelling Semantic Gap and Overflow

Sematic gap and overflow occur between two groups of definitions which should be identical but are not. They measure inconsistencies of meanings of classes between hierarchical levels in a classification system. There exists a semantic gap if the meanings of classes in a level are not completely covered by their nominal descendants. There exists semantic overflow if the meanings of classes in a level exceed the scope of their ancestors. In other words, the semantic gap of a group of classes refers to the uncovered parts by their nominal descendants, while the semantic overflow of a group of classes refers to the parts exceeded by their nominal descendants.

Semantic gap and overflow can occur simultaneously. Measurements of them are proposed as shown in (7.1) and (7.2).

$$\mathrm{GAP}\left(C\right)=\frac{1}{\left|C\right|}\left(\left|C\right|-\left|C\cap\left(\bigcup\left(C^{d}\right)\right)\right|\right) \tag{7.1}$$

$$\mathrm{OF}\left(C\right)=\frac{1}{\left|\bigcup\left(C^{d}\right)\right|}\left(\left|\bigcup\left(C^{d}\right)\right|-\left|C\cap\left(\bigcup\left(C^{d}\right)\right)\right|\right) \tag{7.2}$$

where C is a group of definitions which can be concepts or categories, $GAP\left(C\right)$ is the semantic gap of C, $OF\left(C\right)$ is the semantic overflow of C, C^{d} is one nominal descendant of C, $\bigcup\left(C^{d}\right)$ is the union of all descendants of C, $\cap$ is an intersection operator which returns common parts between C and the union of its descendants, and $\left|\ \right|$ returns the number of compositions. When C is a concept, $\left|C\right|$ returns the number of reference concepts which constitute C. When C is a category, $\left|C\right|$ returns the number of combinations which constitute C.

$\left|C\right|-\left|C\cap\left(\bigcup\left(C^{d}\right)\right)\right|$ is an absolute measurement of uncovered parts of C, which is normalized by $\left|C\right|$. The range of $GAP\left(C\right)$ is $\left[0,1\right]$. 0 means the meaning of C is entirely covered by its descendants, while 1 means the meaning of C is totally uncovered by its descendants, i.e., the larger the value, the larger the gap. $\left|\bigcup\left(C^{d}\right)\right|-\left|C\cap\left(\bigcup\left(C^{d}\right)\right)\right|$ is an absolute measurement of exceeded parts of C, which is normalized by $\left|\bigcup\left(C^d\right)\right|$. The range of $OF\left(C\right)$ is also $\left[0,1\right]$. 0 means the meaning of C includes the meanings of its descendants, while 1 means the meaning of C is totally different from its descendants, i.e., the larger the value, the larger the overflow.

C is identical to the union of its descendants if and only if $GAP\left(C\right)=0$ and $\mathrm{OF}\left(\mathrm{C}\right)=0$. C is completely different from its descendants if $GAP\left(C\right)=1$ or $\mathrm{OF}\left(\mathrm{C}\right)=1$. The smaller the value is, the better the two levels match.

7.3 Case Study

7.3.1 Experimental Data

NLCD2001 CS (see Table 5.4), which owns 8 level-I classes and 16 level-II classes, is selected to evaluate the performance of the proposed measurements. The four level-I classes ‘10’, ‘30’, ‘50’, and ‘9’ are defined as concepts, while other level-I classes and all level-II classes are defined as categories. Every definition is considered independently, although they are arranged in a hierarchical structure. In other words, the characteristics of level-I classes are not inherited by their descendants. The meanings of classes are to be rebuilt by reference concepts based on DSRS, which has been established in Chapter 5. There are a total of 21 reference concepts, of which 9 are located in the artificial zone and 12 are located in the natural zone. A class can be represented by either its characteristic coverings or a category depicted in Table 6.1, Table 6.2, Table 6.9, or Table 6.10. Hence, concept-level measurements are calculated based on characteristic coverings, while category-level measurements are calculated based on reference concept combinations in a category.

7.3.2 Concept-level Gap and Overflow

Classes are represented by their characteristic coverings, which have been depicted in Table 7.1. It is obvious that neither of them completely cover the whole 21 reference concepts. Comparing level-I with level-II, problems occur between ‘40’, ‘50’, and ‘80’ and their descendants. Based on equation (7.1) and (7.2), semantic gaps and overflows are calculated and depicted in Table 7.2, Table 7.3, and Table 7.4.

Table 7.1 Characteristic Coverings of Level-I and Level-II Classes

Level-I Class		Union of Level-II Class	
LI Code	Characteristic Covering	LII Code	Characteristic Covering
10	CL3, CL4	11,12	CL3, CL4
20	CL1, CL2	20	CL1, CL2
30	CL5,CL6	31	CL5,CL6
40	CL11,CL12	41,43,43	CL11, CL12,CL13, CL14
50	CL10, CL13, CL14	52	CL10
70	CL15	71	CL15
80	CL17, CL18, CL19, CL20, CL21	81,82	CL7, CL17, CL18, CL19, CL21
9	CL9, CL16	91,92	CL9, CL16
uncovered	CL7, CL8	uncovered	CL8, CL20

The meanings of union of any level cannot exceed the universe. Hence, compared with the universe, there are 21 reference concepts, and only semantic gap may occur. Reference concepts in the natural zone are entirely covered by both levels, while gaps occur in the artificial zone for both levels. CL7 and CL8 are not characteristic coverings of any level-I classes, and CL8 and CL20 are not characteristic coverings of any level-II classes.

Table 7.2 Semantic Gap to the Universe

	Artificial Zone		Natural Zone	
	Level-I	Level-II	Level-I	Level-II
Gap	0.22	0.22	0	0

Comparing level-I with level-II, both gap and overflow exist in the artificial zone but neither exists in the natural zone. CL20 in level-I is not covered by level-II, and CL7 in level-II is not covered by level-I.

Table 7.3 Semantic Gap and Overlap between Levels

Level-I	Artificial Zone	Natural Zone
Gap	0.14	0
Overflow	0.14	0

Furthermore, semantic gap and overflow are performed for every level-I class. It is found that problems exist in both the artificial zone and the natural zone. There is overflow between ‘40’ and its descendants, and gap between ‘50’ and its descendants in the natural zone. There are also both gap and overflow between ‘80’ and its descendants.

Table 7.4 Semantic Gap and Overlap between Level-I Classes and Their Descendants

Level-I Classes	10	20	30	40	50	70	80	9
Gap	0	0	0	0	0.67	0	0.2	0
Overflow	0	0	0	0.5	0	0	0.2	0

The forehead three types of semantic gap and overflow reflect the details of semantic problems ranging from coarser to finer. The first is a kind of absolute evaluation to reveal whether a possible phenomenon in the real world can be classified by a classification level. The second evaluates the consistency between hierarchical levels to reveal whether meanings described at a level are perfectly described at another. The last evaluates the consistency between a single class and the union of its descendants to reveal whether the meaning of the class is identical to its descendants. In a certain area, some reference concepts may not exist. For example, ‘12. Perennial Ice/Snow’ cannot appear near the equator. Therefore, it is acceptable that there is gap to the universe in some situations. However, the second and third types should always be abandoned.

7.3.3 Category-level Gap and Overflow

In practice, land parcels are rarely pure but are usually a mixture of different reference concepts. Hence, classes are more reasonably represented by categories.

Because four level-I classes are not categories, measurements involving all of level-I are not conducted.

The absolute category-level semantic gap of level-II is depicted in Table 7.5. Although gaps exist, values are very small. This indicates that only very few situations are not considered in level-II classes of NLCD2001 CS. Compared with Table 7.2, they are dramatically different. The reasons are that: (1) land parcels are assumed to be pure in concept-level evaluations, while they are assumed to be mixtures in category-level evaluations; (2) only characteristic coverings are considered in concept-level evaluations, while all possible combinations are involved in category-level evaluations; and (3) although many combinations are not covered in level-II, the relative percentage is small because the total number of combinations is huge.

Table 7.5 Category-level Semantic Gap between Level-II and the Universe

Level-II	Artificial Zone	Natural Zone
gap	0.0018	0.0233

Semantic gap and overflow between level-I categories with their descendants are depicted in Table 7.6. No gaps exist for ‘20’ and ‘40’, and only a very small gap exists for ‘80’. Very large semantic overflow exists for every level-I category. Referencing their definitions in Table 5.4 and formalizations in Table 6.1 and Table 6.2, some findings can be used to partially explain the reasons.

Category ‘20’ and its descendants are defined by declaring an absolute area percentage (relative to the whole land parcel) of the same characteristic coverings but the covering percentage range of the former is smaller than the union of the latter. Category ‘40’ is also defined by declaring an absolute area percentage, while its descendants are defined by declaring a relative area percentage (relative to a certain coverings (such as vegetation) in a land parcel). The same situations occur for ‘70’ and ‘90’. Although it is hard to tell how to remove gaps and overflows, definitions based on relative area percentages are inclined to cause problems.

Table 7.6 Category-level Semantic Gap and Overflow between Level-I Classes and Their Descendants

	20	40	70	80
Gap	0	0	0.3462	0.0117
Overflow	0.8887	0.6862	0.9492	0.7966

7.4 Discussion

Most semantic overflows can be removed if level-I characteristics are inherited by level-II, such as ‘40’. Exceptions only occur if a contradiction is encountered when attributes in level-I are updated by level-II rather than new attributes being added. For example, there is a contradiction between definitions of ‘20’ and ‘21’, and corresponding characteristics of ‘20’ are not inherited by ‘21’. In practice, the products of level-I are not independently obtained, in contrast to level-II. They are an aggregation of level-II products, which prevent the possibility of semantic overflow.

Practically, all parcels are assigned class labels. But it does not indicate that there is no semantic gap in the classification system. Parcels will be allocated to the most similar classes if there are no directly matching class definitions. Semantic similarity will be introduced in Chapter 8.

7.5 Conclusion

This chapter proposed measurements to quantify semantic gap and overflow for the first time. Semantic overflow has not been mentioned by any author until this chapter. The quantitative results of semantic gap and overflow can accurately reflect semantic problems to the universe, between levels, and classes with their descendants, and can help to establish a classification system completely covering the area of interest and perfectly match between levels.

CHAPTER 8 SEMANTIC SIMILARITY

8.1 Introduction

The evaluation of semantic similarities plays an important role in different contexts including: classification definition (Sokal, 1974), categorization (Goldstone and Son, 2005), interpretation (Janowicz, 2008), information retrieval (Janowicz et al., 2011), and information integration (Hakimpour and Geppert, 2001). Previous studies are mainly conducted by psychologists (Gentner and Markman,1994; Goldstone and Son, 2005), and until recently, semantic similarities have been of high concerned in geographic science for reasons such as requirements for interoperation between different systems (Sheth, 1999).

The objectives of this chapter are:

- To propose a new similarity model based on DSRS.
- To make a systematic experiment by employing NLCD2001 CS for demonstration.

The remainder of this chapter is structured as follows. Section 8.2 briefly reviews and compares existing similarity models. Next, a new similarity model based on DSRS is proposed in Section 8.3. A case study on NLCD2001 CS is implemented in Section 8.4, and Section 8.5 provides a conclusion.

8.2 A Brief Review of Semantic Similarity Models

There are mainly four types of semantic similarity models: geometric model, feature model, network model, and transformational model.

8.2.1 Geometric Model

A concept in the geometric model is represented as a region in a multidimensional space. Each dimension is a property of the concept, and the range of the dimension represents all possible values of the property. Instead of measuring semantic similarity directly, geometric models measure the semantic distance between concepts. In analogy to spatial distance, a generic formula for the semantic distance measurement is the Minkowski Metric (equation (8.1)).

$$d\left(c_1, c_2\right) = \left[\sum_{i=1}^{n} \left|C_{1i} - C_{2i}\right|^r\right]^{1/r} \quad (8.1)$$

where C_1 and C_2 are two concepts, n is the number of dimensions used to describe the concept, and C_{1i} (C_{2i}) is the property value of concept C_1 (C_2) in the ith dimension. $r = 2$ results in the Euclidian distance, while $r = 1$ results in the city-block distance.

