UNIVERSITÄT BAYREUTH

Biogeographical Modelling

Quantification of land use and land cover in a Monsoon agricultural mosaic from space

Dissertation

to obtain the academic degree of Doctor of Natural Science (Dr. rer. nat.)

of the Bayreuth Graduate School for Mathematical and Natural Sciences of the University of Bayreuth

presented by

Bumsuk Seo

born 14^{th} December 1978 in Seoul, Republic of Korea

Bayreuth, Aug 2015

This doctoral thesis was prepared at Biogeographical Modelling and Ecological Modelling, University of Bayreuth between April 2009 and Aug 2015 and was supervised by Prof. Dr. Björn Reineking, Prof. Dr. Thomas Köllner, Prof. Dr. John Tenhunen, and Dr. Christina Bogner.

This is a full reprint of the dissertation submitted to obtain the academic degree of Doctor of Natural Sciences (Dr. rer. nat.) and approved by the Bayreuth Graduate School of Mathematical and Natural Sciences (BayNAT) of the University of Bayreuth.

Date of submission: 17 Aug 2015
Date of defense : 15 Dec 2015

Acting Director: Prof. Dr. Stephan Kümmel

Doctoral Committee:

Prof. Dr. Björn Reineking (1^{st} reviewer)

Prof. Dr. Cyrus Samimi (2^{nd} reviewer)

Prof. Dr. Bernd Huwe (chairman)

Prof. Dr. Thomas Köllner

Summary

Land use and land cover (LULC) are fundamental elements of the global ecosystem and LULC changes are key aspects of global change. Information on LULC is essential in a wide range of research fields, including environmental science, ecosystem services, and environmental decision making. The quality of LULC information significantly impacts on the outcomes of these research applications. Hence, acquisition of appropriate LULC data is an important issue for research, especially in complex heterogeneous agricultural landscapes. Particularly in these types of landscapes, the existing global land cover (GLC) products are restricted in their thematic, spatial, and temporal resolution. Therefore, the use of the GLC products may lead to an inadequate representation of the actual landscape. For cultivated landscapes, methods able to retrieve detailed LULC data as well as improvements of GLC products are strongly desired.

This dissertation focuses on enhancing LULC quantification in complex heterogeneous agricultural landscapes. Specifically, extraction of spatially and thematically detailed LULC information from existing, medium resolution, multi-spectral satellite products is pursued. Three main contributions to LULC quantification are presented: ground data collection, derivation of continuous LULC, and classification of multi-crop LULC.

First, high-quality LULC observation data was collected over the study site Haean catchment, South Korea. The observed data illustrates the detailed LULC of the catchment for the three-year study period (2009 – 2011). A comparison with the MODerate Resolution Imaging Spectroradiometer (MODIS) land cover product (MCD12Q1) revealed limitations of this GLC product in spatial and thematic resolution. The limitations were due to the large cell size and the broadly defined cropland classes of the product. This result illustrates the difficulty in using GLC products to monitor LULC changes in complex heterogeneous landscapes.

Second, estimation of continuous LULC was addressed. For the study site, a fractional LULC regression model was developed for 10 LULC classes based on a MODIS multi-spectral dataset (MODIS 13Q1) and Random Forests models. In order to allow for making informed decisions when choosing data-processing options, three key data-processing options were evaluated: selection of spectral predictor sets (NDVI, EVI, surface reflectance, and all combined), time interval (8-day vs. 16-day), and smoothing (no smoothing vs. Savitzky-Golay filter). The models suc-

cessfully reproduced spatial distributions of the LULC fractions, thus illustrated the potential of existing, medium resolution satellite products for continuous LULC estimation. Third, a multi-crop LULC classification model was developed to improve thematic LULC representation. LULC data tends to be imbalanced as majority types dominate over minority types (e.g. unequal distributions of LULC type labels in raster maps). This imbalance is partly a cause of the under-development of multi-crop LULC products. Here, a synthetic sampling method was used to alleviate the problem of data imbalance in the LULC observation data for the study site. Artificial balancing of the training data substantially increased the classification performance of some minority LULC types. However, other minority LULC types remained difficult to classify due to substantial class overlaps (i.e. spectral similarities between LULC types).

For ecosystem research and decision making, continuous representations of LULC and multi-crop LULC are key information sources. In this dissertation, approaches connecting extensive field work, remote sensing and state-of-the-art analysis methods (e.g. Random Forests) are proposed and evaluated. It is shown that a judicious choice of data processing options (e.g. avoiding excessive data smoothing) and synthetic resampling methods can be useful to achieve better LULC presentations from medium resolution remote sensing data in complex cultivated landscapes. The data analysis approach presented in the dissertation was designed to be transferable to other landscapes. The methods can help analysing publicly available remote sensing data for creating detailed spatial and thematic representations of LULC types such as cultivated crops, and enhancing existing global land use and land cover products.

Zusammenfassung

Die Landnutzung/Landbedeckung (LULC: Land Use / Land Cover) ist ein grundlegender Faktor im globalen sozioökologischen System und ihre Veränderung ist ein bedeutender Treiber für den globalen Wandel. Informationen über LULC sind essentiell in Umweltwissenschaften, Forschungen zu Ökosystemleistungen und für Entscheidungsprozesse in der Landschaftsplanung. Die Qualität von Informationen zu LULC beeinflusst deshalb maßgeblich deren Ergebnisse. Daher ist die Akquisition von geeigneten LULC-Daten von entscheidender Bedeutung, insbesondere in komplexen heterogenen Agrarlandschaften. Für diese Landschaften weisen existierende Produkte zur globalen Landbedeckung (GLC) Einschränkungen in ihrer thematischen, räumlichen und zeitlichen Auflösung auf. Die Nutzung dieser Produkte führt daher zu einer schlechten Repräsentation der tatsächlichen Landschaften, was die Entwicklung einer Methode zur Extraktion hochwertiger LULC-Daten als auch die Verbesserung der GLC-Produkte erforderlich macht.

Die vorliegende Dissertation beschäftigt sich mit der Verbesserung der LULC-Quantifizierung in komplexen, heterogenen Agrarlandschaften. Die Gewinnung detaillierter räumlicher und thematischer LULC-Informationen auf Basis vorhandener grob aufgelöster multispektraler Satelliten-Produkte wird angestrebt. Es werden drei wesentliche Beiträge zur LULC-Quantifizierung präsentiert: Erhebung von Felddaten, kontinuierliche LULC-Repräsentation und LULC-Klassifikation von landwirtschaftlichen Systemen mit mehreren Feldfrüchten.

Erstens wurden hochqualitative LULC-Beobachtungsdaten im Forschungsgebiet Haean in Südkorea erhoben. Die Daten spiegeln die detaillierte LULC des Einzugsgebiets über den Zeitraum von drei Jahren (2009 – 2011) wider. Der Vergleich mit dem MODIS Landbedeckungsprodukt (MCD12Q1) offenbarte dessen Einschränkungen der GLC-Repräsentation im Forschungsgebiet. Die Einschränkungen der räumlichen und thematischen Auflösung des GLC-Produkts ergaben sich sowohl durch die große Pixelgröße als auch durch die weit gefassten Nutzpflanzen-Klassen.

Zweitens wurde bisher die Schätzung von kontinuierlichen LULC in Frage gestellt. In dieser Arbeit wurde basierend auf einem MODIS Multispektral-Datensatz (MODIS 13Q1) ein Regressionsmodell für fraktionales LULC für ein 10-Typen-System entwickelt, mit dem die kontinuierliche Repräsentation von LULC im Forschungsgebiet erstellt wurde. Um fundierte Entscheidungen

in Bezug auf die Auswahl geeigneter Optionen der Datenverarbeitung treffen zu können, wurden basierend auf dem Modell drei Schlüssel-Optionen der Datenverarbeitung evaluiert. Da das Modell die räumliche Verteilung von LULC-Fraktionen erfolgreich reproduzierte, hat die vorgeschlagene Methode ein Potential um gut aufgelöste Daten aus grob aufgelösten Satelliten-Produkten zu extrahieren. Die Wirksamkeit der verschiedenen Datenverarbeitung-Optionen in Bezug auf die Sub-Pixel LULC-Modellierung konnte durch deren Vergleich gezeigt werden.
Drittens wird in dieser Arbeit ein Klassifikationsmodell für mehrere Feldfrüchte vorgestellt, welches die thematische LULC-Repräsentation verbessert. LULC-Daten sind oft ungleich verteilt, weil die räumlich häufig angebauten Feldfrüchte die seltener angebauten dominieren. Dies ist einer der Gründe für mangelnde Qualität von LULC-Produkten für landwirtschaftliche Systeme mit mehreren Feldfrüchten. In dieser Arbeit wurde eine synthetische Sampling-Methode angewendet, um das Problem der Ungleichverteilung in den LULC-Daten zu vermindern. Künstliches Ausgleichen der Daten erhöhte die Klassifikationsleistung für einige Beobachtungsklassen erheblich. Die Klassifikation einiger kleinerer LULC-Klassen blieb jedoch auf Grund von substantiellen Informations-Überlappungen zwischen diesen LULC-Klassen schwierig.
Für die Ökosystemforschung und landschaftsplanerische Entscheidungsfindungen in komplexen und heterogenen Landschaften sind kontinuierliche Informationen über Landbedeckung und Landnutzung und Darstellungen von landwirtschaftlichen Systemen mit mehreren Feldfrüchten essentiell. In dieser Dissertation werden dafür Ansätze vorgeschlagen, die extensive Feldarbeit, Fernerkundung und moderne Analysemethoden (z.B. Random Forest) miteinander kombinieren. Es wird gezeigt, dass eine gut gewählte Methode der Datenvorverarbeitung (die z.B. überflüssiges Glätten vermiedet) und synthetisches Resampling zu einer Verbesserungen der LULC-Repräsentationen aus groben Fernerkundungsdaten in komplexen Kulturlandschaften führen kann. Die Modellierungsansätze und Ergebnisse dieser Studie bilden einen hilfreichen Leitfaden für die Entwicklung ähnlicher Modelle in verschiedenen Landschafen. Durch den in dieser Arbeit entwickelten Ansatz können frei verfügbare Fernerkundungsdaten zur detaillierten Identifizierung von LULC-Typen, wie z.B. bestimmter Ackerfrüchte verwendet werden und zur Verbesserung von globalen GLC-Produkten genutzt werden.

Acknowledgements

First, I want to thank Prof. Björn Reineking and Dr. Christina Bogner for being great supervisors. I benefited not only from their insights and knowledge but also from their patience and attentive encouragement. I am also grateful to Prof. John Tenhunen, Prof. Bernd Huwe, and Prof. Thomas Köllner for providing valuable advice throughout the PhD study. Dr. Christ L. Shope and Dr. Dennis Otieno, Prof. Dowon Lee, and Prof. Jeongjeon Rhee deserve co-supervision credits for the time they spent listening with patience to my (usually a bit roughly organised) ideas and even more ideas.