The similarity is a linear or exponentially decaying function of the distance (Melara et al., 1992). Equation (8.2) is a possible exponentially decaying function to transform semantic distance to semantic similarity, which results in a similarity normalized between 0 and 1.

$$s\left(c_1, c_2\right) = \frac{1}{1 + d(c_2, c_2)} \quad (8.2)$$

8.2.2 Feature Model

The basis for the feature model is set theory. The property values of a concept are represented as elements in a feature set. Semantic similarity is computed by taking into account both common and distinct features. Common features increase the similarity, while distinct features decrease the similarity. The most famous feature models are Tversky's (1977) Contrast Model (equation(8.3)) and Ratio Model (equation (8.4)).

$$s\left(c_1, c_2\right) = \theta * f(C_1 \cap C_2) + \alpha * f(C_1 - C_2) + \beta * f(C_2 - C_1) \quad (8.3)$$

$$s\left(c_1, c_2\right) = \frac{f(C_1 \cap C_2)}{f(C_1 \cap C_2) + \alpha * f(C_1 - C_2) + \beta * f(C_2 - C_1)} \quad (8.4)$$

where $f()$ is a function that reflects either the salience or prominence of a set of features (Pirró and Euzenat, 2010), or simply determines the cardinality of the set (Schwering, 2006). $C_1 \cap C_2$ is the set of common features of two concepts, while $C_1 - C_2$ ($C_2 - C_1$) is the set of distinct features that belong to concept C_1 (C_2) but do not belong to concept C_2 (C_1). θ, α and β indicate the importance of different components in the similarity estimation. The sum of α and β should equal 1.

For the contrast model, the similarity value is not bounded between 0 and 1, which makes interpretation difficult.

8.2.3 Network Model

The basis of the network model is graph theory. Concepts are connected through appropriate relations in the semantic network, such as is-a relation. Concepts are represented by nodes, while the relations between them are represented by edges. The similarity in the network model is calculated by graph-theoretic algorithms. A simple algorithm for semantic distance may be the shortest path model (equation (8.5)).

$$d\left(c_1, c_2\right) = \min length(P_{c_1,c_2}) \quad (8.5)$$

where P_{c_1,c_2} is the length between C_1 and C_2.

For a semantic network with only is-a relations, similarity values in the network model are highly sensitive to the predefined hierarchy network (Rodríguez, 2000).

8.2.4 Transformational Model

The similarity in the transformation model is equal to the number of transformations to make one concept identical to the other concept (Hahn et al., 2009). When needed transformations increase, the similarity decreases monotonically. Transformational operations may be counted based on a coding language (Hodgetts et al., 2009). According to Kolmogorov complexity theory, the number of transformations is the smallest number of operations with which a computer program transforms one concept into another (Goldstone, 1999).

Indeed, the calculation of the semantic similarity from the transformational model needs some comparisons (Grimm et al., 2012). Transformational operations are only conducted on distinct components, while identical components are not considered.

8.3 Modelling Semantic Similarity

8.3.1 A New Semantic Similarity Measurement

Based on the Dynamic Semantic Reference System (DSRS), a new semantic similarity model is proposed as shown in equation (8.6).

$$Sim\left(C_1, C_2\right) = \frac{1}{\left|C_1\right|^{\mathrm{L}}} \sum_{i=1}^{\left|C_1\right|^L} \underset{j=1}{\overset{\left|C_2\right|^L}{Max}} \left(\frac{\sum \left(1 - dist^N \left(p\left(C_{1i}^L, C_{2j}^L\right)\right)\right)}{\left|C_{1i}^L\right|^R} \right) \quad (8.6)$$

where C_1 and C_2 are two concepts consisting of leaf concepts C_{1i}^L and C_{2j}^L respectively. $Sim\left(C_1, C_2\right)$ is the similarity between C_1 and C_2 which measures how similar C_1 is to C_2. $\left|\ \right|^{\mathrm{L}}$ is the width of a concept which returns the number of leaf concepts covered by the concept, and $\left|\ \right|^R$ is the depth of a concept which returns the number of attributes used to characterize the concept. $p\left(C_{1i}^L, C_{2j}^L\right)$ is a function to align property pairs according to C_{1i}^L. $dist^N\left(\ \right)$ is a normalized distance.

8.3.2 Characteristics of Semantic Similarity Models

8.3.2.1 Similarity and Dissimilarity

The semantic similarity is obtained directly or converted and normalized from the semantic dissimilarity (distance) indirectly.

The feature model is a typical representative of a direct similarity model. Common and distinct features are combined to evaluate semantic similarities. The more common the features and the less distinct they are, the higher the overall semantic similarity is.

Different kinds of distance, including spatial distance, path length, and transformational complexity, are considered in the geometric model, the network model and the transformational model. The proposed measurement in this chapter is also based on dissimilarity. There is a negative relationship between similarity and distance. The shorter the distance is, the higher the similarity will be.

8.3.2.2 Properties and Relations

Properties describe the characteristics of a concept, while relations describe connections between concepts.

Properties are considered in all models except for the pure network model. The name and the range of the property are explicitly represented in the geometric model. Property values are arranged along the dimension in some rank. In the feature model, property values are simply listed in the feature set. In the transformational model, properties are aligned into two types of properties: matched property and unmatched property. Normally, fewer operations are needed to transform between the matched properties than the unmatched properties. In the proposed measurement, properties are divided into two types of components: contrast components and not-contrast components (see Chapter 5 for details).

Relations are considered in the network model and the transformational model. Relations in the network model are usually hierarchical or associative, while relations are aligned for transformation in the transformational model.

8.3.2.3 Metric and Non-Metric

The metric character is the most important assumption of the geometric model. In metric space, the semantic distance meets the three metric axioms (Gärdenfors, 2000, p.17): minimality (equation (8.7)), symmetry (equation (8.8)), and triangle inequality (equation (8.9)).

$$d\left(c_1, c_2\right) = 0 \Rightarrow c_1 = c_2 \ and \ d\left(c_1, c_2\right) \geq 0 \qquad (8.7)$$

$$d\left(c_1, c_2\right) = d\left(c_2, c_1\right) \qquad (8.8)$$

$$d\left(c_1, c_2\right) + d\left(c_2, c_3\right) \geq d\left(c_1, c_3\right) \qquad (8.9)$$

The axiom of minimality indicates that if the distance between two concepts equals 0, then the concepts are identical. The axiom of symmetry indicates that the order of concepts does not affect the magnitude of distance. The axiom of triangle inequality indicates that the direct distance from one concept to the other is not larger than the sum of the distance from any one of them to an intermediate concept.

The geometric model is set up on the metric space, which has been criticized for disagreement with the human cognitive process. Hence, a feature model that discards the metric character is designed. Network models hold the axiom of minimality and triangle inequality. Undirected network models remain symmetric, and directed network models are asymmetric. The transformational model holds the metric character except for the symmetry.

The metric characteristics of the mentioned models are expressed in Table 8.1. It is clear that the main discrepancy is concentrated on the symmetric axiom, which is preserved by two models and disagreed with by the other four.

Many examples and explanations have been proposed to prove the existence of symmetric similarity (Rada et al., 1989) and asymmetric similarity (Tversky, 1977). For example, North Korea is more similar to China than China to North Korea from human subject tests (Tversky, 1977). Asymmetric similarity is supported in this thesis. Normally, the similarity between a subordinate class with a superordinate class is larger than that between a superordinate class with a subordinate class. In another word, a more specific class is much similar to a general class (Rodr ǵuez and Egenhofer, 2004). Asymmetry can be caused by difference between prototype and variant, expertise, granularity difference, and context variation (Schwering, 2006).

Table 8.1 Metric Characteristics of Semantic Similarity Models

	SYMMETRIC	MINIMALITY	TRIANGLE INEQUALITY
GEOMETRIC MODEL	Yes	Yes	Yes
FEATURE MODEL	No	No	No
NETWORK MODEL (UNDIRECTED)	Yes	Yes	Yes
NETWORK MODEL (DIRECTED)	No	Yes	Yes
TRANSFORMATIONAL MODEL	No	Yes	Yes
Proposed New Model	No	No	No

8.4 Case Study

8.4.1 Experimental Data

We take the NLCD2001 classification system (NLCD2001-CS) as an example to evaluate the performance of the proposed semantic similarity measurement. DSRS has been established in Chapter 5. Constrained definitions are employed. In total, there are 21 leaf concepts and 13 properties.

8.4.2 Normalized Distance

There are many distance functions, such as the Minkowsky (Batchelor, 1978), Mahalanobis (Nadler and Smith, 1993), hyperrectangle distance functions (Salzberg, 1991), and others. Many of them work well for numerical attributes but are not well-suited to nominal attributes (Wilson and Martinez, 1997). In land class definitions, there are many nominal attributes. The normalized distance between any attributes is defined as:

$$dist^N(x,y) = \begin{cases} 1, \ if \quad condition1, \text{otherwise} \\ 0, \ if \quad condition2, \text{otherwise} \\ \frac{1}{|p|}, \ if \quad \text{condition 3}, \quad otherwise \\ \frac{abs(\text{x}-\text{y})}{|p|} \end{cases} \qquad (8.10)$$

where x and y are two attributes, p is the property which contains x and y, $|p|$ is the number of attributes in p, and $abs(\text{x}-\text{y})$ is the absolute value of $(\text{x}-\text{y})$. $|p|$ equals the number of attributes if p is a nominal property. $|p|$ equals to the maximal ordinal number if p is a ordinal property. $|p|$ equals the range (maximum-minimum) of p if p is an interval property or a ratio property. Condition1 refers that x and y belong to different properties. Condition2 refers that x and y are identical. And condition3 refers that x and y are nominal attributes. Hence, the distance between attributes of a nominal property is identical.

Normalized distance for every property is illustrated in Table 8.2. All properties are nominal properties except for ‘p4’. Property ‘p4’ is treated as an ordinal property. Hence the distance between ‘>5m’ and ‘<6m’ equals 2/3 and others are equal to 1/3.

Table 8.2 Normalized Distance in Every Property

	Attributes of Property	$\lvert p \rvert$	$dist^N(\ )$
p1	{water, vegetation, soil, constructed material}	4	1/4
p2	{open, ice/snow}	2	1/2
p3	{woody, herbaceous}	2	1/2
p4	{>6m, 5m-6m, <5m}	3	1/3 or 2/3
p5	{deciduous, evergreen}	2	1/2
p6	{natural or semi-natural, planted or managed}	2	1/2
p7	{not for man use, for man use}	2	1/2
p8	{saturated with water, not saturated with water}	2	1/2
p9	{for residence, not for residence}	2	1/2
p10	{extensive mining, not extensive mining}	2	1/2
p11	{for row crops, for small grains}	2	1/2
p12	{fallow, not fallow}	2	1/2
p13	{for urban/recreational, not for urban/recreational}	2	1/2

8.4.3 Leaf Concept Similarity

The similarity between leaf concepts is degraded from equation (8.6) to equation (8.11) as they are covered by themselves.

$$\mathrm{sim}\left(C_1, C_2\right) = \frac{\sum\left(1 - dist^N\left(p\left(C_{1i}^L, C_{2j}^L\right)\right)\right)}{\left|C_{1i}^L\right|^R} \qquad (8.11)$$

The results are presented in Table 8.3. The diagonal is self-similarity which is always equal to 1. Referencing Figure 5.4, leaf concepts with the same superordinate concept are filled with light grey. It is found that similarities within a superordinate class are not always the largest. For example, $sim\left(CL11, CL13\right)$ is larger than $sim\left(CL11, CL12\right)$. One reason is that although leaf concepts are unique, intermediate nodes are not. If the property ‘p5’ is applied prior to ‘p4’, then CL11 and CL13 are inherited from a common superordinate concept as shown in Figure 8.1. Another reason is that the differences between the applied properties, which can be found in Table 8.2, cause the difference in the normalized distance. Therefore, it seems that the similarity of concepts within a superordinate concept may not be the largest, but it is always relatively large. Other bold values in Table 8.2 which look unbelievable are also caused by a combination of the two reasons mentioned above.

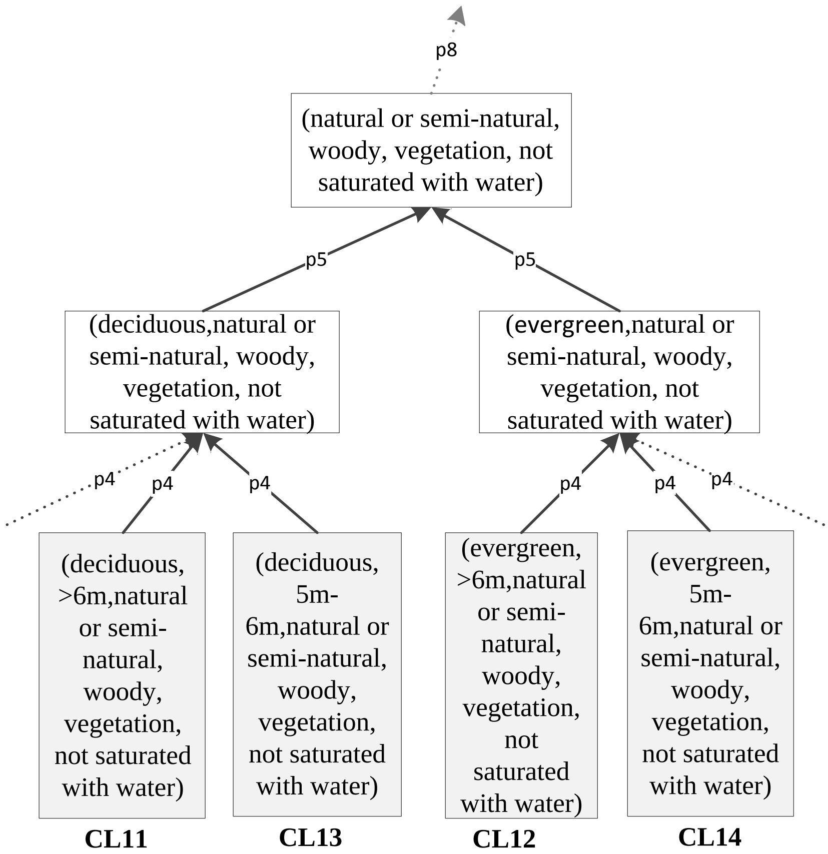


Figure 8.1 An Alternative Tree Structure for a Part of Figure 5.4

8.4.4 Concept-level Similarity

Characteristic covering concepts of NLCD2001 classes are depicted in Table 6.1 and Table 6.2. The results of similarities between them are presented in Table 8.4. There is no doubt that the self-similarity along the diagonal is always equal to 1. There are some identical classes which have been highlighted in bold and underlined, such as ‘30’ and ‘31’. As only one class inherits from class ‘30’ and

‘70’, respectively, it is obvious and reasonable that ‘31’ should equal ‘30’, and that ‘71’ should equal ‘70’. The subordinate classes of ‘20’ are defined with the same properties as ‘20’. Hence, they are characterized by the same characteristic coverings.

Not all classes are entirely similar to their superordinate class, including classes ‘41’, ‘42’, ‘43’, and ‘82’. The reason is that they are characterized by some coverings that are beyond their superordinate classes, which can be directly found in Table 6.1 and Table 6.2. Surprisingly, class ‘40’ is a subordinate class of ‘43’, because class ‘43’ owns more characteristic coverings than ‘40’. Similarities between classes owning the same superordinate class are normally largest.

In level-II, for classes beyond the superordinate class, they are equally similar to classes ‘11’, ‘12’, ‘21’, ‘22’, ‘23’, ‘24’, and ‘31’. Classes ‘41’, ‘42’, and ‘43’ are more similar to ‘52’, and vice versa. Class ‘71’ is more similar to ‘41’, ‘42’, ‘43’, and ‘92’. Class ‘81’ and ‘82’ are more similar to ‘71’ and ‘92’. Class ‘91’ is more similar to ‘41’, ‘42’, ‘43’, and ‘52’. Class ‘92’ is more similar to ‘71’.

8.4.5 Category-level Similarity

Category-level similarity is only implemented on level-II classes. Although equation (8.6) is still applied to calculate the category-level similarity, the meaning of elements is different. C_1 and C_2 represent two categories consisting of leaf concepts C_{1i}^{L} and $C_{2,j}^{L}$ respectively. $| \; |^{L}$ returns the number of possible combinations of the category. $dist^{N}\left(\; \right)$ is a normalized distance weighted by difference of area percentage. Other elements keep same meanings as concept-level similarity.

The results are presented in Table 8.5. In contrast to values in Table 8.4, they are more varied and more surprising. There are no identical classes. Similarities between classes with the same superordinate class are largest for classes ‘23’, ‘24’, ‘41’, ‘42’, and ‘92’. Classes ‘11’, ‘21’, ‘22’, ‘43’, ‘81’, and ‘82’ also perform well in that fewer than 3 classes exceed inner-class similarity. Class ‘12’ and ‘92’ perform poorly in that more than 5 classes exceed inner-class similarity. A commonality of the first group is that the area percentage of characteristic coverings is very large while the area percentage of characteristic coverings for the third group is very small which leaves enough opportunity to mix with other coverings.

Table 8.3 Leaf Concept Similarity

	CL1	CL2	CL3	CL4	CL5	CL6	CL7	CL8	CL9	CL10	CL11	CL12	CL13	CL14	CL15	CL16	CL17	CL18	CL19	CL20	CL21
CL1	1.00	0.75	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
CL2	0.75	1.00	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
CL3	0.38	0.38	1.00	0.75	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
CL4	0.38	0.38	0.75	1.00	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
CL5	0.38	0.38	0.38	0.38	1.00	0.75	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
CL6	0.38	0.38	0.38	0.38	0.75	1.00	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
CL7	0.19	0.19	0.19	0.19	0.19	0.19	1.00	0.88	0.63	0.63	0.63	0.63	0.63	0.63	0.50	0.50	**0.88**	**0.88**	**0.88**	0.75	0.75
CL8	0.19	0.19	0.19	0.19	0.19	0.19	0.88	1.00	0.63	0.63	0.63	0.63	0.63	0.63	0.50	0.50	0.75	0.75	0.75	**0.88**	**0.88**
CL9	0.19	0.19	0.19	0.19	0.19	0.19	0.63	0.63	1.00	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	0.75	**0.88**	0.50	0.50	0.50	0.50	0.50
CL10	0.15	0.15	0.15	0.15	0.15	0.15	0.50	0.50	0.70	1.00	0.87	0.87	**0.93**	**0.93**	0.70	0.60	0.40	0.40	0.40	0.40	0.40
CL11	0.13	0.13	0.13	0.13	0.13	0.13	0.42	0.42	0.58	0.72	1.00	0.92	**0.94**	0.86	0.58	0.50	0.33	0.33	0.33	0.33	0.33
CL12	0.13	0.13	0.13	0.13	0.13	0.13	0.42	0.42	0.58	0.72	0.92	1.00	0.86	**0.94**	0.58	0.50	0.33	0.33	0.33	0.33	0.33
CL13	0.13	0.13	0.13	0.13	0.13	0.13	0.42	0.42	0.58	0.78	**0.94**	0.86	1.00	0.92	0.58	0.50	0.33	0.33	0.33	0.33	0.33
CL14	0.13	0.13	0.13	0.13	0.13	0.13	0.42	0.42	0.58	0.78	0.86	**0.94**	0.92	1.00	0.58	0.50	0.33	0.33	0.33	0.33	0.33
CL15	0.19	0.19	0.19	0.19	0.19	0.19	0.50	0.50	0.75	**0.88**	**0.88**	**0.88**	**0.88**	**0.88**	1.00	0.88	0.63	0.63	0.63	0.63	0.63
CL16	0.19	0.19	0.19	0.19	0.19	0.19	0.50	0.50	**0.88**	0.75	0.75	0.75	0.75	0.75	0.88	1.00	0.63	0.63	0.63	0.63	0.63
CL17	0.15	0.15	0.15	0.15	0.15	0.15	0.70	0.60	0.40	0.40	0.40	0.40	0.40	0.40	0.50	0.50	1.00	**0.90**	**0.90**	0.70	0.70
CL18	0.13	0.13	0.13	0.13	0.13	0.13	0.58	0.50	0.33	0.33	0.33	0.33	0.33	0.33	0.42	0.42	0.75	1.00	0.92	0.58	0.58
CL19	0.13	0.13	0.13	0.13	0.13	0.13	0.58	0.50	0.33	0.33	0.33	0.33	0.33	0.33	0.42	0.42	0.75	0.92	1.00	0.58	0.58
CL20	0.15	0.15	0.15	0.15	0.15	0.15	0.60	0.70	0.40	0.40	0.40	0.40	0.40	0.40	0.50	0.50	0.70	0.70	0.70	1.00	0.90
CL21	0.15	0.15	0.15	0.15	0.15	0.15	0.60	0.70	0.40	0.40	0.40	0.40	0.40	0.40	0.50	0.50	0.70	0.70	0.70	0.90	1.00

Table 8.4 Concept-level Similarity between NLCD2001 Classes

	10	11	12	20	21	22	23	24	30	31	40	41	42	43	50	52	70	71	80	81	82	9	91	92
10	1.00	**0.88**	**0.88**	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
11	**1.00**	1.00	*0.75*	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
12	**1.00**	*0.75*	1.00	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
20	0.38	0.38	0.38	1.00	**1.00**	**1.00**	**1.00**	**1.00**	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
21	0.38	0.38	0.38	**1.00**	1.00	**1.00**	**1.00**	**1.00**	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
22	0.38	0.38	0.38	**1.00**	**1.00**	1.00	**1.00**	**1.00**	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
23	0.38	0.38	0.38	**1.00**	**1.00**	**1.00**	1.00	**1.00**	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
24	0.38	0.38	0.38	**1.00**	**1.00**	**1.00**	**1.00**	1.00	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
30	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	1.00	**1.00**	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
31	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	**1.00**	1.00	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
40	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	1.00	0.96	0.96	**1.00**	0.94	0.72	0.58	0.58	0.33	0.33	0.42	0.58	0.58	0.50
41	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.97	1.00	0.92	**1.00**	0.97	0.75	0.58	0.58	0.33	0.33	0.42	0.58	0.58	0.50
42	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.97	0.92	1.00	**1.00**	0.97	0.75	0.58	0.58	0.33	0.33	0.42	0.58	0.58	0.50
43	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.97	0.96	0.96	1.00	0.97	0.75	0.58	0.58	0.33	0.33	0.42	0.58	0.58	0.50
50	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.92	0.95	0.95	**0.98**	1.00	0.85	0.62	0.62	0.36	0.36	0.44	0.62	0.62	0.53
52	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.87	0.93	0.93	0.93	**1.00**	1.00	0.70	0.70	0.40	0.40	0.50	0.70	0.70	0.60
70	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.88	0.88	0.88	0.88	0.88	0.88	1.00	**1.00**	0.63	0.63	0.63	0.88	0.75	0.88
71	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.88	0.88	0.88	0.88	0.88	0.88	**1.00**	1.00	0.63	0.63	0.63	0.88	0.75	0.88
80	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.37	0.37	0.37	0.37	0.37	0.37	0.47	0.47	1.00	0.75	**0.88**	0.47	0.37	0.47
81	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.40	0.40	0.40	0.40	0.40	0.40	0.50	0.50	**1.00**	1.00	0.70	0.50	0.40	0.50
82	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.42	0.42	0.42	0.42	0.42	0.42	0.46	0.46	**0.97**	0.65	1.00	0.49	0.42	0.46
9	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.56	0.56	0.63	1.00	**0.94**	**0.94**
91	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.88	0.88	0.88	0.88	0.88	0.88	0.75	0.75	0.50	0.50	0.63	**1.00**	1.00	0.88
92	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.75	0.75	0.75	0.75	0.75	0.75	0.88	0.88	0.63	0.63	0.63	**1.00**	0.88	1.00