I thank all the members of TERRECO, Biogeograhical Modelling, Plant Ecology and Ecological Modelling groups for the great research environment with interesting discussions. I could not do anything without the marvellous department members and university staffs Ralf Geyer, Margarete Wartinger, Friederike Rothe, Dongjae Otto Lee, Pedro Gerstberger, Yongdoo Kim, Bärbel Heindl-Tenhunen, Ingeborg Vogler, Ellen Gossel. I am also obliged to Cornelia Nicodemus for her support via the international center. I learned a great deal from working in the fields with Emily Martin, Svenja Bartsch, Patrick Poppenborg, Sebastian Arnhold, Bora Lee, Eunyoung Jung, Kiyong Kim, Janine Kettering, Marianne Ruidisch, Sina Berger, Peng Zhao, Hamada Elsayed Ali, Mathias Hoffmeister, Steve Linder, Susann Schäfer, Stefan Strohmeier, Kati Wenzel, Youngsun Kim, Hyungjoon Moon, Feelgeun Song, Dr. Nguyen Trung Thanh, Axel Müller, Balint Jakli, Melanie Hauer, Daeun Ki, Bastian Göldel, Corrina Dinkel, Christian Thoma, Jongyol Park, Bongjae Gu, Miyeon Park and Doyeon Hwang; we had lively times together in the beloved complex agricultural catchment. I shared precious half-science/half-beer moments in Bayreuth with David Harter, Andereas Schweiger, Gwanyong Jeong, Severin Iri, Manuel Steinbauer, Julian Garvia, Yohannes Ayanu, Julienne Schiebold, Timothy Thrippleton, Michael Ewald, Kwanghun Choi, Klara Dolos, and Adriana Silva; my life in Bayreuth would have been extremely boring without them. I found that it is possible to share real friendships without beer. at least in Freie Christengemeinde Bayreuth. It was so nice to have Daniel Mario Reim, Lohna Bonkat, Tary Areka, Robert Owino, Girum Getachew, Indra Yohannes and Mrs. Yohannes, Chikas Danfulani, and all the church members in my life. Without their prayers, I was not able to come to this end. Also I thank you for the KNU members in TERRECO for helping me as well as the project to the most extent.

I would also like to thank my family members for their warm hearted care and support. They believed in me and allowed me to do whatever I wanted, which resulted in this dissertation. I would not have been courageous enough to come to Bayreuth and climb this small but not humble mountain. I would also like to extend thank my best friend, Heera Lee, for her being (as her) and for priceless support and encouragement.

Grant information

This research was supported by the Deutsche Forschungsgemeinschaft as an activity of the Bayreuth Center for Ecology and Environmental Research (BayCEER) in the context of the International Research Training Group TERRECO: Complex Terrain and Ecological Heterogeneity (GRK 1565/1) at the University of Bayreuth, Germany and by the Korean Research Foundation (KRF) at Kangwon National University, Chuncheon, South Korea.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Information on land, or the Earth's terrestrial surface is key to understand human–environment interactions. Land is an interface of social and environmental systems in which the vast majority of human activities occurs such as agriculture, habitation, industry, and various cultural and recreational practices. It supports the structure and functions of ecosystems across different spatial and temporal scales, consequently ecosystem services are also tightly connected to land and its changes (Müller et al., 2014; Tolvanen et al., 2014). Land has been and will remain a central theme in the study of human-environment systems (Müller et al., 2014). Availability and quality of information on land are important for ecosystem services research, decision making and studies on global change in general (Hansen et al., 2013; Schulp et al., 2011) and influence significantly the outcomes of environmental and ecological models (Mahecha et al., 2010; Matthews, 1983) as well as decision making studies.

This dissertation deals with quantification of land use and land cover (LULC) in complex heterogeneous agricultural landscapes. Specifically, this study searches for methodological advances in retrieving LULC information principally from pre-existing satellite data. In this chapter, a short introduction to the dissertation will be given. First, background and motivation regarding remote sensing of LULC in agricultural landscapes will be given. State-of-the-art of current research on global land cover products and quantification techniques are reviewed especially concerning complex heterogeneous agricultural landscapes. Then, the research gaps in current research and objectives and concepts of this dissertation will be articulated. The study site is briefly introduced at the end of the introduction.

1.1 Background and motivation

1.1.1 Land use and land cover (LULC)

Land use and land cover (LULC) is a term jointly denoting land use and land cover. Land cover denotes the bio-physical cover of the earth, which is a basis of the human and physical environments as well as a fundamental part of the global ecosystem (Di Gregorio, 2005; Herold et al., 2009; Loveland et al., 2010) and change of land cover is an important driver in global environmental changes (Goldewijk, 2001; Herold et al., 2009; Sterling et al., 2012; Vitousek, 1994). Land use denotes human activities taking place on a spatial unit that are directly related to the land surface itself (Comber, 2008). Land use has a direct link to land cover as it occurs in a certain land cover type to produce, change or maintain it. For example, "bare soil" is a land cover term as it refers to the earth's surface which outcrops bare soil or rocks. In contrast, "construction area" is a land use term as it describes how people use the bare soil cover. Often the land use and the land cover for a unit area are mixed. For example, the land cover "forest" is most commonly used as the land use "forest" (or "forestry"). Often the distinction of the two concepts is difficult, thus the use of the term LULC is prevalent in the research community (Comber, 2008).

A growing body of literature emphasises that LULC changes have impacted on Earth's climate (e.g. Chhabra et al., 2006; Foley et al., 2005; Turner et al., 2007), biodiversity (e.g. Dawson et al., 2011; Hoffmann et al., 2010), water cycle (e.g. Sterling et al., 2012), and ecosystem services (e.g. Poppenborg et al., 2013) across different spatial, temporal, and thematic scales. For example, Fu (2003) claimed that more than 60% of the East Asian natural vegetation has been affected by human-induced LULC changes (e.g. forest conversion and desertification). Such (human-induced) LULC changes result in significant changes of ecosystem functions and services at various scales (e.g. local, regional, and global scale).

LULC is a key input for ecosystem services research, decision making and studies on global change in general and influence significantly the outcomes of environmental and ecological models as well as decision making studies (Hansen et al., 2013; Matthews, 1983; Schulp et al., 2011; Vitousek et al., 1997). LULC is recognised as one of the most important spatial data in global initiatives such as the United Nations Framework Convention on Climate Change (UNFCCC) and global organisations such as Food and Agriculture Organization of the United Nations (FAO) and the United Nations Environment Programme (UNEP) (Di Gregorio, 2005; Mora et al., 2014). For example, many studies infer biodiversity information (e.g. habitat type)

indirectly from land cover maps, which is often derived from the remote sensing images (e.g Boyd et al., 2011; Tomaselli et al., 2013). The quality of LULC information is important for these applications – acquisition of appropriate LULC data is an essential issue.

Accurate assessment of LULC and its changes are fundamental factors to sustainable management of natural resources, societal goods and services (Di Gregorio, 2005). Therefore, quantification of LULC is a critical research topic for a wide range of public, private and governmental communities (Müller et al., 2014; Rindfuss et al., 2008). Thus, obtaining appropriate LULC information is critical to secure the quality of the outcomes of the applications using the data (Hansen et al., 2013; Mahecha et al., 2010; Matthews, 1983; Poppenborg et al., 2013; Schulp et al., 2011).

1.1.2 Land use and land cover in cultivated landscapes

Cultivated (managed) landscapes refer to managed vegetated areas where the natural vegetation is replaced by various vegetative LULC types of anthropogenic origin (e.g. dry field crops), livestock grazing, or forestry (Di Gregorio, 2005). Cultivated ecosystems constitute an essential form of human land use. These types of landscapes occupy 34% of the Earth's land areas (Chhabra et al., 2006) and differ greatly from unmanaged landscapes such as natural forest. Land use practices in cultivated landscapes affects functions and services of the embedded agro-ecosystem such as pest control, pollination or control of soil erosion (e.g. Nguyen et al., 2014). An inappropriate land use practice in these type of landscapes can lead to serious damages on those components.

LULC patterns in cultivated landscapes are complex and heterogeneous. In cultivated landscapes, agricultural land use is particularly dominant over any other land use type. Agricultural land uses in a cultivated landscape cause often complex and heterogeneous LULC patterns both spatially and temporally. In spatial aspect, spatial configuration of the agricultural land use is fundamentally artificial and can occur very heterogeneous and complex patterns. For example, a mosaic of crop/non–crop land use (e.g. mixed dry field) can occur in the landscape unlike a homogeneous unmanaged landscape (e.g. natural forest). In temporal aspect, land surface of these agricultural land uses is ceaselessly modified (e.g. tillage and irrigation) and occasionally with no (above ground) vegetation (e.g. harvest) due to constant human management activities. Moreover, these land uses can be converted to different type of land uses in an extremely short time frame (e.g. farmlands conversion). These complex and heterogeneous LULC patterns and their rapid changes in cultivated landscapes are fundamentally affecting the related ecosystem

functions and services at various scales such as local, regional, and global scale.

1.1.3 Global land cover products and its limitations in cultivated landscapes

Despite the significance of LULC information in studies on cultivated(Bartholomé et al., 2005) landscapes, available LULC information is generally limited (Fritz et al., 2013). A data collection of site-specific LULC (i.e. LULC survey) is generally uncommon as it is usually an expensive and laborious task. Instead, pre-existing LULC databases such as satellite-borne global land cover (GLC) products are frequently used as LULC input data in research on cultivated landscapes.

In the last two decades, advancements of remote sensing technologies have supported the derivation of LULC information (Bontemps et al., 2011; De Fries et al., 2010; Defries et al., 1994; Loveland et al., 2000; Mora et al., 2014) and have led to the production of several GLC databases. GLC data provides valuable information about various land systems such as urban, forested, shrubland, and agriculture. It remains a key data source for scientific/non-scientific decision making applications.

Even though GLC remains a key dataset for many applications and studies, existing GLC products have limitations and there are unmatched users' need in the existing GLC datasets (Herold et al., 2008; Mora et al., 2014; Müller et al., 2014). Due to their coarse resolution the GLC products are limited in representing spatial and temporal patterns of LULC, particularly in cultivated landscapes. Such a landscape, especially with frequently changing land use, would not be sufficiently represented by GLC products due to the aforementioned thematically, spatially, and temporally complex nature of the LULC of the landscape.