Table 8.5 Category-level Similarity

	11	12	21	22	23	24	31	41	42	43	52	71	81	82	91	92
11	1	0.7958	0.4507	0.3886	0.3886	0.4570	0.4637	0.7833	0.7833	**0.7985**	**0.9527**	0.4761	0.4570	0.4570	0.9513	**0.9527**
12	0.7546	1	0.6508	0.6487	0.6478	0.6485	0.6728	**0.9246**	**0.9246**	**0.9424**	**0.9729**	0.7328	0.6501	0.6501	**0.9731**	**0.9199**
21	0.6804	0.7018	1	0.9614	0.9023	0.8189	0.6800	0.7027	0.7027	0.7018	0.7027	0.6874	**0.9631**	**0.9694**	0.7027	0.6966
22	0.6867	0.6941	0.9333	1	**0.9478**	0.8819	0.6867	0.6941	0.6941	0.6941	0.6941	0.6867	**0.9128**	**0.9128**	0.6941	0.6905
23	0.6200	0.6209	0.8200	0.9200	1	**0.9364**	0.6200	0.6209	0.6209	0.6209	0.6209	0.6200	0.8063	0.8063	0.6209	0.6200
24	0.5235	0.5235	0.6219	0.7615	**0.8810**	1	0.5235	0.5235	0.5235	0.4924	0.5235	0.5235	0.6205	0.6205	0.5235	0.5235
31	0.4947	0.7467	0.4160	0.3704	0.3704	0.4413	1	0.7467	0.7467	0.7467	**0.8933**	0.4947	0.4413	0.4413	**0.8933**	**0.8933**
41	0.7372	0.9169	0.6408	0.6372	0.6299	0.6254	0.6547	1	0.9578	**0.9639**	0.9574	0.6861	0.6338	0.6338	0.9566	0.7443
42	0.7372	0.9169	0.6408	0.6372	0.6299	0.6254	0.6547	0.9578	1	**0.9639**	0.9574	0.6861	0.6338	0.6338	0.9566	0.7443
43	0.8182	0.9456	0.7156	0.7156	0.7141	0.7127	0.7355	0.9542	0.9542	1	**0.9720**	0.7453	0.7140	0.7140	**0.9736**	0.7908
52	0.7953	0.9380	0.7130	0.7119	0.7065	0.7037	0.7456	0.9438	0.9438	0.9607	1	0.7604	0.7078	0.7078	**0.9755**	0.8383
71	0.4160	0.7600	0.6320	0.5262	0.4203	0.3273	0.3100	**0.8992**	**0.8992**	0.7600	**0.8992**	1	0.6572	0.6572	0.8220	**0.8992**
81	0.6304	0.6558	**0.9804**	0.9304	0.8573	0.7589	0.6304	0.6565	0.6565	0.6558	0.6565	0.6473	1	0.9802	0.6565	0.6620
82	0.6253	0.6463	**0.9777**	0.9271	0.8499	0.7531	0.6253	0.6470	0.6470	0.6463	0.6470	0.6413	0.9594	1	0.6470	0.6526
91	0.7928	**0.9364**	0.7206	0.7191	0.7127	0.7041	0.7456	**0.9427**	**0.9427**	**0.9601**	**0.9770**	0.7621	0.7118	0.7118	1	0.8769
92	0.6950	0.9093	0.6342	0.6232	0.6048	0.5893	0.6974	0.8924	0.8924	0.8829	0.9456	0.7297	0.6360	0.6360	**0.9614**	1

8.5 Discussion and Conclusion

This chapter proposes a new measurement to quantify semantic similarity between classes in a classification system based on the Dynamic Semantic Reference System. NLCD2001 CS classes are employed for demonstration. An interesting finding is that semantic similarity between classes inherited from a common nominal parent may not be the largest, which betrays a common assumption in classification. For example, in remote sensing, it is assumed that 'pixels within classes are spectrally more similar to one another than they are to pixels in other classes' (http://www.learnremotesensing.org/modules/image_classification/index.php?target=image_class accessed on 27 August, 2014). One reason is that a class may have different parents.

CHAPTER 9 SEMANTIC INTEROPERABILITY

9.1 Introduction

An increasing number of different classification systems have been established to obtain and regularly update land cover and land use databases at global, continental, national, and municipal levels. Inconsistent definitions of classes lead to difficulties in data share and data interoperability. For applications where data need to be translated from other systems or where more than one system is used, the relationships between these systems need to be explicitly correlated (Scholes et al., 2012).

This chapter proposes a new method to bridge two classification systems based on DSRS. Interoperability by standardization and harmonization is introduced in Section 9.2. Section 9.3 presents the method to bridge different classification systems. In Section 9.4, NLCD1992 CS and NLCD2001 CS are used as examples for demonstration. Discussion and analysis are provided in Section 9.5. Section 9.6 concludes this research.

9.2 Standardization and Harmonization

Solutions to reach consistent semantic contents of classes are divided into two contrary directions: standardization and harmonization. Standardization creates uniform definitions and methods which should be mandatory in an international, national, or industrial domain. The application of standards makes it possible to consistently meet predefined purposes and conduct activities with a well-defined practice. Standardization reduces the risk of failure. It is a compromise among different producers and users and follows a top-down process. Divergences are removed and commonalities are held and emphasized by standardization. In the geospatial domain, OGC (Open Geospatial Consortium) and ISO/TC211 (the International Organization for Standardization, Technical Committee 211) are the two main organizations working on international standards. ISO 19144, which is the result of FAO LCCS, is the first international standard to address classification systems. It is divided into two parts: the ISO 19144-1 classification system structure and ISO 19144-2 Land Cover Meta Language (LCML).

Harmonization involves how to make different standards fit together without making them uniform. The ultimate goal is to allow direct comparison between different systems (Herold et al., 2006b). Hence, harmonization is not removing differences, but making an explicit distinction between commonalities with differences. Translations and comparative analysis reveal inconsistencies between different classification systems. These harmonizing results improve our understandings of which part of classes can be directly compared and where incompatibility occurs. Harmonization is helpful for the revision of current projects and the designation of future projects.

Standardization is not static; it changes when the conceptualization of the underlying phenomenon changes (Comber et al., 2008a). Changes in policy, purpose, method, and context encourage the use of different standards. A more acceptable solution is to maintain existing systems but to build an interchanging platform to combine separate systems (wyatt et al., 1997, p.6).

9.3 Methodology

Interoperability is achieved in three main steps: decompose, rebuild, and compare. Definitions of classes cannot be compared directly. The first prominent issue is to constitute a reference platform to decompose original classes into comparable units which are either identical or totally different. The DSRS introduced in Chapter 5 is such a system. All classes can be represented optimally. In other words, a merge of any leaf concepts will cause problems and a further breakup of leaf concepts is meaningless in the DSRS. The reference platform, DSRS, is the foundation for interoperability in this research.

In the second step, original classes are rebuilt by leaf concepts. Although a land class is a mixture of different coverings, it is characterized by some characteristic coverings. For example, agricultural land is characterized by crops and it is what it is even if some trees may be included. Hence, to some extent, characteristic coverings represent a class. It is meaningful to match between characteristic coverings also. Indeed, matches in routine communication and most researches are more between characteristic coverings than mixtures, because characteristic coverings are relatively easy to understand in contrast to a black box mixture.

In the final step, direct comparison is allowed to be implemented between classification systems. The overlaps between classes between different systems are calculated. The output of interoperability in this research is a quantitative numerical table rather than a qualitative matching table. Overlap measurements introduced in Chapter 6 are employed, as shown in equation (6.1). But C_1 and C_2 are classes of different classification systems.

The whole process is illustrated in Figure 9.1.

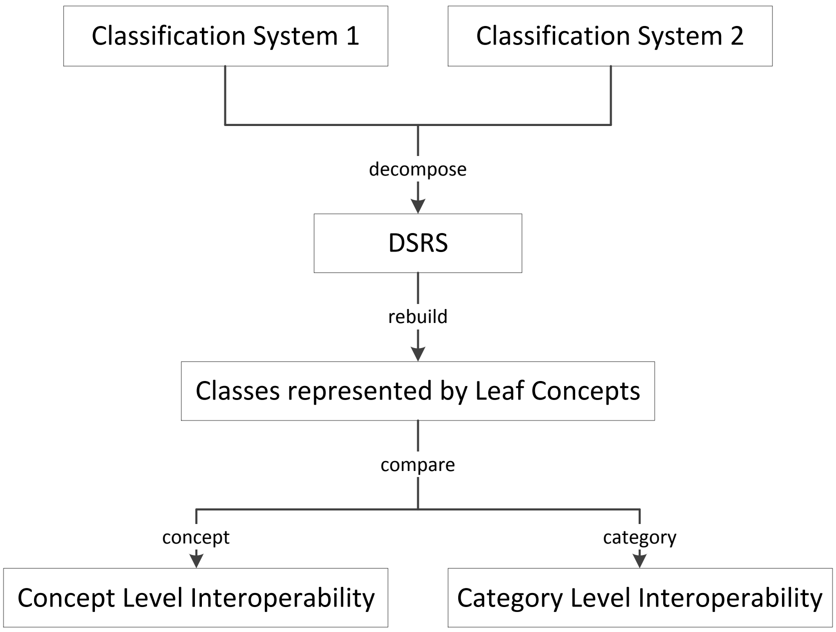


Figure 9.1 Workflow of Semantic Interoperability between Different Classification Systems

9.4 Case Study and Analysis

9.4.1 Experimental Data

Table 9.1 Constrained Level-II Classes in Artificial Zone (NLCD1992 CS)

Code	Characteristic Covering	Condition
20	CL1, CL2	AP(CL1+CL2)>=30%
21	CL1	(1) 30%<=AP(CL1)<80%; (2) AP(CL1+CL2)>=30%
22	CL1	(1) AP(CL1)>=80%; (2) AP(CL1+CL2)>=30%
23	CL2	(1) AP(CL2)>=80%; (2) AP(CL1+CL2)>=30%
60	CL7,CL8	AP(CL7+CL8)>=25%
61	CL7,CL8	
80	CL17,CL18,CL19,CL20,CL21	AP(CL17+CL18+CL19+ CL20+CL21)>=75%
81	CL21	
82	CL18	
83	CL19	
84	CL17	
85	CL20	

NLCD1992 CS and NLCD2001 CS are used as examples for demonstrations. Definitions of the classes are described in Chapter 5. DSRS has been established in Chapter 5. NLCD 2001 CS has been rebuilt in Table 6.1, Table 6.2, Table 6.9, and Table 6.10 in Chapter 6. NLCD1992 CS is rebuilt in Table 9.1 and Table 9.2. The whole zone is still divided into artificial zones and natural zones. Two types of matches will be applied: concept-level (i.e., characteristic covering) and category-level (i.e., mixture). The level-II classes are constrained by the corresponding parent level-I class definitions, because characteristics of a superordinate class should be inherited by its descendants in a hierarchical structure. Grey highlighted conditions are inherited from the corresponding superordinate class as shown in Table 9.1 and Table 9.2.

Table 9.2 Constrained Level-II Classes in Natural Zone (NLCD1992 CS)

Code	Characteristic Covering	Condition
10	CL3, CL4	
11	CL3	AP(CL3)>=75
12	CL4	
30	CL5,CL6	
31	CL6	
32	CL5	
33	CL9,CL10,CL 11,CL12,CL13 ,CL14,CL15,C L16	0<(CL9+CL10+CL11+CL12+CL13+CL14+CL15+CL 16)<25%
40	CL11,CL12	AP(CL11+CL12)>=25%
41	CL11	(1) AP(CL11)/AP(CL11+CL12)>=75%, (2) CL11>0; and (3) AP(CL11+CL12)>=25%
42	CL12	(1) AP((CL12))/AP((CL11+CL12))>=75%, (2) CL12>0; and (3) AP(CL11+CL12)>=25%
43	CL11, CL12	(1) AP(CL11)/AP(CL11+CL12)<75% , (2) AP(CL12)/AP(CL11+CL12)<75%, (3) CL11+CL12>0, and (4) AP(CL11+CL12)>=25%
50	CL10,CL13,C L14	
51	CL10,CL13,C L14	(1)AP(CL10+CL13+CL14)>=25%,and AP(CL11+CL12)<25%; and (2)AP(CL10+CL13+CL14)<25% and AP(CL10+CL13+CL14)=Max(AP(CL3+CL4), AP(CL5+CL6), AP(CL10+CL13+CL14), AP(CL11+CL12), AP(CL15),AP(CL9+CL16))
70	CL15	AP(CL15)>=75%
71	CL15	(1) AP(CL15)> AP(CL9+CL10+CL11+CL12+CL13+CL14); (2) AP(CL15)>=75%
9	CL9,CL16	
91	CL9	AP(CL9>=25%
92	CL16	AP(CL16)>=75%

All classes are characterized by some characteristic coverings but not all classes are defined as a mixture. Hence, all classes are involved in the concept-level interoperable process and only some level-II classes are concerned in the category-level interoperable process. In NLCD1992 CS, there are 3 level-I classes and 9 level-II classes in the artificial zone and 6 level-I classes and 12 level-II classes in the natural zone. 3 level-II classes in the artificial zone and 9 level-II classes in the natural zone are defined as a mixture. In NLCD2001 CS, there are 2 level-I classes and 6 level-II classes in the artificial zone and 6 level-I classes and 10 level-II classes in the natural zone. All level-II classes are defined as a mixture. Furthermore, the translation is asymmetric, and the results have been presented separately.

9.4.2 Concept-level Interoperability (NLCD1992 CS to NLCD2001 CS)

9.4.2.1 Artificial Zone

The results of concept-level interoperability from NLCD1992 CS to NLCD2001 CS in the artificial zone are presented in Table 9.3. The value 0 means the class cannot be translated between the class pair, 1 means the class is completely translated into the target class, and other values mean the class can be partially translated into the target class. The larger the value is, the more a class is translated.

In level-I, class ‘60’ in NLCD1992 CS (92_60) fails to match any class in NLCD2001 CS. It is obvious by directly comparing level-I class definitions. There exists a many-to-many relationship between class ‘20’ (including its descendants) from these two CSs correspondingly. Half of class ‘92_61’ is converted to ‘01_82’, and the rest do not match any. Every level-II class under ‘92_80’ is completely converted to a certain subordinate class of ’01-80’, except that there does not exist a match for ‘92_85’. However, cross level matches exist for ‘92_60’ and ‘92_85’. Half of ‘92_60’ can be converted to ‘01_82’, and ‘92_85’ can be entirely converted to ‘01_80’. In total, only ‘92_60’ and ‘92_61’ cannot be completely represented by NLCD2001 CS.

Table 9.3 Concept-level Interoperability (92-01) in Artificial Zone

	01_20	**01_21**	**01_22**	**01_23**	**01_24**	**01_80**	**01_81**	**01_82**
92_20	1	1	1	1	1	0	0	0
92_21	1	1	1	1	1	0	0	0
92_22	1	1	1	1	1	0	0	0
92_23	1	1	1	1	1	0	0	0
92_60	0	0	0	0	0	0	0	0.5
92_61	0	0	0	0	0	0	0	0.5
92_80	0	0	0	0	0	1	0.2	0.6
92_81	0	0	0	0	0	1	1	0
92_82	0	0	0	0	0	1	0	1
92_83	0	0	0	0	0	1	0	1
92_84	0	0	0	0	0	1	0	1
92_85	0	0	0	0	0	1	0	0

9.4.2.2 Natural Zone

The results of concept-level interoperability from NLCD1992 CS to NLCD2001 CS in the natural zone are presented in Table 9.4. Level-I classes in NLCD1992 CS are completely and uniquely represented by level-I classes in NLCD2001 CS. A complete and unique translation also occurs on ‘92_11’, ‘92_12’, ‘92_31’, ‘92_32’, ‘92_71’, ‘92_91’, and ‘92_92’ in level-II classes. Class ‘92_33’ has the maximal number of corresponding classes, which is in accordance with its name ‘Transitional’. Class ‘92_41’ can be represented by ‘01_41’ and ‘01_43’, while ‘92_42’ can be represented by ‘01_42’ and ‘01_43’. Class ‘92_43’ can be represented by ‘01_43’, and a part of it can also be represented by ‘01_42’ and ‘01_43’. Class ‘92_51’ can only partially be represented by ‘01_52’ in NLCD2001 CS.

Table 9.4 Concept-level Interoperability (92-01) in Natural Zone

	01_10	01_11	01_12	01_30	01_31	01_40	01_41	01_42	01_43	01_50	01_52	01_70	01_71	01_9	01_91	01_92
92_10	1	0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
92_11	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
92_12	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
92_30	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
92_31	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
92_32	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
92_33	0	0	0	0	0	0.25	0.25	0.25	0.5	0.38	0.13	0.13	0.125	0.25	0.125	0.125
92_40	0	0	0	0	0	1	0.5	0.5	1	0	0	0	0	0	0	0
92_41	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
92_42	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0
92_43	0	0	0	0	0	1	0.5	0.5	1	0	0	0	0	0	0	0
92_50	0	0	0	0	0	0	0	0	0	1	0.33	0	0	0	0	0
92_51	0	0	0	0	0	0	0	0	0	1	0.33	0	0	0	0	0
92_70	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
92_71	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
92_9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.5	0.5
92_91	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
92_92	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

9.4.3 Concept-level Interoperability (NLCD2001 CS to NLCD1992 CS)

9.4.3.1 Artificial Zone

The results of concept-level interoperability from NLCD2001 CS to NLCD1992 CS in the artificial zone are presented in Table 9.5. Level-I classes completely match. Only ‘01_81’ is entirely converted to one class ‘92_81’. A part of level-II classes of ‘01_20’ are converted to level-II classes of ‘92_20’ respectively. A part of ‘01_82’ is even represented by class ‘92_61’.

Table 9.5 Concept-level Interoperability (01-92) in Artificial Zone

	92_20	**92_21**	**92_22**	**92_23**	**92_60**	**92_61**	**92_80**	**92_81**	**92_82**	**92_83**	**92_84**	**92_85**
01_20	1	0.5	0.5	0.5	0	0	0	0	0	0	0	0
01_21	1	0.5	0.5	0.5	0	0	0	0	0	0	0	0
01_22	1	0.5	0.5	0.5	0	0	0	0	0	0	0	0
01_23	1	0.5	0.5	0.5	0	0	0	0	0	0	0	0
01_24	1	0.5	0.5	0.5	0	0	0	0	0	0	0	0
01_80	0	0	0	0	0	0	1	0.2	0.2	0.2	0.2	0.2
01_81	0	0	0	0	0	0	1	1	0	0	0	0
01_82	0	0	0	0	0.25	0.25	0.75	0	0.25	0.25	0.25	0

Comparing Table 9.5 with Table 9.3, identical classes include ‘92_20’ and ‘01_20’, ‘92_80’ and ‘01_80’, and ‘92_81’ and ‘01_81’. There are subordinate and superordinate relationships between ‘92_21’, ‘92_22’, ‘92_23’ and ‘01_21’, ‘01_22’, ‘01_23’, ‘01_24’, respectively, and also between ‘92_82’, ‘92_83’, ‘92_84’ and ‘01_82’, respectively.

9.4.3.2 Natural Zone

The results of concept-level interoperability from NLCD2001 CS to NLCD1992 CS in the natural zone are presented in Table 9.6. In level-I, all classes in NLCD2001 CS are completely and uniquely represented by level-I classes in NLCD1992 CS. However, cross-level complete matches also exist between ‘01_40’ and ‘92_33’, ‘01_40’ and ‘92_43’, ‘01_50’ and ‘92_33’, ‘01_50’ and

‘92_51’, and ‘01_70’ and ‘92_71’, respectively. A complete and unique translation in level-II occurs only on ‘01_11’ and ‘01_12’. Most of level-II classes in NLCD2001 CS are completely covered by class ‘92_33’, including ‘01_41’, ‘01_42’, ‘01_43’, ‘01_52’, ‘01_71’, ‘01_91’ and ‘01_92’. Ignoring the situation covered by ‘92_33’, the latter four classes are completely and uniquely covered by one level-II class in NLCD1992 CS. A part of ‘01_31’ is covered by ‘92_31’ and ‘92_32’, respectively. Level-II classes of ‘01_40’ are partially translated into level-II classes of ‘92_40’ correspondingly.

Comparing Table 9.6 with Table 9.4, identical classes include ‘92_10’ and ‘01_10’, ‘92_30’ and ‘01_30’, ‘92_40’ and ‘01_40’, ‘92_50’ and ‘01_50’, ‘92_70’ and ‘01_70’, ’92-9’ and ‘01_9’, ‘92_11’ and ‘01_11’, ‘92_12’ and ‘01_12’, ‘92_71’ and ‘01_71’, ‘92_91’ and ‘01_91’, and ‘92_92’ and ‘01_92’. There are subordinate and superordinate relationships between ‘92_31’, ‘92_32’ and ‘01_31’, respectively, between ‘92_41’, ‘92_42’, ‘92_43’, ‘92_51’, ‘92_71’, ‘92_91’, ‘92_92’, and ‘01_33’, respectively, between ‘92_41’ and ‘01_41’, between ‘92_41’ and ‘01_43’, between ‘92_42’ and ‘01_42’, between ‘92_42’ and ‘01_43’, between ‘92_43’ and ‘01_43’, and between ‘01_52’ and ‘92_51’.

Table 9.6 Concept-level Interoperability (01-92) in Natural Zone

	92_10	92_11	92_12	92_30	92_31	92_32	92_33	92_40	92_41	92_42	92_43	92_50	92_51	92_70	92_71	92_9	92_91	92_92
01_10	1	0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
01_11	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
01_12	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
01_30	0	0	0	1	0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0
01_31	0	0	0	1	0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0
01_40	0	0	0	0	0	0	1	1	0.5	0.5	1	0	0	0	0	0	0	0
01_41	0	0	0	0	0	0	1	0.5	0.5	0	0.5	0	0	0	0	0	0	0
01_42	0	0	0	0	0	0	1	0.5	0	0.5	0.5	0	0	0	0	0	0	0
01_43	0	0	0	0	0	0	1	0.5	0.25	0.3	0.5	0	0	0	0	0	0	0
01_50	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0
01_52	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0
01_70	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
01_71	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
01_9	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0.5	0.5
01_91	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0
01_92	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1

9.4.4 Category-level Interoperability (NLCD1992 CS to NLCD2001 CS)

The category-level interoperability, which is a transformation between mixtures, is more complicated than the concept-level interoperability. Overlaps exist between categories of a classification system as demonstrated in Chapter 6. An indicator, called the transform rate, is proposed to reflect how much the category has been transformed. The transform rate ranges from 0 to 1. The larger the value, the more can be transformed. A value of 0 means the class cannot be represented by the target classification system, whereas 1 means it can be entirely represented. As some classes are not defined as categories, the-category level interoperability is only performed on a part of classes.

9.4.4.1 Artificial Zone

The results of category-level interoperability from NLCD1992 CS to NLCD2001 CS in the artificial zone are presented in Table 9.7. All class categories in NLCD1992 CS can be entirely represented. Classes ‘92_22’ and ‘92_23’ are completely and uniquely represented by ‘01_24’, respectively. Class ‘92_21’ is partitioned and represented by ‘01_22’, ‘01_23’, and ‘01_24’.

Table 9.7 Category-level Interoperability (92-01) in Artificial Zone

Transform rate		01_21	01_22	01_23	01_24	01_81	01_82
1	92_21	0	0.6208	0.3688	0.0103	0	0
1	92_22	0	0	0	1	0	0
1	92_23	0	0	0	1	0	0

9.4.4.2 Natural Zone

The results of category-level interoperability from NLCD1992 CS to NLCD2001 CS in the natural zone are presented in Table 9.8. Out of 9 class categories, 5 can be completely transferred, but none of them are uniquely transferred. Class

‘92_11’ can be entirely transferred to ‘01_11’, but also more than 25% can be transferred to ‘01_52’ and ‘01_91’. Semantic overlaps exist within a classification system and have been evaluated in Chapter 6. Only 62% of ‘92_41’ and ‘92_42’ are represented by ‘01_41’ and ‘01_42’, respectively. More than 90% of ‘92_43’ is represented by ‘01_43’. Class ‘92_91’ can be entirely represented by ‘01_91’ but some of it can also be represented by other classes, such as ‘01_52’.

Around 70% of ‘92_33’ can be transferred, and mainly covered by ‘01_52’ and ‘01_91’ (33% respectively). Around 50% of ‘92_51’ can be transferred and mainly covered by ‘01_52’ (38%). Almost all of ‘92_71’ and ‘92_92’ that can be transferred is covered by ‘01_71’ and ‘01_92’, respectively.

Table 9.8 Category-level Interoperability (92-01) in Natural Zone

Transform Rate		**01_11**	**01_12**	**01_31**	**01_41**	**01_42**	**01_43**	**01_52**	**01_71**	**01_91**	**01_92**
1	**92_11**	1	0	0	0.0005	0.0005	0.0005	0.2644	0	0.2644	0.0160
0.7171	**92_33**	0.0068	0.3075	0.0097	0	0	0	0.3280	0	0.3280	0.0096
1	**92_41**	0	0.0267	0	0.6286	0	0.3714	0.1372	0	0.1372	0
1	**92_42**	0	0.0267	0	0	0.6286	0.3714	0.1372	0	0.1372	0
1	**92_43**	0	0.0254	0	0.0306	0.0306	0.9387	0.1322	0	0.1322	0
0.5184	**92_51**	0	0.0368	0	0	0	0	0.3769	0	0.1801	0
0.6575	**92_71**	0	0	0	0.0005	0.0005	0.0005	0.0011	0.6538	0.0011	0
1	**92_91**	0	0.0217	0	0.0358	0.0358	0.0936	0.1170	0	1	0
0.6575	**92_92**	0	0	0	0.0005	0.0005	0.0005	0.0011	0	0.0011	0.6538

9.4.5 Category-level Interoperability (NLCD2001 CS to NLCD1992 CS)

9.4.5.1 Artificial Zone

The results of category-level interoperability from NLCD2001 CS to NLCD1992 CS in the artificial zone are presented in Table 9.9. All class categories are partially transferred. Only 13% of ‘01_22’ is transferred to ‘92_21’. Half of ‘01_23’ is transferred to ‘92_21’. Around 75% of ‘01_24’ is transferred to ‘92_21’, ‘92_22’, and ‘92_23’ correspondingly. It is important to note that the transform rate is 0 for ‘01_21’, ‘01_81’, and ‘01_82’. The reason for ‘01_21’ is that only one property is employed to define class ‘20’ and its descendants in both systems. Its attribute range is beyond others. The reason for ‘01_81’ and ‘01_82’ is mainly that subordinate classes of ‘92_80’ are defined as concepts rather than categories.