First, the existing GLC products are limited thematically (i.e. excessively generalised LULC types). Cultivated landscapes are often made up of spatial mosaic of different crop types. In contrast, typically GLC products have few generalised cropland types. Moderate-resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q2) product, for instance, provides two cropland types (Bontemps et al., 2011; Loveland et al., 2000; U.S. Geological Survey, 2012); GlobCover 2000 is provided with two generalised cropland classes. This limited GLC information makes it difficult to monitor crop production, land degradation, and other agriculture associated land use.

Second, the existing GLC products are also limited in spatial resolution as those are coarse raster maps with large cell sizes (e.g. 1 km). Use of a GLC product in complex heterogeneous

landscapes may lead to a poor LULC representation as the LULC mosaic can be smaller than the cell size. Therefore, the GLC products are generally limited in representing mixture classes (i.e. unable to discriminate mixed trees, shrubs, and herbaceous vegetation) (Herold et al., 2008). Inability to deal with small-scaled linear elements (e.g. small streams) could also lead to a substantial misrepresentation of a target landscape.

Third, the GLC products are poor in temporal resolution and imprecise about temporal reference (Thackway et al., 2013). GLC products are commonly unspecific/unclear about temporal reference and have 2-3 years lag between the data acquisition and the releasing date of it. Most GLC data products are released in an irregular interval (e.g. 5–10 years). This is because the LULC data products are released few years after the satellite images were taken. Longitudinal land cover data constitutes an important element especially where land use changes rapidly. However, MODIS Land Cover Type (MCD12Q1) is the only product that provides annual information and is widely used for analysing land cover changes. Consequently, timely new and accurate information is generally lacking in GLC products.

As discussed above, the use of the existing GLC products may be inappropriate in complex heterogeneous agricultural landscapes. Under this circumstance, researchers often inevitably use improperly represented LULC data in their model. If the model is sensitive to LULC input, an inconsistent and imprecise outcome will be produced. Interpretation of the result will be also difficult since the system and its dynamics are poorly described by the model.

There are needs to improve accuracy, stability, spatial resolution, and thematic content of the current GLC datasets (Bontemps et al., 2011; Mora et al., 2014). On one hand, these limitations are due to the low-quality training data (i.e. ground LULC observation) and the input spectral data which are coarse in spatial and temporal resolution (e.g. 500 m 16-day surface reflectance) (Mora et al., 2014; Müller et al., 2014). On the other hand, such limitations may have been unavoidable since the GLC product entails subjective processes such as abstraction, aggregation, classification, and simplification (Comber, 2008; Comber et al., 2005). Nevertheless, more attention is needed to improve accuracy and overall information contents of the existing GLC products.

1.1.4 Towards better LULC quantification in cultivated landscapes

In total, appropriate land cover type information is often unavailable and the use of the current GLC products may be inappropriate in complex agricultural landscapes. Therefore, acquisition of appropriate LULC data is an important issue for research in complex heterogeneous agricultural

landscapes. Also, improvements of GLC products in thematic, spatial, and temporal scales is desired.

To produce better LULC data, one can either use a new high-quality input data (e.g. high-resolution satellite data) product or use a new methodology which can additionally extract information from the existing input data (e.g. medium resolution satellite data). Using new high-resolution data demands an additional campaign (i.e. satellite sensor) and increases computational burden. In contrast, further extracting information from existing satellite data can enrich the information contents with little additional cost. Also it can be applied to the past-time satellite data.

Towards better quantification of LULC with relatively small cost, an appropriate modelling framework for LULC quantification should be developed. To deal with complex heterogeneous agricultural landscapes, such a model development process should thoroughly incorporate (1) high-quality ground truth (i.e. LULC survey) with appropriate meta-information, (2) statistical methods appropriate for the data and the research goal, and (3) a model evaluation scheme to adequately assess model performance and select model and modelling options.

1.2 State-of-the-art and research gaps

This section contains literature review concerning the LULC quantification in cultivated landscapes and the relevant methods. Necessity of the high-quality LULC ground truth data and feasible modelling approaches to expand the volume of the ground truth will be driven from the review.

1.2.1 Remote sensing of LULC and global land cover (GLC) products

Remote sensing of LULC refers to an estimation of LULC types based on the remotely sensed data (e.g. satellite images) using image processing (Anderson et al., 1976; Moody et al., 1995). Estimation of land cover is a common application of remote sensing (Foody et al., 2006). Indeed, remote sensing is an essential tool of land use science as it enables observations over large extents of the Earth. In the last two decades, advancements of remote sensing technologies have supported the derivation of LULC information about various LULC types such as urban, forested, shrubland, and agriculture (Defries et al., 1994) and have nurtured ecosystem research and its applications extensively (Bartholomé et al., 2005; Friedl et al., 2002; Mora et al., 2014).

The first 1 km resolution GLC dataset International Geosphere-Biosphere Programme Data and Information System's GLC map (IGBP – DISCover) was produced for the 1992–1993 period and used in a great variety of applications. Aided by the development of satellite data products, continuous efforts to improve the LULC products are being made (Fritz et al., 2013; Mora et al., 2014) and there are elaborated GLC products available (Herold et al., 2009; Masson et al., 2003; Mora et al., 2014). Currently, the MODIS land cover type, GLC2000, and GlobCover products are available at moderate spatial resolution down to 300 m. GLC-2000 is a global land cover map for year 2000, produced by an international partnership of 30 institutions (Bartholomé et al., 2005). Globcover is a global land cover map for 2005 at 300 m resolution using ENVISAT MERIS data (Bontemps et al., 2011) and adopted FAO Land Cover Classification System (LCCS) to describe land cover types. Annual land cover data is supplied by the MODIS land cover product (MCD12Q1) for the period beginning from 2002. For natural vegetation, higher-resolution surface information databases become globally available (e.g. Hansen et al., 2013; Sexton et al., 2015).

There are currently no global available land cover products on finer than 300 m spatial resolution (Herold et al., 2009). For enhancing GLC products spatially and thematically, GLC mapping projects based on higher-resolution data such as Landsat are being developed by land use science communities (e.g. Chen et al., 2015; Gutman et al., 2012). These new developments aim to provide GLC products with an elaborated information on LULC and overcome the limitations based primarily on 30 m Landsat in combination with high-resolution images such as QuickBird and Worldview-2 (Gutman et al., 2012).

The Land Cover Classification System (LCCS) has been developed by FAO (Food and Agricultural Organization of the United Nations) for a consistent and complete land cover description universally applicable for the whole globe (Di Gregorio, 2005). Using the LC Metadata Language (LCML – LCCS v.3), it describes LULC in a comprehensive and standardised way. It is flexible and allows a dynamic creation of LULC types, which is very useful in heterogeneous landscapes (i.e. users can create own classes by a dynamic combination of land cover attributes). It is also powerful in describing multiple information layers for a single LULC type. The LCCS, as a universal legend definition, has a huge potential in quantifying thematically rich land use and land cover types and there has been thorough LULC quantification studies based on the system either globally (Bartholomé et al., 2005; Bontemps et al., 2011) and locally (Cord et al., 2010).

1.2.2 GLC products in cultivated landscapes

Identification of LULC and its changes in cultivated landscapes is an important issue from regional to global scales (Fritz et al., 2013). However, for cultivated landscapes, acquisition of detailed LULC data is not sufficiently fulfilled by the use of the existing GLC products. While GLC data provides valuable information about various LULC types such as urban, forested, shrubland, and agriculture, however, fine-quality GLC data is untenable in cultivated landscapes due to general inability of GLC products in dealing with heterogeneous agricultural LULC types (Fritz et al., 2013; Herold et al., 2008; Seo et al., 2014). Cultivated landscapes are frequently made up of a spatial mosaic of agricultural land use types. In contrast, the most frequently used global land cover databases like GlobCover or MODIS Land Cover Type contain only few crop-related classes (Bontemps et al., 2011; Loveland et al., 2000; U.S. Geological Survey, 2012). For instance, GlobCover 2000 is provided at 300 m resolution and has four cropland or relevant mixture types, and MODIS Land Cover Type (MCD12Q2) product provides five raster land cover layers at 500 m (Bontemps et al., 2011; Loveland et al., 2000; U.S. Geological Survey, 2012). There are ongoing efforts to extend GLC databases in this context (e.g. Biggs et al., 2006; Gumma et al., 2011; He et al., 2011; Pittman et al., 2010; Potgieter et al., 2007; Wardlow et al., 2007; Wardlow et al., 2008). Enhancement of the quality and usability of the GLC products in cultivated landscapes would be an essential aid to scientific, governmental and non-governmental communities.

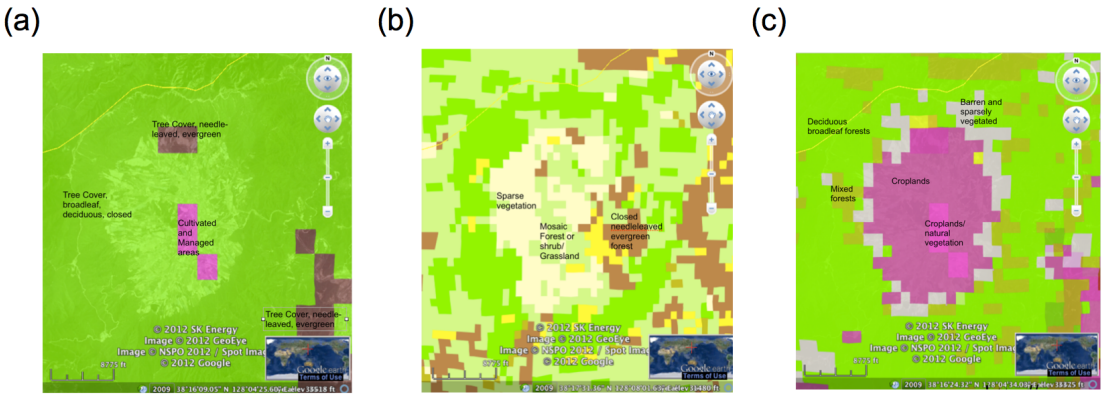

Fig. 1.1 Existing global land use and land cover databases in Haean catchment, South Korea (2009): (a) GLC-2000, (b) GlobCover and (c) MODIS Land Cover Type. LULC becomes overly simplified compared to the real landscape (Figure 1.3). Images courtesy of Geo-Wiki Project (http://geo-wiki.org) (Fritz et al., 2009).