Table 9.9 Category-level Interoperability (01-92) in Artificial Zone

Transform rate		**92_21**	**92_22**	**92_23**
0	**01_21**	0	0	0
0.1341	**01_22**	0.1341	0	0
0.5066	**01_23**	0.5066	0	0
0.7429	**01_24**	0.5714	0.0857	0.0857
0	**01_81**	0	0	0
0	**01_82**	0	0	0

Comparing Table 9.9 with Table 9.7, there are no identical classes between NLCD1992 CS and NLCD2001 CS at the category-level. There exists a subordinate and superordinate relationship between ‘92_22’ and ‘01_24’, and between ‘92_23’ and ‘01_24’.

9.4.5.2 Natural Zone

The results of category-level interoperability from NLCD2001 CS to NLCD1992 CS in the natural zone are presented in Table 9.10. Out of 10 class categories, 5 can be completely transferred and only ‘01_71’ is uniquely transferred. Class ‘01_11’ can be entirely transferred to ‘92_11’ but there is also a large overlap (81%) between ‘01_11’ and ‘92_33’. Class ‘01_41’ and class ‘01_42’ are mainly (94%) transferred to ‘92_41’ and ‘92_42’, respectively. Only around 60% of ‘01_43’ is transferred to ‘92_43’, and others (19%) are mainly transferred to ‘92_41’ and ‘92_42’.

Around 50% of ‘01_12’ is transferred, although ‘92_12’ is absent. Approximately 30% of ‘01_12’ is transferred to ‘92_51’. 96% of ‘01_31’ can be entirely transferred to ‘92_33’. Around 85% of ‘01_52’ can be transferred and mainly (68%) transferred to ‘92_51’. 70% of ‘01_91’ can be transferred, and only around 40% is transferred to '92_91'. Surprisingly, only 14% of '01_92' can be transferred, and only 5% of it is transferred to '92_92'.

Table 9.10 Category-level Interoperability (01-92) in Natural Zone

Transform rate		**92_11**	**92_33**	**92_41**	**92_42**	**92_43**	**92_51**	**92_71**	**92_91**	**92_92**
1	**01_11**	1	0.8059	0.0005	0.0005	0.0005	0.0048	0	0.0002	0
0.4970	**01_12**	0	0.0357	0.0433	0.0433	0.0548	0.2876	0	0.0377	0
0.9636	**01_31**	0	0.9636	0	0	0	0	0	0	0
1	**01_41**	0	0	0.9391	0	0.0608	0	0	0.0572	0
1	**01_42**	0	0	0	0.9391	0.0608	0	0	0.0572	0
1	**01_43**	0	0	0.1865	0.1865	0.6269	0	0	0.0502	0
0.8639	**01_52**	0.0001	0.0088	0.0512	0.0512	0.0656	0.6772	0	0.0467	0
1	**01_71**	0	0	0	0	0	0	1	0	0
0.7068	**01_91**	0.0001	0.0088	0.0512	0.0512	0.0656	0.3236	0	0.3989	0
0.1395	**01_92**	0.0012	0.0886	0	0	0	0	0	0	0.0508

Comparing Table 9.10 with Table 9.8, there exists one identical class pair, '92_11' and '01_11'. There exists a subordinate and superordinate relationship between '01_71' and '92_71', and between '92_91' and '01_91'.

9.5 Discussion

9.5.1 Uncertainty Propagated from Uncertainties within a CS

Semantic problems hidden in a classification system, such as semantic overlap and gap, make interoperability more complex and confusing. They increase the difficulty of interpreting the experimental results. Ideally, the transform rate applied in section 9.4.4 and section 9.4.5 should be equal to the sum of all overlaps with that class category, i.e., it should be identical to the sum of rows except for the transform rate. However, it is seldom the case in practice, which has been demonstrated in the case study.

Level-II classes in section 9.4.2.1 and section 9.4.3.1 are used to demonstrate how the elimination of such semantic problems will benefit the interoperability process. Only one group of identical concepts exists within NLCD1992 CS and

NLCD2001 CS respectively: class '92_21' and '92_22' in NLCD1992 CS, which will be uniquely labelled as '92_21,22', and class '01_21', '01_22', '01_23', and '01_24' in NLCD2001 CS, which will be uniquely labelled as '01_21,22,23,24'. The results of concept-level interoperability are presented in Table 9.11 and Table 9.12. It is obvious that all classes, except for class '92_61' and class '92_85', can be completely transferred between these two systems and there is only one identical pair: class '92_81' and '01_81'. All classes in NLCD1992 CS, except for '92_61', '92_81', and '92_85', are subordinate classes of corresponding classes in NLCD2001 CS.

Table 9.11 Example of Level-II Classes Concept level Interoperability (92-01) in Artificial Zone after Eliminating Semantic Problems within a Classification System

	01_21,22,23,24	**01_81**	**01_82**
92_21,22	1	0	0
92_23	1	0	0
92_61	0	0	0.5
92_81	0	1	0
92_82	0	0	1
92_83	0	0	1
92_84	0	0	1
92_85	0	0	0

Table 9.12 Example of Level-II Classes Concept level Interoperability (01-92) in Artificial Zone after Eliminating Semantic Problems within a Classification System

	92_21,22	**92_23**	**92_61**	**92_81**	**92_82**	**92_83**	**92_84**	**92_85**
01_21,22,23,24	0.5	0.5	0	0	0	0	0	0
01_81	0	0	0	1	0	0	0	0
01_82	0	0	0.25	0	0.25	0.25	0.25	0

The benefits resulting from the elimination of semantic problems within a classification system include: (1) confusion hidden in values is avoided as all classes within a classification system are clearly separately with each other; (2) the results become easy to read and compare; and (3) more information is revealed which will help to make consolidated conclusions and solve existing problems.

9.5.2 Discussion on Concept-level and Category-level Interoperability

Concept-level interoperability results only partially reflect the relationship between NLCD1992 CS and NLCD2001 CS from the perspective of characteristic coverings. What has been mostly compared and communicated is actually this kind of information. Even a match table proposed based on expert knowledge is a concept-level match. The reason is that the concept-level interoperability is easy to understand, operate, and communicate. However, the full version of the story is the category-level interoperability.

Different aspects are revealed by concept-level interoperability and category-level interoperability. Hence, it is meaningless to make a direct comparison between them. However, the pattern of values can be cross-analysed. In contrast to the results of concept-level interoperability, the results of category-level interoperability contain fewer 0 and 1 values, fewer identical values, and more different values. The results of concept-level are more in agreement with what we thought about these classes. The results of category-level are more approximate of the actual truth, although some results may be shocking, such as the transfer of '01_92'.

Although characteristic coverings of '92_92' and '01_92' are identical to 'CL16', differences in category-level interoperability are surprising. 65% of '92_92' can be represented by '01_92', but only 5% of '01_92' can be represented by '92_92'. A direct difference in their definitions is that the area accounts for 75% to 100% in NLCD1992 CS, whereas it accounts for greater than 80% in NLCD2001 CS. But the main reason is the selection of different referencing coverings. 75% to 100% is relative to the whole land, including vegetation-covered land and non-vegetation-covered land, while 80% is relative to only vegetation-covered land. Hence, only a small part of '01_92' coincides with

‘92_92’. For example, a piece of land where vegetation dominates 50% and ‘CL16’ dominates 45% belongs to ‘01_92’, but does not belong to ‘92_92’.

9.6 Conclusion

The method for semantic interoperability proposed in this chapter is based on DSRS. Overlap measurements are employed to evaluate the magnitude of transformation. It is hoped that larger overlaps occur on classes with similar names. However, the case study demonstrates that there are vast discrepancies in definitions of corresponding classes which are hoped to be identical. Instead of trying to eliminate overlaps within a classification system, overlap is positively employed in the interoperability process.

CHAPTER 10 SUMMARY AND FUTURE WORK

10.1 Summary

'Current mapping techniques of land cover would not be possible today without milestones such as James Anderson's 1976 publication entitled 'A Land Cover Classification System for Use with Remote Sensor Data'' (USGS, http://landcover.usgs.gov/usgslandcover.php, accessed on 27 August 2014). More and more classification systems have been established to fulfil different purposes. They are normally applied to land inventory projects without validation. However, semantic problems, such as semantic overlap, have been found in classification systems. These problems will reduce the quality of classification products and lead to confusion. Currently, semantic problems are only analysed qualitatively in very few studies. A systematic study of semantic uncertainty has not been found. The objective of this thesis is to systematically and quantitatively assess semantic uncertainties of classes in classification systems based on a common foundation. There are four main contributions.

Firstly, a new method is proposed to formalize class. Classes are divided into concepts and categories based on their definitions. A concept is only constituted by complete symbols, while a category contains at least one incomplete symbol. A category can always be transferred to a group of concepts by methods such as exhaustive enumeration. Concepts are classified into Single-concept (S-concept), Union-concept (U-concept), Join-concept (J-concept), and Complex-concept (C-concept). A formalization of them by applying product operation and union operation has been proposed. An S-concept needs neither of them. A U-concept is formalized only by union operations, while a J-concept is formalized only by product operations. A C-concept is formalized by both product operations and union operations. The formalization of class provides a method to quantitatively evaluate a class.

Secondly, a new semantic reference system, Dynamic Semantic Reference System (DSRS), is proposed to uniquely represent classes on a common foundation. Instead of establishing a top-level ontology, such as the Semantic Reference System (Kuhn, 2003) and the Land Cover Classification System (Di Gregorio, 2005) by a top-down method, DSRS is built based on 'contrast' (Nida, 1975, p.31) by a bottom-up method. The result of DSRS is a collection of reference concepts. All classes concerned can be uniquely rebuilt by these reference concepts. Along with more classes concerned, the DSRS approaches a top-level ontology. Because 'standards are themselves problematic as they are frequently revised' (Comber et al., 2008a), the prosed DSRS seems more reasonable theoretically and practically.

Thirdly, semantic uncertainties are systematically studied, and corresponding models are proposed for quantitative evaluation. Semantic uncertainties of classification systems include semantic overlap, semantic similarity, semantic gap, semantic overflow, and semantic interoperability. Semantic overlap and semantic similarity exist between different classes. Semantic gap and semantic overflow occur between different hierarchical levels in a classification system. Semantic interoperability applies to classes between different classification systems. Models are proposed to measure semantic overlap, semantic similarity, semantic gap, and semantic overflow based on DSRS. The model to measure semantic interoperability is the same with semantic overlap but applies to classes between different classification systems.

Fourthly, semantic issues related to NLCD CSs are systematically studied for demonstration. NLCD CSs include NLCD1992 CS and NLCD2001 CS, which applied to produce products NLCD1992, NLCD2001, NLCD2006, and NLCD2011. In NLCD1992 CS, 4 classes are concepts, and 5 classes are categories in level-I, and 9 classes are concepts, and 12 classes are categories in level-II. In NLCD2001 CS, only 4 classes in level-I are concepts, other level-I classes and all level-II classes are categories. The DSRS concerning NLCD CSs contains 21 reference concepts. Experimental results reveal that semantic uncertainties are widespread in NLCD CSs. Semantic uncertainties of classes represented by characteristic coverings are closer to what has been mentioned by other authors. However, land classes are mixtures which should be categories. More semantic uncertainties are found when classes are treated as mixtures. Furthermore, because all five measurements are calculated based on the same DSRS, they can work together to reveal different aspects of the classification systems systematically and to obtain a measure avoiding of unexpected influences. For example, classes in a land classification system should not be overlapped in a level. In the experiments, overlaps widely occur (see section 10.2 as examples). By the method proposed in this thesis, overlaps can be eliminated before similarity is measured. Hence, a similarity measurement without the influence of overlaps can be acquired.