1.2.3 LULC quantification in GLC products

The quantification of LULC is a major application of remote sensing. It is based on images such as satellite imagery, RADAR, LiDAR datasets, and aerial photographies. These images are captured by sensors mounted on satellites, airplanes, and drones. Different data sources and algorithms have been used to map global land cover worldwide. Input data used for global LULC quantification vary from low- to high-resolution in spatial (250 m – 1km), temporal (daily – annual), and spectral resolution (1–15 bands). Despite of its lower-resolution, main observation sensors for the existing global LULC monitoring have been mid to coarse spatial resolution multi-spectral data such as Advanced Very High Resolution Radiometer (AVHRR), MODIS, LANDSAT, SPOT-Vegetation, and MERIS (Masson et al., 2003; Mora et al., 2014). High-resolution datasets such as IKONOS and Quickbird are produced in irregular time interval, which causes difficulty in continuous observation of LULC. In contrast, medium to coarse resolution datasets (> 30 m) such as MODIS are at regular time intervals (e.g. 16-day for Landsat, near-daily for MODIS) (e.g. Doraiswamy et al., 2006; Vittek et al., 2014; Watts et al., 2010).

Most importantly, MODIS datasets are produced on a near daily basis on the entire Earth and play an important role in LULC monitoring (e.g. Doraiswamy et al., 2006; Franklin et al., 2002; Pittman et al., 2010; Thenkabail et al., 2005). Moreover, due to its acquisition interval and composition procedure (NASA Land Processes Distributed Active Archive Center (LP DAAC), 2013a), MODIS 8-day and 16-day products are robust to cloud contamination in monsoonal regions.

GLC products have been developed and validated using varying reference datasets (Bartholomé et al., 2005; Bontemps et al., 2011; Friedl et al., 2010; Sulla-Menashe et al., 2011). For example, the MODIS land cover product is trained using System for Terrestrial Ecosystem Parameterization (STEP) database (Sulla-Menashe et al., 2011) which has approximately 2000 training locations for the whole terrestrial cover (Friedl et al., 2010; Sulla-Menashe et al., 2011). However, in general, global LULC ground truth datasets are still lacking (Herold et al., 2008). The STEP version 6 database includes approximately 500 pixels for cultivated zones (i.e. > 60% agriculture), however, specific crop type information is missing. Instead, five broadly defined crop type classes, namely cereal crop, broadleaf crop, mixed crop, rice, and orchards/vineyards are recorded. These limitations in training data restrict the thematic quality (i.e. simplified agricultural LULC types) in most of the GLC databases.

The lack of training/validation data is partially responsible for the simplified land cover types in

the existing global landcover databases. To enhance the situation, collaborative efforts are being made to expand coverage and increase information contents of the global LULC ground truth databases (e.g. Fritz et al., 2011). Collaborative and open-access mapping of LULC would be useful to develop and validate high-resolution LULc datasets in future. These data can be also useful to regional environmental modelling, ecosystem services research and decision making analysis as high-quality LULC input.

A variety of supervised/unsupervised classification algorithms have been applied to quantify LULC in GLC products (e.g. Herold et al., 2009; Mora et al., 2014, and references therein). For example, the collection 5 MODIS land cover product (MCD12Q1) is based on the decision tree method (Friedl et al., 2002) and Globcover on supervised spatio-temporal clustering. Typically, automated classification procedure are combined with expert opinions from local/regional researchers.

In the recent years, more elaborated machine learning algorithms become popular in LULC quantification as they can handle highly correlated input data (e.g. spectral data) in an explicit way; incorporate data from various sources; deal with mass amount of data and easily amend the missing data. Random Forest (RF) has been used to classify land cover (Clark et al., 2010; Ghimire et al., 2010; Gislason et al., 2006; Hüttich et al., 2009; Nitze et al., 2015; Rodriguez-Galiano et al., 2012; Thenkabail et al., 2005), vegetation type (Hüttich et al., 2009; Immitzer et al., 2012; Senf et al., 2013), and also crop type (Nitze et al., 2015). RF is a decision-tree based ensembling algorithm that uses bootstrap aggregation (bagging) and the random subspace method (Breiman, 2001). Similarly, Support Vector Machines (SVM) have also gained increasing attention (Attarchi et al., 2014; Mountrakis et al., 2011; Vuolo et al., 2012) and used extensively to quantify LULC (e.g. Pal, 2006; Senf et al., 2015; Vuolo et al., 2012). For example, Vuolo et al. (2012) used SVM with MODIS data to evaluate existing GLC products. These two algorithms are comparable in performance to the other state-of-the-art learning algorithms such as neural networks (Attarchi et al., 2014; Gislason et al., 2006; Schwieder et al., 2014).

LULC quantification studies often determine (hyper) parameters of statistical learning algorithms based on literature values or simplified preliminary runs, occasionally without cross-validation (e.g. Dennison et al., 2003; Xiao et al., 2005). However, optimal data-processing options are case-specific (i.e. dependent on the purpose, cost and processing capacities) (Thackway et al., 2013) thus should be site-specifically evaluated. Improperly selected data-processing options can degrade the model performance by reducing information contained in the data.

1.2.4 Fractional LULC regression

Fractional cover is the proportions of non-overlapping land cover types in pixels of a given raster grid (Defries et al., 2000; Price, 1992; Smith et al., 1990). It is defined as the sum of patches covered covered by a land cover type divided by the total area (Asner et al., 2000; Smith et al., 1990). It is also called sub-pixel land cover as it can be conceived as one way to interpret sub-pixel cover labelling (Fernandes et al., 2004). In a satellite image, it is calculated per pixel and ranges from 0 (0% cover) to 1 (100% cover) (Obata et al., 2012). As it contains information for which discrete raster land cover maps, it is increasingly used as a key descriptor of ecosystem and its functions (e.g. Fernandes et al., 2004; Johnson et al., 2012; Pittman et al., 2010; Schwieder et al., 2014; Zhang et al., 2013). For instance, Bevanda et al. (2014) used fractional cover to add structure to land cover for animal habitat modelling.

Similarly, fractional LULC can be defined as the sum of the LULC patch area divided by the total area in each pixel of a given raster grid (Fernandes et al., 2004). Estimating fractional LULC from available coarse resolution satellite data can be a useful strategy (e.g. Schwieder et al., 2014). There have been studies intended to retrieve LULC fractions from spectral data (e.g. Colditz et al., 2011; Guerschman et al., 2009; Obata et al., 2012) and continuous efforts to derive fractional land cover information from existing satellite data (e.g. Defries et al., 2000). Nevertheless, fractional LULC as continuous LULC representation, especially with multiple land cover types is still underdeveloped.

Fractional cover regression can be implemented via various techniques. The techniques include the fuzzy classifier (Foody et al., 1996), the time series model (Lu et al., 2003), linear models (DeFries et al., 1995; Schwarz et al., 2005), data mining algorithms (Fernandes et al., 2004; Schwieder et al., 2014), and spectral mixture analysis (Asner et al., 2000; Guerschman et al., 2009). Spectral mixture analysis (SMA) has been frequently used in fractional cover studies using spectral data (Obata et al., 2012). In this approach, mixed spectral signals are decomposed into spectral endmembers and by which sub-pixel fractions of land cover types are estimated (Guerschman et al., 2009; Lobell et al., 2004; Obata et al., 2012). However, the SMA approach generally favours hyperspectral data over multi-spectral data (i.e. MODIS reflectance data) (Asner et al., 2000; Guerschman et al., 2009), which is still deficient at the global scale. Moreover, the method is under the assumption that there are linear relationships between the area fractions of spectral sources (e.g. land cover types) and spectral signals (e.g. surface reflectances) (Asner et al., 2000; Lobell et al., 2004; Xiao et al., 2005). This assumption is violated when non-linear functions such as NDVI or EVI are used as predictor (Lobell et al., 2004).

Instead, there are studies using RF to quantify fractional cover (e.g. Colditz et al., 2011; Guerschman et al., 2009; Lu et al., 2003; Obata et al., 2012; Schwieder et al., 2014). RF can deal with a large number of highly correlated features (e.g. spectral data) and non-linear relationships (Immitzer et al., 2012) as it tends not to overfit the data (Breiman, 2001; Segal, 2004). Moreover it is convenient to set up compared to other data mining algorithms as it has a small number of training parameters (Liaw et al., 2002).

1.2.5 Multi-crop LULC classification

Quantifying multi-crop LULC is a multinomial classification task. In cultivated ecosystems, LULC data type labels are often imbalance since, when aggregated, minor LULC types occupy a substantial portion in this type of landscape. This cause data imbalance when organised for LULC classification using statistical learning techniques (i.e. classification and regression algorithms). In this case, training data sets are imbalanced.

Generally learning algorithms require balanced training data (e.g. Chawla et al., 2002; Fernández et al., 2011). For instance, support vector machine (SVM) assumes training dataset is balanced and known to be biased to major types otherwise (Akbani et al., 2004). Therefore, under a data imbalance, rare or minor LULC types are more difficult to classify. This can be avoided by doing a binary classification (e.g. vegetation and non-vegetation) via reclassification of the data. However, as indicated, this imbalance may be a major challenge for multi-crop LULC mapping which inevitably incorporates many LULC types including presumably minor LULC types such as crop species.

In general, there are three major ways to cope with imbalanced data sets. The first is to adapt the classification algorithm to reinforce learning of the minor classes (e.g. Bruzzone et al., 1997; Williams et al., 2009). The second is to adjust the classifier by assigning different costs to misclassification in rare versus frequent classes (e.g. Sun et al., 2007). The third is by re-sampling the data set (e.g. García et al., 2011; He et al., 2009; Waske et al., 2009, and references therein). This last approach has the advantage to be independent from the classifier used.

Oversampling of the rare classes with replacement or undersampling of the major class have been discussed by several authors (Japkowicz et al., 2002; Ling et al., 1998; Schistad Solberg et al., 1996). However, the potential of these approaches to improve the classification accuracy of rare classes seems to be limited. In particular random oversampling with replacement can lead to overfitting (Chawla, 2010).

To overcome the issue of overfitting, Chawla et al. (2002) proposed to generate new minority

instances by a synthetic minority oversampling technique (SMOTE) instead of oversampling with replacement. They reported that the synthetic points created by SMOTE forced the classifier to learn larger and less specific regions and thus changed the boundaries between classes. SMOTE performs better than oversampling the minority class by replacement and can be combined with undersampling of the majority class.

There are other obstacles in classifying multi-crop LULC from spectral data. First, spectral characteristics of LULC types are often altered by the spatial and temporal mixture of LULC types. Second, vegetation development phase varies by socio-economic factors (e.g. cropping schedule) as well as natural factors (e.g. climatic and topographic conditions), hence spectral characteristics are highly heterogeneous even within a single type. In addition, the use of coarse spatial resolution datasets (e.g. 250 m) induces the presence of mixed LULC types in one pixel especially in transition zones (Foody et al., 1996). Since the pixel size of data from many remote sensing systems is relatively large, many pixels are of mixed in LULC composition.

1.2.6 Research gaps and objectives

Although the remote sensing is widely used to retrieve LULC information for the globe, its practical implementation does not suffice for the need of LULC information in heterogeneous agricultural landscapes due to the limitation of the data, method and its operational difficulties.

- Lack of detailed ground observation

  Although LULC distribution and its change over time is essential, detailed LULC observation data is scarce or even non-existent especially for complex agricultural landscapes. It restricts building a statistical model to estimate LULC information. Hence, collection of detailed LULC data is necessary for a complex heterogeneous landscape.

- Deficit of continuous representation of LULC

  Available global LULC data is coarse raster maps and continuous representation of LULC is still lacking. It is primarily due to the source remote sensing data is coarse in spatial resolution. Developing a model retrieving continuous LULC from coarse remote sensing data would be a useful strategy.

- Deficit of multi-crop LULC

  However, albeit acclaimed theoretically, the concept is still far away from being incorporated routinely into practical decision-making. In cultivated ecosystems, crop LULC types are often important as they have significant impact on the system. LULC classification

with multiple crop types is still underdeveloped. Most of the existing LULC databases use lumped agricultural classes (e.g. “croplands”) and lack detailed information such as crop species. However, model-based estimation of multi-crop LULC is challenging due to severe imbalance in distribution of crop-type LULC types (i.e. majority types dominating over minor types) because most standard classifiers assume a balanced distribution of training data.

The goal of this dissertation is to enhance LULC mapping in a complex agricultural landscape of South Korea. Specifically, extraction of spatially (i.e. continuous representation) and thematically (i.e. multi-crop types) detailed LULC information from existing, medium resolution, multi-spectral satellite products are pursued. Regarding the aforementioned research questions and research gaps, three objectives were formulated.

1. Collection of detailed ground LULC observation

   To establish better LULC models for complex heterogeneous landscapes, high-quality observation data is essential. During a field campaign, a high-quality ground LULC data was collected over the entire study area. In chapter 2, the observed data is introduced and compared with a GLC product for the three-year study period.

2. Modelling of continuous LULC representation

   To obtain spatially improved LULC representation, extraction of continuous LULC representation can be a good approach. In chapter 3, extraction of fractional LULC as continuous LULC representation is pursued. A Random Forest (RF) regression model was developed to extract fractional LULC from satellite products. To attain optimal performance of the model, various data processing options were tested and chosen based on its impact on the performance.

3. Classification of multi-crop LULC

   To thematically improve LULC mapping, LULC classification with multi-crop LULC data is important yet underdeveloped. Often unequal distribution of LULC types prevents classifiers from successfully recognising minor LULC types. In chapter 4, multi-crop LULC classification using Support Vector Machine (SVM) and Random Forest (RF) is presented. A synthetic sampling technique was applied to improve the classification performance of minor types by artificially balancing training data.

The three objectives altogether intend to find lessons and strategies to enhance LULC quantification in agricultural landscapes and consequently global land use and land cover databases. In order to soundly relate the three objectives in an integrated manner, this dissertation pursues to design a combined framework covering the three topics. The data obtained in the course of

achieving the first goal is used as training data in the next modelling studies. The limitations of the existing GLC products are revealed and clarified in the first study and are methodologically addressed afterwards. Each of the LULC modelling study is dealing with one specific limitation of the existing GLC products (i.e. coarse spatial resolution and limited thematic resolution). The general underlying data rarity/imbalance problem of this LULC studies is addressed in the last study by means of a data resampling technique.

1.3 Concept of the dissertation

Corresponding to the research gaps and objectives, the dissertation includes three major components: high-quality LULC data collection, continuous representation of LULC, and classification of multi-crop LULC. These three components are presented in the next three chapters.
In chapter 2, the ground LULC census data from the field campaign is introduced. The observed LULC data and the collection and the processing strategies are illustrated. Additionally, the census data is compared with the MODIS Land Cover Type product (MCD12Q1). In chapter 3, a data mining model retrieving continuous representation of LULC is presented. It seeks after feasible strategies to achieve spatially and thematically improved LULC representation based on existing satellite products. In chapter 4, to thematically improve LULC representation, a multi-crop LULC classification model is presented. In the two modelling chapters, two statistical learning techniques (i.e. RF and SVM) and various data-processing techniques (e.g. SMOTE) are used to improve the performance of the LULC quantification models.

Deriving a per-field land use and land cover map in an agricultural mosaic catchment (Chapter 2)

During three years of field campaign (2009–2011), LULC of the study area was thoroughly surveyed. In this chapter, the census data, its collection strategy and post-processing protocol are introduced. The raw LULC information and various reclassified class labels are provided with meta information.
The collected data is transformed into a thematically and spatially rich LULC data for the area. The dataset is a 'per-field' data as the unit entity of the dataset is a polygon corresponding to an actual land parcel. Based on the field observation, the polygons are associated with ecological and physical traits. The dataset is available at the public repository Pangaea (Seo et al., 2014).
Furthermore, a comparison between the census data and the MODIS Land Cover Type product

(MCD12Q1) was made to quantitatively evaluate available land use and land cover products. This comparison would clarify the potential and limitations of the GLC products in the agricultural landscape.

This type of vector-form data should be produced extensively to develop high-resolution LULC datasets. The detailed crop type information described with FAO-LCCS can be used for regional environmental modelling as well as for ecosystem services research and decision making analysis. The chapter is published at the data journal Earth System Science Data (Seo et al., 2014).

Mapping Fractional Land Use and Land Cover in a Monsoon Region: The Effects of Data Processing Options (Chapter 3)

In chapter 3, an empirical model for deriving continuous representation of LULC is developed. We hypothesised that, with proper methods, existing satellite products can be a useful source of information about LULC at sub-pixel level. In that context, a fractional cover regression for 10-type system was carried out based on a MODIS multi-spectral data (MODIS 13Q1). Based on the regression model, an evaluation framework was created for making informed decisions in choosing data-processing options. By using the framework the effect of three key data-processing options were evaluated by its impact on the regression performance: selection of spectral predictor sets (NDVI, EVI, surface reflectance, and all combined), time interval (8-day vs. 16-day), and smoothing (no smoothing vs. Savitzky-Golay filter). The mechanism affecting the model performance was investigated by looking at the correlation between observed and predicted LULC fractions. Additionally, relative importance of the spectral bands and the data acquisition dates were estimated. Through the study, cross-validation was used to rigorously calibrate and evaluate the models. The chapter has been submitted to IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Improving the classification of rare land use and land cover types using synthetic data (Chapter 4)

LULC classification with multiple crop types is yet under-developed. One of the reasons is data imbalance. LULC data tends to be imbalanced with majority types dominating over minor types. Generally speaking, data mining algorithms perform best on equally distributed training data. However, most standard classifiers implicitly assume a balanced distribution of LULC types. Therefore, they fail to detect minor LULC types under severe imbalance. When imbalanced, the learning of the minor LULC types will be hampered as well as the overall performance. In cultivated ecosystems, minority types are often more important than major types as they might

indicate changes of LULC. In chapter 4, the synthetic oversampling technique (SMOTE) is applied to the training LULC data. It balances training data, hence improve model performance. Among various approaches to cope with imbalanced data sets (e.g. cost sensitive learning), a resampling approach was preferred as it is advantageous as being independent from the used learning algorithms. The goal of the study was to improve the classification of rare LULC types in an agricultural mosaic catchment by using SMOTE. Support Vector Machine (SVM) and Random Forest (RF) classifiers were used to classify a multi-majority and multi-minority dataset using the MODIS spectral product (MOD13Q1). Four scenarios were formulated to reveal the effect of SMOTE. The mechanism affecting the model performance was investigated as well. The chapter is under review in the journal ISPRS Journal of Photogrammetry and Remote Sensing.

Overall concept and the guiding hypothesis

Overall, this dissertation seeks after how to better quantify LULC of a complex agricultural landscape primarily relying on globally available multi-spectral satellite data. For guiding this line, one working hypothesis is developed: the combination of high-quality LULC data, recent machine learning techniques, and data-processing techniques can substantially increase the amount of LULC information we can extract from the medium resolution satellite products. This main hypothesis is pursued because GLC products are mostly based on the same type of data (e.g. global satellite data) and similar algorithms (e.g. decision trees). Thus, retrieving detailed LULC information from existing satellite products would be a convenient way to obtain new information. An overview of the dissertation work is shown in 1.2.

These three chapters are written as part of this cumulative dissertation. Before going into the main chapters, a description of the study site will be given. At the end of the dissertation, a synopsis will be provided to summarise the results and discuss current capacity and potential of the LULC science for complex heterogeneous landscapes.

All the three studies were conducted within and supported by the International Research Training Group between Germany and South Korea (DFG/KOSEF, Complex TERRain and ECOlogical Heterogeneity - TERRECO, GRK 1565/1). In the overarching project TERRECO, this dissertation work aimed to supply LULC information for the other research projects evaluating various ecosystem functions and ecosystem services.