Moreover, classification is employed in every discipline. The most popular and normal method to define these classes is natural language which is very complex. How to decompose natural language into basic elements for comparison and quantitative analysis plays a key role in scientific research. A meaningful and feasible method has not been achieved from the author's perspective. This thesis proposes a method to construct a Dynamic Semantic Reference System based on 'contrast'. A unique and quantitative reference is obtained. The application of this method to other disciplines will also construct a dynamic and common foundation for information retrieving, comparison, sharing, and analysis.

10.2 Example Problems in NLCD2001 CS and Suggestions

10.2.1 Examples

Classes are mixtures other than pure covering materials. Problematic semantic overlaps must be abandoned from a classification system. Examples of problematic semantic overlaps between classes in NLCD2001 CS are listed as follows.

Table 10.1 Examples of Problematic Semantic Overlaps between Classes in NLCD2001 CS (in Artificial Zone)

	Class Code		Area Percentage of Reference Concepts（%）								
	C1	C2	CL1	CL2	CL7	CL8	CL17	CL18	CL19	CL20	CL21
1	20	81	0	30	0	0	0	0	0	0	70
2	20	82	0	30	0	0	0	0	15	0	55
3	21	80	0	5	0	0	0	0	0	0	95
4	21	81	0	5	0	0	0	0	0	0	95
5	21	82	0	5	0	0	0	0	20	0	75
6	22	80	0	20	0	0	0	0	0	0	80
7	22	81	0	20	0	0	0	0	0	0	80
8	22	82	0	20	0	0	0	0	20	0	60
9	23	81	0	50	0	0	0	0	0	0	50
10	23	82	0	50	0	0	0	0	15	0	35
11	24	81	0	80	0	0	0	0	0	0	20
12	24	82	0	80	0	0	0	0	5	0	15
13	81	82	0	0	0	0	0	0	25	0	75

Interpretations of Table 10.1 and Table 10.2 should reference on Figure 5.4, Table 5.4, and Table 5.5. For example, a land can be classified to both class ‘21’ and class ‘81’, if the land is covered by 5% of CL2 and 95% of CL21. In other words, if 5% of the land is covered by constructed materials which are not for residence, and the rest 95% of the land is covered by planted or managed herbaceous vegetation which are not for man use and not for urban/recreational uses, the land can be classified either ’21. Developed, Open Space’ or ‘81. Pasture/Hay’.

Table 10.2 Examples of Problematic Semantic Overlaps between Classes in NLCD2001 CS (in Natural Zone)

	Class Code		Area Percentage of Reference Concepts（%）											
	C1	C2	CL3	CL4	CL5	CL6	CL9	CL10	CL11	CL12	CL13	CL14	CL15	CL16
1	11	40	75	0	0	0	0	0	0	25	0	0	0	0
2	11	41	75	0	0	0	0	0	0	0	10	0	0	15
3	11	42	75	0	0	0	0	0	0	0	0	10	0	15
4	11	43	75	0	0	0	0	0	0	0	5	5	0	15
5	11	52	75	0	0	0	0	10	0	0	0	0	0	15
6	11	71	75	0	0	0	0	0	0	0	0	0	25	0
7	11	91	75	0	0	0	10	0	0	0	0	0	0	15
8	11	92	75	0	0	0	0	0	0	0	0	0	0	25
9	12	40	0	30	0	0	0	0	0	25	0	0	0	45
10	12	41	0	30	0	0	0	0	0	0	15	0	0	55
11	12	42	0	30	0	0	0	0	0	0	0	15	0	55
12	12	43	0	30	0	0	0	0	0	0	5	10	0	55
13	12	52	0	30	0	0	0	15	0	0	0	0	0	55
14	12	71	0	30	0	0	0	0	0	0	0	0	60	10
15	12	91	0	30	0	0	15	0	0	0	0	0	0	55
16	12	92	0	30	0	0	0	0	0	0	0	0	0	70
17	31	41	0	0	0	85	0	0	0	0	5	0	0	10
18	31	42	0	0	0	85	0	0	0	0	0	5	0	10
19	31	43	0	0	0	85	0	0	0	0	5	5	0	5
20	31	52	0	0	0	85	0	5	0	0	0	0	0	10
21	31	71	0	0	0	85	0	0	0	0	0	0	15	0
22	31	91	0	0	0	85	5	0	0	0	0	0	0	10
23	31	92	0	0	0	85	0	0	0	0	0	0	0	15
24	40	52	0	0	0	0	0	25	0	25	0	0	0	50
25	40	91	0	0	0	0	25	0	0	25	0	0	0	50
26	41	52	0	0	0	0	0	25	0	0	20	5	0	50
27	41	91	0	0	0	0	25	0	0	0	20	5	0	50
28	42	52	0	0	0	0	0	25	0	0	0	25	0	50
29	42	91	0	0	0	0	25	0	0	0	0	25	0	50
30	43	52	0	0	0	0	0	25	0	0	10	15	0	50
31	43	91	0	0	0	0	25	0	0	0	10	15	0	50
32	52	91	0	0	0	0	25	25	0	0	0	0	0	50
33	70	40	0	0	0	0	0	0	0	25	0	0	75	0
34	70	41	0	0	0	0	0	0	0	0	20	5	75	0
35	70	42	0	0	0	0	0	0	0	0	0	25	75	0
36	70	43	0	0	0	0	0	0	0	0	10	15	75	0
37	70	52	0	0	0	0	0	25	0	0	0	0	75	0
38	70	91	0	0	0	0	25	0	0	0	0	0	75	0

10.2.2 Suggestions

It has been demonstrated that semantic problems are popular in existing land classification systems. Quantified evaluations are executed to find these problems accurately. Results of measurements can be used to guild the design of future classification systems. From the author's knowledge and experience, to avoid semantic problems, a land classification system is better to:

- based on classifiers other than names;
- include concept tree which obviously reflect what kinds of materials cover the land and the relationships among these materials;
- clearly set priorities for all classes at a level;
- only define the finest level of classes which can aggregate to the higher level classes.

10.3 Future Work

Semantic uncertainties have been studied in this thesis purely based on classification systems. Future works can be conducted in the following three directions.

- Improvements can be made in the designation of a classification system. On one hand, semantic uncertainties can be evaluated for a potential classification system. The semantic problems found can help designers to make better criteria to improve the classification system. The process can be implemented iteratively until a satisfying classification system comes out. On the other hand, the representation form of a classification system may also be improved. Conventional classification systems are normally based on nomenclatures and definitions in natural languages, which are prone to cause uncertainties. LCCS is based on classifiers other than nomenclatures, which are better to keep consistency but lack feasibility. A new trial may describe a classification system by a concept tree, certain popular illustrations, and percentages of concepts. The concept tree explicitly determines the structure of all concerned concepts. Illustrations are used to help understanding. Percentages of concepts are used when they are categories.

- Combination of results of this thesis with practical situations can be made. Semantic uncertainties in this thesis are only theoretical values. Practically, they are different when characteristics of land vary. Influences of semantic uncertainty are constrained by spatial characteristics of a region. Even there is high semantic uncertainty between two classes, if they do not exist in the region, they do not have influence in this region. For example, 'Perennial Ice/Snow' (Code: 12) does not exist in Hong Kong, but exists in the northwest of China (e.g., Urumqi). Semantic uncertainties of NLCD1992 CS applied in Hong Kong should be different from those in Urumqi.

- The propagation of semantic uncertainties could be studied. A classification system is established at the beginning of a land inventory project and applied throughout the project until the products are out of use. Hence, how they are propagated in the process should be studied. For example, a study can be conducted on the propagation of semantic uncertainties to area measurement uncertainty. Semantic uncertainties constrained by spatial characteristics may make it possible to predict area measurement uncertainty.

- Semantic uncertainties of more classification systems should be conducted. The methods proposed in this thesis can be applied to other land classification systems directly. It is meaningful to evaluate the semantic issues of other systems, such as Current Land Cover System in China. Furthermore, a more general DSRS will be established by adding more class definitions of other systems. Thereafter, semantic uncertainties of different systems can be revealed and compared based on the same foundation. And semantic interoperability between different systems can also be executed, which will benefit information sharing.

REFERENCES

Agarwal, P., 2005. Ontological considerations in GIScience. International Journal of Geographical Information Science, 19(5), 501-536.

Ahlqvist, O., 2004. A parameterized representation of uncertain conceptual spaces. Transactions in GIS, 8(4), 493-514.

Ahlqvist, O., and Gahegan, M., 2005. Probing the relationship between classification error and class similarity. Photogrammetric Engineering & Remote Sensing, 71(12), 1365-1373.

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. Washington, DC: Government Printing Office (US Geological Survey, Professional Paper 964).

Bartholome, E., and Belward, A.S., 2005. GLC2000: A new approach to global land cover mapping from earth observation data. International Journal of Remote Sensing, 26(9), 1959-1977.

Batchelor, B.G., 1978. Pattern recognition: Ideas in practice. New York: Plenum Press, 71-72.

Bian, L., and Hu, S., 2007. Identifying components for interoperable process models using concept lattice and semantic reference system. International Journal of Geographical Information Science, 21(9), 1009-1032.

Bicheron, P., Defourny, P., Brockmann, C., Schouten, L., Vancutsem, C., Huc, M., Bontemps, S., Leroy, M., Achard, F., Herold, M., Ranera, F., and Arino, O., 2008. GLOBCOVER: Products description and validation report. Tolouse: MEDIAS-France.

Burley, T.M., 1961. Land use or land utilization?, Prof. Geographer, 13(6), 18-20.

Büttner, G.; Feranec, G.; and Jaffrain, G., 2002. Corine land cover update 2000, Technical guidelines. EEA Technical report No 89. http://reports.eea.europa.eu/technical_report_2002_89/en.

Büttner, G., Feranec, J., Jaffrain, G., Man, L., Maucha, G., and Soukup, T., 2004. The CORINE land cover 2000 project. In: Reuter, R. (Ed.), EARSeL eProceedings, 3(3). Paris (EARSeL): 331–346.

Cadenasso, M.L., Pickett, S.T.A., and Schwarz, K., 2007. Spatial heterogeneity in urban ecosystems: reconceptualizing land cover and a framework for classification. Frontiers in Ecology and the Environment 5(2), 80–88.

Cihlar J., 2000. Land cover mapping of large areas from satellites: Status and research priorities. International Journal of Remote Sensing, 20(6), 1093-1114.

Clawson, M., and Stewart, C.L., 1965. Land use information. A critical survey of U.S. statistics including possibilities for greater uniformity: Baltimore, Md., The Johns Hopkins Press for Resources for the Future, Inc..

Comber, A.J., Fisher, P.F., and Wadsworth, R.A., 2008a. Land cover: To standardise or not to standardise? Comment on “Evolving standards in land cover characterization” by Herold et al. Journal of Land Use Science, 2(4), 283-287.

Comber, A.J., Fisher, P.F., and Wadsworth, R.A., 2008b. Semantics, Metadata, Geographical Information and Users. Transactions in GIS, 12(3), 287-291.

Crovello, T.J., 1970. Theory and practice of classification in diverse disciplines. Science, 169(3944), 505-506.

Di Gregorio, A., 2005. Land cover classification system: classification concepts and user manual: LCCS. Food and Agriculture Organization of the United Nations.

Di Gregorio, A., 2011. Overview on land cover classifications and their interoperability: The FAO LCML (Land Cover Meta-Language). In Expert

Meeting on Ecosystem Accounts, ISSUE3: Land cover mapping, land cover classifications, and accounting units, London, UK, 5-7 December 2011.

Di Gregorio, A., and Jansen, L.J.M., 1998. A new concept for a land cover classification system. The Land, 2(1), 55-65.

DRDLR, 2013. Report on the workshops on land use classification standards, methodology and symbology, Spatial Planning and Information (SPI). Department of Rural Development and Land Reform (DRDLR), Republic of South Africa, April 2013.

EPA, 2003. CORINE land cover 2000 update (Ireland) final report, Environment Protection Agency, Ireland.

EPA, 2006. CORINE land cover - Ireland, land cover update for 2006 final report, Environment Protection Agency, Ireland.

FAO, 2001. FRA2000 Global Forest Cover Mapping Final Report. Rome.

Feddema, J.J., Oleson, K.W., Bonan, G.B., Mearns, L.O., Buja, L.E., Meehl, G.A., and Washington, W.M., 2005. The importance of land-cover change in simulating future climates. Science, 310(5754), 1674-1678.

Foley, J.A., DeFries, R. Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., and Snyder, P.K., 2005. Global consequences of land use. Science, 309(5734), 570-574.