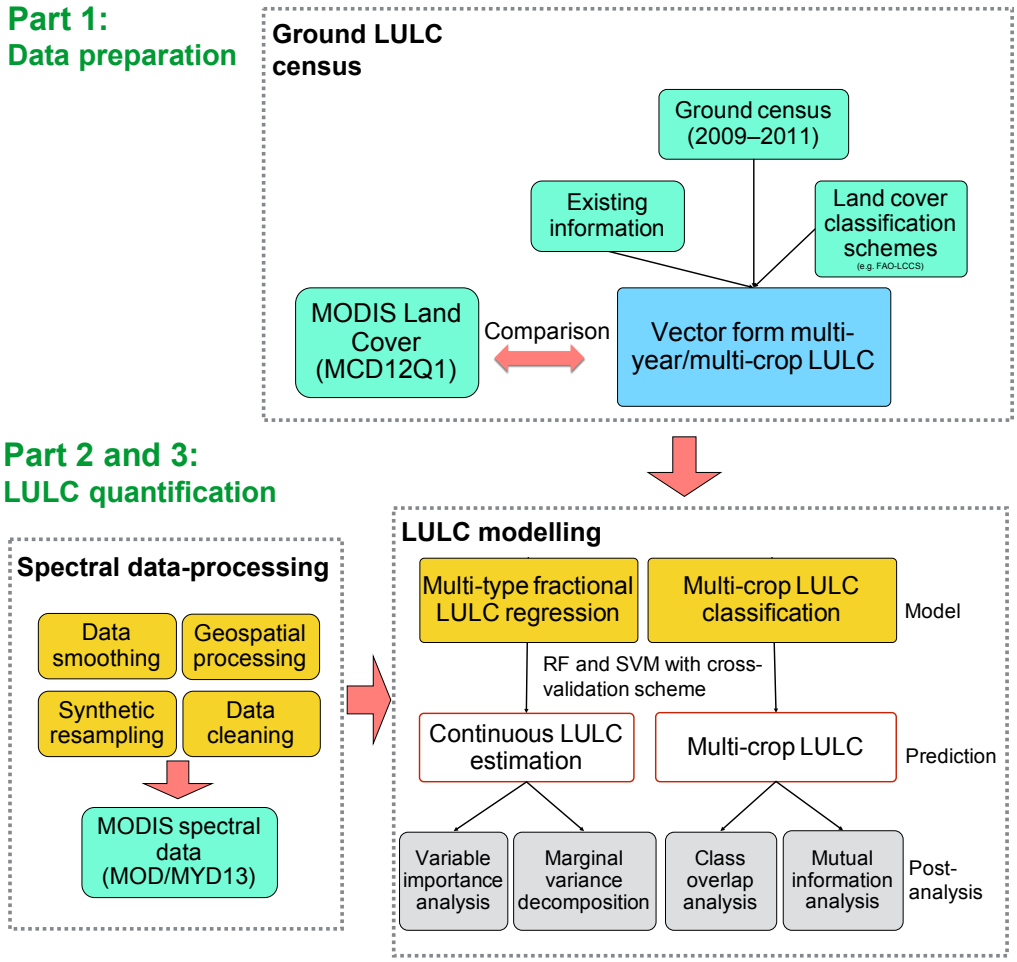

Fig. 1.2 Structure of the disseration and connections of different parts

1.4 Study site

The study site Haean-myeon is located at the border between North and South Korea (128°1′33.101″E, 38°28′6.231″N, 450 - 1200 m a.s.l.) (Figure 1.3). It is a small heterogeneous agricultural catchment (64.4 km^2) consisting of 58% of surrounding forested area, 30% of agricultural area in the centre, and 12% of the remaining area as residential and other semi natural LULC area such as riparian and farm road area.

Due to its characteristic bowl shape, the land use changes from predominantly rice paddies at the valley bottom to dry field farming on moderate slopes. The higher altitudes are covered by deciduous and mixed forests. The agricultural area in the centre is characterised as a mosaic of non-crop and crop patches (Figure 1.4). The LULC types are unevenly distributed due to its topography with an imbalance ratio up to 100 : 1 (Figure 1.5).

The upland forest is at higher elevations, predominately composed of mixed deciduous oak (Quercus spp.) and maple forest. Other major species include the Japanese Red Pine (Pinus densiflora), Japanese Larch (Larix leptolepis), Pitch Pine (Pinus rigida), Korean Pine (Pinus koraiensis),

Japanese Cedar (Cryptomeria japonica), Japanese Cypress (Chamaecyparis obtuse), Japanese Chestnut (Castanea crenata). Five dominant weeds in the crop fields were Digitaria sanguinalis, Conyza canadensis var. canadensis, Cyperus difformis, Cyperus orthostachyus and Cyperus amuricus. Weeds such as Artemisia montana, Elsholtzia splendens, Aster scaber and Persicaria nepalensis that are typically found in mountainous areas and such as Conyza canadensis var. canadensis, Senecio vulgaris, Aster pilosus and Bidens frondosa were also found.

The study region belongs to East Asian summer monsoon (EASM) region (Yihui et al., 2005) and shows persistent and intensive raining period in Summer. This period is called "Changma" (long lasting rain) in Korean literature (Kang et al., 2009). The average air temperature of the study area is 8.5° C at the central plateau (Korean Meteorological Administration, http://web.kma.go.kr/eng). The annual average rainfall equals 1599 mm and the maximum daily rainfall was 223 mm between 1999 and 2010. Due to the raining period in which more than 60% of annual precipitation is concentrated and extreme rainfall events occur frequently, acquisition of cloud-free spectral data during summer is generally difficult (Guerschman et al., 2009; Yihui et al., 2005).

The catchment is located in the Soyang river watershed, South Korea (127′43″ to 128′35″ E and 37′41″ to 38′29″ N). Total 85% of the watershed is covered by deciduous forest but with a few agricultural hotspots. There was a severe downstream water quality degradation by the agricultural activity occurring in the catchment (Meusburger et al., 2013; Shope et al., 2014). Accordingly, a series of policy measures including land use conversion were initiated by the local government in 2008 (Jun et al., 2010). The land use conversion policy aimed to reduce soil erosion by converting annual dry field crops to perennial crops such as "ginseng" by subsidising perennial crops. This policy caused rapid LULC changes in land use and land cover (Seo et al., 2014).

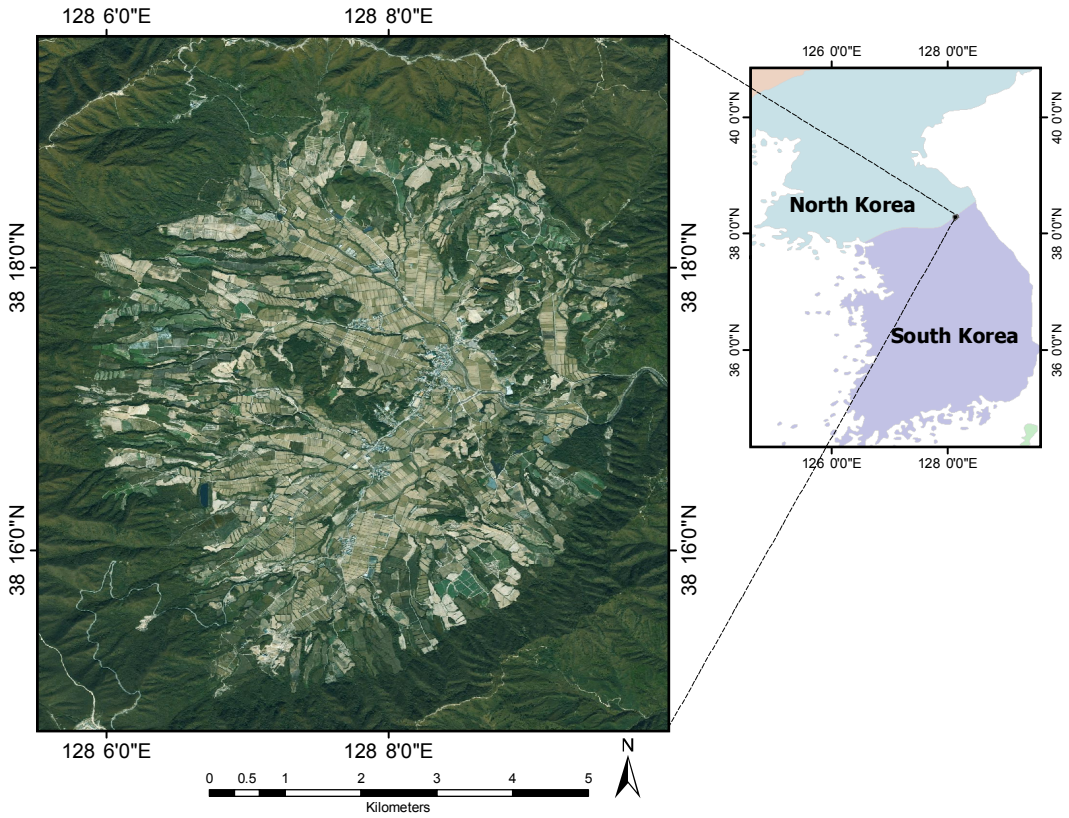

Fig. 1.3 Map and the location of the study site 'Haean' on the Korean peninsula. The catchment is an agricultural hotspot located in the protected forested watershed. Satellite image a SPOTMaps mosaic product (Astrium Services, http://www.astrium-geo.com) acquired in 2009.

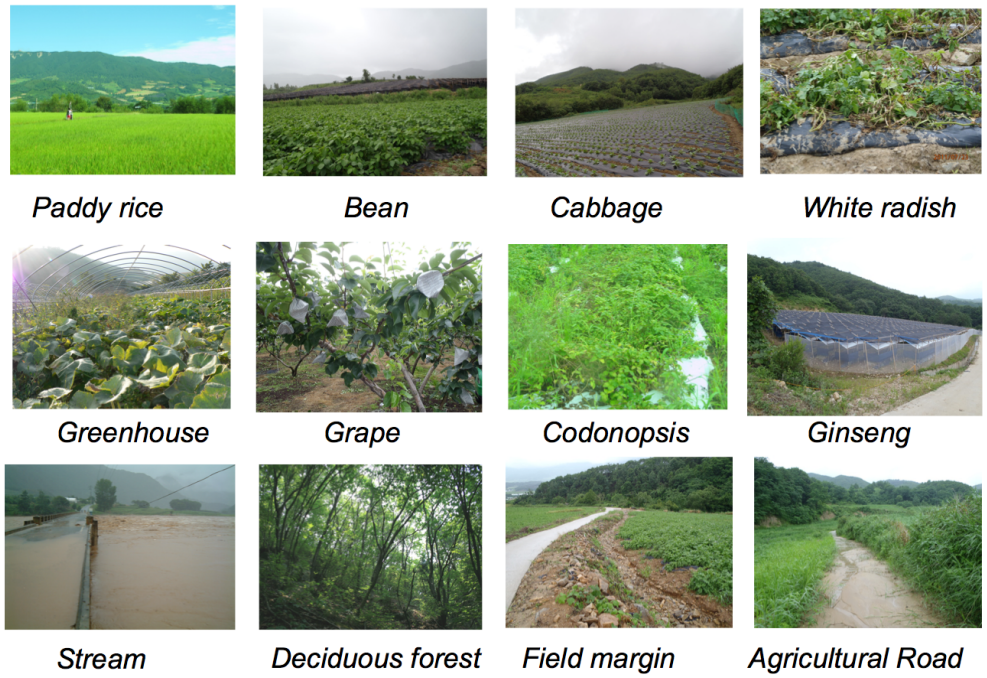

Fig. 1.4 Pictures of the observed LULC types taken during the three-year study period (2009–2011). In the relatively small study area, a huge variety of crop/non-crop LULC types occurred. By means of technical and financial aids such as strong subsidisation, the local management promoted alternative crops such as ginseng and orchards which caused rapid changes in LULC.