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

Frumkina, R.M., and Mikhejev, A.V., 1996. Meaning and categorization. New York: Nova Science Publishers.

Fry, J.A., Xian, G., Jin, S., Dewitz, J.A., Homer, C.G., Yang, L., Barnes, C.A., Herold, N.D., and Wickham, J.D., 2011. Completion of the 2006 National Land Cover Database for the conterminous United States. Photogrammetric Engineering & Remote Sensing, 77(9), 858-864.

Gärdenfors, P., 2000. Conceptual space: The geometry of thought. Cambridge, MA: MIT Press.

Ganter, B., and Wille, R., 1999. Formal concept analysis: Mathematical foundations, Springer-Verlag.

GB/ T 21010-2007, 2007. Current land use classification (Chinese).

Gentner, D. and Markman, A.B., 1994. Structural alignment in comparison: No difference without similarity. Psychological Science, 5(3), 152–158.

GEOSS, 2005. The Global Earth Observation System of Systems (GEOSS) 10-year implementation plan. http://www.earthobservations.org/documents.shtml.

Goel N.S., and Hsu S., 1992. Overview of automated target recognition and tracking. Remote Sensing Reviews, 6(1), 1-10.

Goldstone, R.L., 1999. Similarity. In: Wilson, R.A. and Keil, F.C. (eds.), Mit Encyclopedia of the Cognitive Sciences. pp. 763-765. Cambridge, MA: MIT Press.

Goldstone, R.L., and Son, J., 2005. Similarity. In: Holyoak, K. and Morrison, R. (eds.), Cambridge Handbook of Thinking and Reasoning, pp. 13–36. Cambridge University Press.

Grimm, N.B., Faeth, S.H., Golubiewski, N.E, Redman, C.L., Wu, J., Bai, X., and Briggs, J.M., 2008. Global change and the ecology of cities. Science, 319(5864), 756-760.

Grimm, L.R., Rein, J.R., and Markman, A.B., 2012. Determining transformation distance in similarity: Considerations for assessing representational changes a priori. Thinking & Reasoning, 18(1), 59-80.

Hahn, U., Close, J., and Graf, M., 2009. Transformation direction influences shape similarity judgements. Psychological Science, 20, 447-454.

Hakimpour, F., and Geppert, A., 2001. Resolving semantic heterogeneity in schema integration: An ontology based approach. In: Welty, C. and Smith, B. (eds.), Proceedings of International Conference on Formal Ontologies in Information Systems FOIS'01. ACM Press.

Hansen, M.C., Defries, R.S., Townshend, J.R.G., and Sohlberg, R., 2000. Global land cover classification at 1 km spatial resolution using a classification tree approach. International Journal of Remote Sensing, 21(6-7), 1331-1364.

Herold, M., Hubald, R., and Di Gregorio, A., 2009. Translating and evaluating land cover lengends using the UN Land Cover Classification System (LCCS). GOFC-GOLD Report No. 43, Land Cover Project Office, Jena, Germany, June, 2009.

Herold, M., Latham, J.S., Di Gregorio, A., and Schmullius, C.C., 2006a. Evolving standards in land cover characterization. Journal of Land Use Science, 1(2-4), 157-168.

Herold, M., Mayaux, P., Woodcock, C.E., Baccini, A., and Schmullius, 2008. Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. Remote Sensing of Environment, 112(5), 2538-2556.

Herold, M., and Schmullius, C., 2004. Report on the harmonization of global and regional land cover products meeting, Rome, Italy.

Herold, M., Woodcock, C.E., Di Gregorio, A., Mayaux, P., Belward, A.S., Latham, J., and Schmullius, C.C., 2006b. A joint initiative for harmonization and validation of land cover datasets. IEEE Transactions on Geoscience and Remote Sensing, 44(7), 1719-1727.

Hodgetts, C.J., Hahn, U., and Chater, N., 2009. Transformation and alignment in similarity. Cognition, 113, 62-79.

Homer, C., Huang, C., Yang, L., Wylie, B., and Coan, M., 2004. Development of a 2001 national land-cover database for the United States. Photogrammetric Engineering & Remote Sensing, 70(7), 829-840.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., Mckerrow, A., VanDriel, J.N., and Wickham, J., 2007. Completion of the 2001 National Land Cover Database for the Conterminous United States. Photogrammetric Engineering & Remote Sensing, 73(4), 337-341.

Houghton, R.A., and Hackler, J.L., 2001. Carbon flux to the atmosphere from land-use changes: 1850 to 1990. ORNL/CDIAC-131, NDP-050/R1 (Oak Ridge National Laboratory, Oak Ridge, TN,2001).

Hu S., and Wang L., 2013. Automated urban land-use classification with remote Ssensing. International Journal of Remote Sensing, 34(3), 790-803.

Hudak, P., 2000. The Haskell school of expression: Learning functional programming through multimedia. Cambridge University Press, New York.

Hunter, G.J., 2002. Understanding semantics and ontologies: They're quite simple really – If you know what i mean. Transactions in GIS, 6(2), 83-87.

ISO 19144, 2009. ISO 19144:2009, Geographic information -- Classification systems -- Part 1: Classification system structure.

Janowicz, K., 2008. Kinds of contexts and their impact on semantic similarity measurement. In Proc. 5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea), 6th IEEE International Conference on Pervasive Computing and Communication (PerCom), IEEE Computer Society.

Janowicz, K., Raubal, M., and Kuhn, W., 2011. The semantics of similarity in geographical information retrieval. Journal of Spatial Information Science, 2, 29-57.

Jansen, L.J.M., and Di Gregorio, A., 2002. Parametric land cover and land-use classifications as tools for environmental change detection. Agriculture, Ecosystems and Environment, 91(1-3), 89-100.

Jansen, L.J.M., Groom, G., and Carrai, G., 2008. Land-cover harmonization and semantic similarity: some methodological issues. Journal of Land Use Science, 3(2-3), 131-160.

Joseph, G., 1996. Imaging sensors for remote sensing. International Journal of Remote Sensing, 13(3-4), 257-342.

Kavouras, M., and Kokla, M., 2008. Theories of geographic concepts: Ontological approaches to semantic integration, CRC Press.

Kleeschulte, S., and Büttner, G., 2008. European land cover mapping – The CORINE experience. In 'North America Land Cover Summit' edited by Campbell J.C., et al.. Washington, DC: Association of American Geographers. pp. 31-43.

Kosolapoff, G.M., 1945. Basis for scientific terminology and classification. Science, 101(2613), 89-90.

Kuhn, W., 2003. Semantic reference systems. International Journal of Geographical Information Science, 17(5), 405-409.

Kuhn, W., and Raubal, M., 2003. Implementing semantic reference systems. In M. Gould, R. Laurini and S. Coulondre (Eds), Proceedings of the 16th AGILE Conference on Geographic Information Science, AGILE 2003, Lyon, France, 63-72.

Lam, D.K., and Remmel, T.K., 2010. Effect of configuration on spectral signatures. Canadian Journal of Remote Sensing, 36(1), 48-54.

Landgrebe, D.A., 2003. Signal theory methods in multispectral remote sensing (Hoboken, NJ: John Wiley and Sons).

Latham, J., 2008. FAO land cover mapping initiatives. In ‘North America Land Cover Summit’ edited by Campbell J.C., et al.. Washington, DC: Association of American Geographers. pp. 75-96.

Lin G.C.S, and Ho S.P.S., 2003. China’s land resources and land-use change: insights from the 1996 land survey. Land Use Policy, 87-107.

Lin, W-S., Wu, J-Y, Chiu, H-J, Chen, C-S, and Chang, Y-J., 1999. Radiometric characterization of the ROCSAT ocean color imager. International Journal of Remote Sensing, 20(17), 3247-3263.

Liu, T., and Yang, X., 2013. Mapping vegetation in an urban area with stratified classification and multiple endmember spectral mixture analysis. Remote Sensing of Environment, 133, 251-264.

Lu, D., and Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing, 28(5), 823-870.

Melara, R.D., Marks, L.E., and Lesko, K.E., 1992. Optional processes in similarity judgements. Perception & Psychophysics, 51(2), 123-133.

Nadler, M., and Smith, E.P., 1993. Pattern recognition engineering. New York: Wiley, 293-294.

NASA news feature, 2005. Tropical deforestation affects rainfall in the U.S. and around the globe. http://www.nasa.gov/centers/goddard/news/topstory/2005/deforest_rainfall.html.

Nelson, G.C., and Robertson, R.D., 2007. Comparing the GLC2000 and GeoCover LC land cover datasets for use in economic modelling of land use. International Journal of Remote Sensing, 28(19), 4342-4262.

Nida, E.A., 1975. Componential analysis of meaning: An introduction to semantic structures. The Hague: Mouton.

Pirró, G., and Euzenat, J., 2010. A feature and information theoretic framework for semantic similarity and relatedness. In: Patel-Schneider et al. (eds.), ISWC 2010, Part I. LNCS, vol. 6496, 615-630.

Rada, R., Mili, H., Bicknell, E., and Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1), 17-30.

Ritter, W.E., 1916. The place of description, definition and classification in philosophical biology. The Scientific Monthly, 3(5), 455-470.

Rodríguez, M.A., 2000. Assessing semantic similarity among spatial entity classes. Ph.D. Thesis, University of Maine.

Rodríguez, M.A., and Egenhofer, M.J., 2004. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. International Journal of Geographical Information Science, 18(3), 229-256.

Salzberg, S., 1991. A nearest hyperrectangle learning method. Machine Learning, 6, 277-309.

Scholes, R.J., Walters, M., Turak, E., Saarenmaa, H., Heip, C.H.R., O'Tuama, E., Faith, D.P., Mooney, H.A., Ferrier, S., Jongman, R.H.G., Harrison, I.J., Yahara, T., Pereira, H.M., Larigauderie, A., and Geller, G., 2012. Building a global observing system for biodiversity. Current Opinion in Environmental Sustainability, 4, 139-146.

Schwering, A., 2006. Semantic similarity measurement including spatial relations for semantic information retrieval of geo-spatial data. Ph.D. Thesis, University of Münster.

Sheth, A., 1999. Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In: Goodchild, M., Egenhofer, M., Fegeas, R. and Kottman, C. (eds.). Interoperating Geographic Information Systems. pp. 2-30. Kluwer Academic Publisher, Norwell, MA.

Shi, W.Z., 2008. Principles of modeling uncertainties in spatial data and spatial analyses. CRC Press.

Silla, C.N., and Freitas, A.A., 2010. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, 22(1-2), 31-72.

Smith, B., 2004. Beyond concepts: Ontology as reality representation. In Proc. Third International Conference on Formal Ontology In Information Systems (FOIS 2004), ed. A. Verzi and L. Vieu, 73-84. Amsterdam: IOS Press.

Smith, B., and Mark, D.M., 2001. Geographical categories: An ontological investigation. International Journal of Geographical Information Science, 15(7), 591-612.

Sokal, R.R., 1974. Classification: Purpose, principles, prospects. Science, 185 (4157), 1115-1123.

SPLUMB, 2012. Spatial planning and land use management bill. Minister of Rural Development and Land Reform, Republic of South Africa, 2012.

Tversky, A., 1977. Features of similarity. Psychological Review, 84, 327-352.

Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., and Driel, N.V., 2001. Completion of the 1990s National Land Cover Data Set for the Conterminous United States from Landsat Thematic Mapper Data and Ancillary Data Sources. Photogrammetric Engineering & Remote Sensing, 67, 650-662.

Wagle, P., Xiao, X., Tom, M.S., Cook, D.R., Matamala, R., Fischer, M.L., Jin, C., Dong, J., and Biradar, C., 2014. Sensitivity of vegetation indices and gross primary production of tallgrass prairie to severe drought. Remote Sensing of Environment, 152, 1-14.

Wille, R., 1992. Concept lattices and conceptual knowledge systems. Computers and Mathematics with Applications, 23(6-9), 493-515.

Wilson, D.R., and Martinez, T.R., 1997. Improved heterogeneous distance function. Journal of Artificial Intelligence Research, 6, 1-34.

Wood, E.M., Pidgeon, A.M., Radeloff, V.C., and Keuler, N.S., 2012. Image texture as a remotely sensed measure of vegetation structure. Remote Sensing of Environment, 121, 516-526.

Wyatt, B.K., Billington, C., De Bie, K., De Leeuw, J., Davies, N.G., and Luxmoore, R., 1997. Guidelines for land use and land cover description and classification (draft final report).