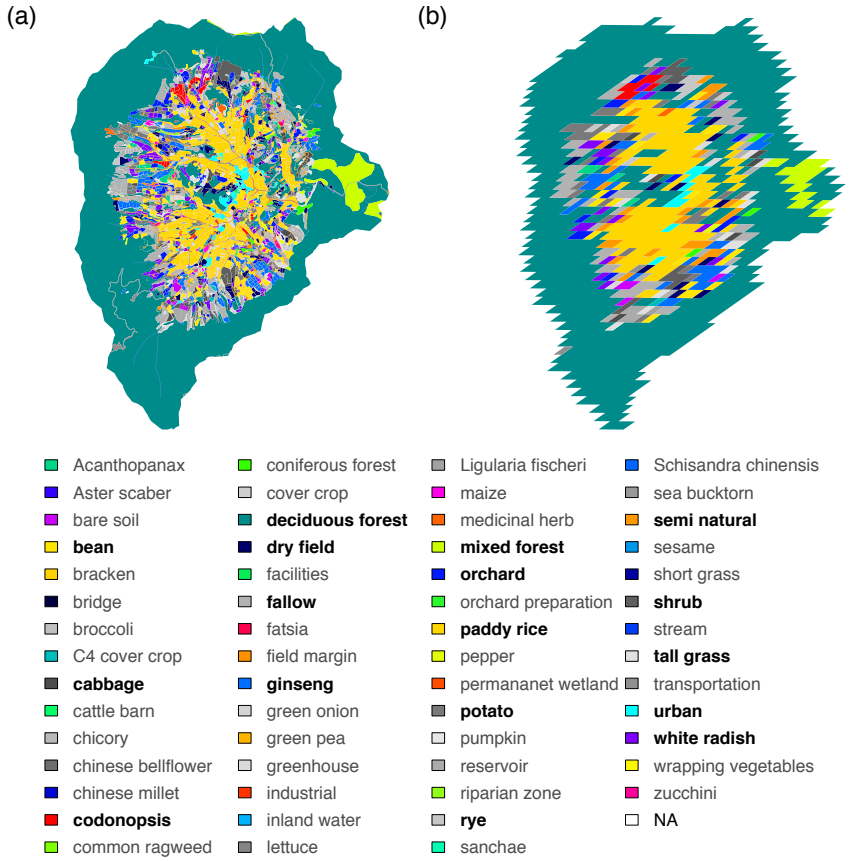

Fig. 1.5 Land use and land cover of the Haean catchment surveyed in 2010. (a) Original polygon data with 59 LULC types and (b) rasterised LULC upon the MODIS sinusoidal grid (H28V5) with 28 remained types after rasterisation. The LULC types are according to the classification scheme S1 of the original survey data and the names in bold indicate the dominant LULC types (Seo et al., 2014).

1.5 Record of contributions to this thesis

The three studies described in this thesis refer to three manuscripts. The first manuscript is published at Earth System Science Data, the second manuscript is submitted to IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, and the third to ISPRS Journal of Photogrammetry and Remote Sensing. The following list specifies the contributions of the individual authors to each manuscript.

Manuscript 1 (Chapter 2):

Authors: Bumsuk Seo, Christina Bogner, Patrick Poppenborg, Emily Martin, Mathias Hoffmeister, Mansig Jun, Thomas Koellner, Björn Reineking, Christopher Shope, and John Tenhunen

Title: Deriving a per-field land use and land cover map in an agricultural mosaic catchment

Status: Published (2014)

Journal: *Earth System Science Data*

Contributions:

Bumsuk Seo	60%, idea, data collection, methods, data analysis, figures, tables, manuscript writing, discussion, manuscript editing, corresponding author
Christina Bogner	20%, idea, methods, data analysis, figures, tables, manuscript writing, discussion, manuscript editing
Patrick Poppenborg	2%, data collection
Emily Martin	3%, idea, discussion, data collection
Mathias Hoffmeister	2%, data collection
Mansig Jun	2%, idea, data collection, discussion
Thomas Koellner	2%, idea, discussion
Björn Reineking	2%, idea, data collection, discussion, manuscript editing
Christopher Shope	2%, idea, data collection
John Tenhunen	5%, idea, data collection

Dataset (Chapter 2):

Authors: Bumsuk Seo, Patrick Poppenborg, Emily Martin, Mathias Hoffmeister, Christina Bogner, Hamada Elsayed Ali, Björn Reineking and John Tenhunen

Title: Per-field land use and land cover data set of the Haean catchment, South Korea

Status: Published (2014)

Publisher: *PANGAEA*

Contributions:

Bumsuk Seo	65%, idea, data collection, data structuring, figures, metadata writing, data editing, corresponding author
Patrick Poppenborg	5%, data collection
Emily Martin	5%, idea, discussion, data collection
Mathias Hoffmeister	5%, data collection
Christina Bogner	5%, data structuring, figures, metadata writing
Hamada Elsayed Ali	5%, data collection
Björn Reineking	5%, idea, data collection, discussion
John Tenhunen	5%, idea, data collection

Manuscript 2 (Chapter 3):

Authors: Bumsuk Seo, Christina Bogner, Thomas Koellner, and Björn Reineking

Title: Mapping Fractional Land Use and Land Cover in a Monsoon Region: The Effects of Data Processing Options

Status: Submitted

Journal: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*

Contributions:

Bumsuk Seo	75%, idea, data collection, methods, data analysis, figures, tables, manuscript writing, discussion, manuscript editing, corresponding author
Christina Bogner	10%, idea, methods, data analysis, figures, tables, discussion, manuscript editing
Thomas Koellner	5%, idea, discussion
Björn Reineking	10%, idea, methods, data analysis, figures, tables, discussion, manuscript editing

Manuscript 3 (Chapter 4):

Authors: Christina Bogner, Bumsuk Seo, and Björn Reineking

Title: Improving the classification of rare land use and land cover types using synthetic data

Status: Under review

Journal: *ISPRS Journal of Photogrammetry and Remote Sensing*

Contributions:

Christina Bogner	50%, idea, methods, data analysis, figures, tables, manuscript writing, discussion, manuscript editing, corresponding author
Bumsuk Seo	45%, idea, data collection, methods, data analysis, figures, tables, manuscript writing, discussion, manuscript editing
Björn Reineking	5%, idea, methods, discussion, manuscript editing

References

Akbani, R., S. Kwek & N. Japkowicz (2004). "Applying support vector machines to imbalanced datasets". In: *Machine Learning: ECML 2004*. Springer, pp. 39–50 (cit. on pp. 12, 107).

Anderson, J. R., E. E. Hardy, J. T. Roach & R. E. Witmer (1976). *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*. Vol. 964. US Government Printing Office (cit. on p. 6).

Asner, G. & D. B. Lobell (2000). "A biogeophysical approach for automated SWIR unmixing of soils and vegetation". In: *Remote Sensing of Environment* 74.1, pp. 99–112 (cit. on pp. 11, 67, 72, 157).

Attarchi, S. & R. Gloaguen (2014). "Classifying complex mountainous forests with L-Band SAR and Landsat data integration: A comparison among different machine learning methods in the hyrcanian forest". In: *Remote Sensing* 6.5, pp. 3624–3647 (cit. on pp. 10, 72, 107).

Bartholomé, E & A. S. Belward (2005). "GLC2000: a new approach to global land cover mapping from Earth observation data". In: *International Journal Of Remote Sensing* 26.9, pp. 1959–1977 (cit. on pp. 4, 6, 7, 9, 37).

Bevanda, M., N. Horning, B. Reineking, M. Heurich, M. Wegmann & J. Mueller (2014). "Adding structure to land cover - using fractional cover to study animal habitat use". In: *Movement Ecology* 2.1, p. 26 (cit. on pp. 11, 63).

Biggs, T. W., P. S. Thenkabail, M. K. Gumma, C. A. Scott, G. R. Parthasaradhi & H. N. Turral (2006). "Irrigated area mapping in heterogeneous landscapes with MODIS time series, ground truth and census data, Krishna Basin, India". In: *International Journal Of Remote Sensing* 27.19, pp. 4245–4266 (cit. on pp. 8, 56, 64).

Bontemps, S., P. Defourny, E. Bogaert, O. Arino, V. Kalogirou & J. Perez (2011). *GLOBCOVER 2009 - Products Description and Validation Report*. Tech. rep. European Space Agency (cit. on pp. 4, 5, 7–9, 38, 63, 64, 105).

Boyd, D. S. & G. M. Foody (2011). "An overview of recent remote sensing and GIS based research in ecological informatics". In: *Ecological Informatics* 6.1, pp. 25–36 (cit. on p. 3).

Breiman, L (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32 (cit. on pp. 10, 12, 72, 73, 76, 107).

Bruzzone, L. & S. B. Serpico (1997). "Classification of imbalanced remote-sensing data by neural networks". In: *Pattern Recognition Letters* 18.11, pp. 1323–1328 (cit. on pp. 12, 106).

Chawla, N. V. (2010). "Data mining for imbalanced datasets: An overview". In: *Data Mining and Knowledge Discovery Handbook*. Springer, pp. 875–886 (cit. on pp. 12, 106).

Chawla, N. V., K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357 (cit. on pp. 12, 106, 111, 118).

Chen, J., J. Chen, A. Liao, X. Cao, L. Chen, X. Chen, C. He, G. Han, S. Peng, M. Lu, W. Zhang, X. Tong & J. Mills (2015). "Global land cover mapping at 30m resolution: A POK-based operational approach". In: *Isprs Journal of Photogrammetry and Remote Sensing* 103.C, pp. 7–27 (cit. on p. 7).

Chhabra, A., H. Geist, R. A. Houghton, H. Haberl, A. K. Braimoh, P. L. Vlek, J. Patz, J. Xu, N. Ramankutty, O. Coomes & others (2006). "Multiple impacts of land-use/cover change". In: *Land-use and land-cover change*. Springer, pp. 71–116 (cit. on pp. 2, 3, 105).

Clark, M. L., T. M. Aide, H. R. Grau & G. Riner (2010). "A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America". In: *Remote Sensing of Environment* 114.11, pp. 2816–2832 (cit. on pp. 10, 64, 72, 106).

Colditz, R. R., M Schmidt, C Conrad, M. C. Hansen & S Dech (2011). "Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions". In: *Remote Sensing of Environment* 115.12, pp. 3264–3275 (cit. on pp. 11, 12, 37, 38, 64, 71, 84, 85).

Comber, A. J. (2008). "The separation of land cover from land use using data primitives". In: *Journal of Land Use Science* 3.4, pp. 215–229 (cit. on pp. 2, 5).

Comber, A., P. Fisher & R. Wadsworth (2005). "What is land cover?" In: *Environment and Planning B: Planning and Design* 32.2, pp. 199–209 (cit. on p. 5).

Cord, A, C Conrad, M Schmidt & S Dech (2010). "Standardized FAO-LCCS land cover mapping in heterogeneous tree savannas of West Africa". In: *Journal of Arid Environments* 74.9, pp. 1083–1091 (cit. on pp. 7, 154).

Dawson, T. P., S. T. Jackson, J. I. House, I. C. Prentice & G. M. Mace (2011). "Beyond predictions: biodiversity conservation in a changing climate". In: *Science* 332.6025, pp. 53–58 (cit. on pp. 2, 105).

De Fries, R. S., M Hansen, J. R. G. Townshend & R Sohlberg (2010). "Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers". In: *International Journal Of Remote Sensing* 19.16, pp. 3141–3168 (cit. on p. 4).

Defries, R. S. & J. R. G. Townshend (1994). "NDVI-derived land cover classifications at a global scale". In: *International Journal Of Remote Sensing* 15.17, pp. 3567–3586 (cit. on pp. 4, 6).

Defries, R. S., M. C. Hansen & J. Townshend (2000). "Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8 km AVHRR data". In: *International Journal Of Remote Sensing* 21.6-7, pp. 1389–1414 (cit. on pp. 11, 63, 64, 157).

DeFries, R. S., C. B. Field, I. Fung, C. O. Justice, S. Los, P. A. Matson, E. Matthews, H. A. Mooney, C. S. Potter, K. Prentice, et al. (1995). "Mapping the land surface for global atmosphere-biosphere models: Toward continuous distributions of vegetation's functional properties". In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 100.D10, pp. 20867–20882 (cit. on pp. 11, 72, 157).

Dennison, P. & D. Roberts (2003). "Endmember selection for multiple endmember spectral mixture analysis using endmember average RMSE". In: *Remote Sensing of Environment* 87, pp. 123–135 (cit. on pp. 10, 85).

Di Gregorio, A (2005). *Land Cover Classification System: Classification Concepts and User Manual: LCCS*. Rome (Italy). Food and Agriculture Organization of the United Nations (FAO) (cit. on pp. 2, 3, 7, 43, 52, 56, 154).

Doraiswamy, P., B Akhmedov & A. Stern (2006). "Improved Techniques for Crop Classification using MODIS Imagery". In: *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on*, pp. 2084–2087 (cit. on p. 9).

Fernandes, R., R. Fraser, R. Latifovic, J. Cihlar, J. Beaubien & Y. Du (2004). "Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region". In: *Remote Sensing of Environment* 89.2, pp. 234–251 (cit. on pp. 11, 37, 63, 68, 72, 75, 85, 157).

Fernández, A., S. García & F. Herrera (2011). "Addressing the classification with imbalanced data: open problems and new challenges on class distribution". In: *Hybrid Artificial Intelligent Systems*. Springer, pp. 1–10 (cit. on pp. 12, 106).

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs & others (2005). "Global consequences of land use". In: *Science* 309.5734, pp. 570–574 (cit. on pp. 2, 105).

Foody, G & P Atkinson (2006). "Current Status of Uncertainty Issues in Remote Sensing and GIS". In: *Uncertainty in Remote Sensing and GIS*. John Wiley & Sons, Ltd. Chap. 17, pp. 287–302 (cit. on p. 6).

Foody, G. M. & M. K. Arora (1996). "Incorporating mixed pixels in the training, allocation and testing stages of supervised classifications". In: *Pattern Recognition Letters* 17.13, pp. 1389–1398 (cit. on pp. 11, 13, 72, 157).

Franklin, S. E. & M. A. Wulder (2002). “Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas”. In: *Progress in Physical Geography* 26.2, pp. 173–205 (cit. on p. 9).

Friedl, M. A., D. K. McIver, J. Hodges & X. Y. Zhang (2002). “Global land cover mapping from MODIS: algorithms and early results”. In: *Remote Sensing of Environment* 83.1-2, pp. 287–302 (cit. on pp. 6, 10, 44, 54).

Friedl, M. A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley & X. Huang (2010). “MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets”. In: *Remote Sensing of Environment* 114.1, pp. 168–182 (cit. on pp. 9, 37, 43, 54, 158).

Fritz, S., I. McCallum, C. Schill, C. Perger, R. Grillmayer, F. Achard, F. Kraxner & M. Obersteiner (2009). “Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover”. In: *Remote Sensing* 1.3, pp. 345–354 (cit. on p. 8).

Fritz, S., I. McCallum, C. Schill, C. Perger, L. See, D. Schepaschenko, M. van der Velde, F. Kraxner & M. Obersteiner (2011). “Geo-Wiki: An online platform for improving global land cover”. In: *Environmental Modelling and Software*, pp. 1–14 (cit. on pp. 10, 158).

Fritz, S., L. See, L. You, C. Justice, I. Becker Reshef, L. Bydekerke, R. Cumani, P. Defourny, K. Erb, J. Foley, S. Gilliams, P. Gong, M. Hansen, T. Hertel, M. Herold, M. Herrero, F. Kayitakire, J. Latham, O. Leo, I. McCallum, M. Obersteiner, N. Ramankutty, J. Rocha, H. Tang, P. Thornton, C. Vancutsem, M. Velde, S. Wood & C. Woodcock (2013). “The Need for Improved Maps of Global Cropland”. In: *Eos, Transactions American Geophysical Union* 94.3, pp. 31–32 (cit. on pp. 4, 7, 8, 38).

Fu, C. (2003). “Potential impacts of human-induced land cover change on East Asia monsoon”. In: *Asia Monsoon Environment System and Global Change* 37.3–4, pp. 219–229 (cit. on p. 2).

García, V., J. S. Sánchez & R. Mollineda (2011). “Classification of high dimensional and imbalanced hyperspectral imagery data”. In: *Pattern Recognition and Image Analysis*. Ed. by J. Vitrià, J. M. Sanches & M. Hernández. Vol. 6669. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 644–651 (cit. on pp. 12, 106).

Ghimire, B, J Rogan & J Miller (2010). “Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic”. In: *Remote Sensing Letters* 1.1, pp. 45–54 (cit. on pp. 10, 72).

Gislason, P. O., J. A. Benediktsson & J. R. Sveinsson (2006). “Random Forests for land cover classification”. In: *Pattern Recognition Letters* 27.4, pp. 294–300 (cit. on pp. 10, 72).

Goldewijk, K. K. (2001). "Estimating global land use change over the past 300 years: The HYDE Database". In: *Global Biogeochemical Cycles* 15.2, pp. 417–433 (cit. on p. 2).

Guerschman, J. P., M. J. Hill, L. J. Renzullo, D. J. Barrett, A. S. Marks & E. J. Botha (2009). "Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors". In: *Remote Sensing of Environment* 113.5, pp. 928–945 (cit. on pp. 11, 12, 19, 63, 65, 71, 72, 157).

Gumma, M. K., P. S. Thenkabail & A. Nelson (2011). "Mapping Irrigated Areas Using MODIS 250 Meter Time-Series Data: A Study on Krishna River Basin (India)". In: *Water* 3.1, pp. 113–131 (cit. on pp. 8, 56, 64, 86).

Gutman, G., C. Justice, E. Sheffner & T. Loveland (2012). "The NASA Land Cover and Land Use Change Program". In: *Land-Use and Land-Cover Change Pathways and Impacts*. Dordrecht: Springer Netherlands, pp. 17–29 (cit. on p. 7).

Hansen, M. C., P. V. Potapov, R Moore, M Hancher, S. A. Turubanova, A Tyukavina, D Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A Kommareddy, A Egorov, L Chini, C. O. Justice & J. R. G. Townshend (2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change". In: *Science* 342.6160, pp. 850–853 (cit. on pp. 1–3, 7, 37, 56, 161).

He, H. & E. A. Garcia (2009). "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9, pp. 1263–1284 (cit. on pp. 12, 106, 124).

He, Y. & Y. Bo (2011). "A consistency analysis of MODIS MCD12Q1 and MERIS Globcover land cover datasets over China". In: *Geoinformatics, 2011 19th International Conference on*. IEEE, pp. 1–6 (cit. on pp. 8, 56).

Herold, M, P Mayaux, C. E. Woodcock, A Baccini & C Schmullius (2008). "Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets". In: *Remote Sensing of Environment* 112.5, pp. 2538–2556 (cit. on pp. 4, 5, 8, 9, 63).

Herold, M, C Woodcock, M Cihlar, M Wulder & O Arino (2009). *Assessment of the status of the development of the standards for the Terrestrial Essential Climate Variables: T9 Land Cover*. FAO (cit. on pp. 2, 7, 10).

Hoffmann, M., C. Hilton-Taylor, A. Angulo, M. Böhm, T. M. Brooks, S. H. Butchart, K. E. Carpenter, J. Chanson, B. Collen, N. A. Cox & others (2010). "The impact of conservation on the status of the world's vertebrates". In: *Science* 330.6010, pp. 1503–1509 (cit. on pp. 2, 105).

Hüttich, C., U. Gessner, M. Herold, B. J. Strohbach, M. Schmidt, M. Keil & S. Dech (2009). "On the Suitability of MODIS Time Series Metrics to Map Vegetation Types in Dry Savanna Ecosystems: A Case Study in the Kalahari of NE Namibia". In: *Remote Sensing* 1.4, pp. 620–643 (cit. on pp. 10, 64, 72, 86, 106).

Immitzer, M., C. Atzberger & T. Koukal (2012). "Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data". In: *Remote Sensing* 4.12, pp. 2661–2693 (cit. on pp. 10, 12, 72, 107).

Japkowicz, N. & S. Stephen (2002). "The class imbalance problem: A systematic study". In: *Intelligent data analysis* 6.5, pp. 429–449 (cit. on pp. 12, 106, 126).

Johnson, B., R. Tateishi & T. Kobayashi (2012). "Remote Sensing of Fractional Green Vegetation Cover Using Spatially-Interpolated Endmembers". In: *Remote Sensing* 4.12, pp. 2619–2634 (cit. on pp. 11, 85).

Jun, M. & J. Kang (2010). *Muddy Water Management and Agricultural Development Measures in the Watershed of Soyang Dam: Focused on Haean-myeon, Yanggu-gun*. Tech. rep. Chuncheon (cit. on pp. 19, 47).

Kang, M., S Park, H Kwon, H. T. Choi, Y. J. Choi & J Kim (2009). "Evapotranspiration from a deciduous forest in a complex terrain and a heterogeneous farmland under monsoon climate". In: *Asia-Pacific Journal of Atmospheric Sciences* 45.2, pp. 175–191 (cit. on pp. 19, 65).

Liaw, A. & M. Wiener (2002). "Classification and Regression by randomForest". In: *R news* 2.3, pp. 18–22 (cit. on pp. 12, 72, 117).

Ling, C. X. & C. Li (1998). "Data mining for direct marketing: Problems and solutions." In: *Proceedings of the Fourth International Conference of Knowledge Discovery and Data Mining (KDD-98)*. Vol. 98, pp. 73–79 (cit. on pp. 12, 106).

Lobell, D. B. & G. Asner (2004). "Cropland distributions from temporal unmixing of MODIS data". In: *Remote Sensing of Environment* 93.3, pp. 412–422 (cit. on p. 11).

Loveland, T. R., B. C. Reed, J. F. Brown, D. O. Ohlen, Z Zhu, L Yang & J. W. Merchant (2000). "Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data". In: *International Journal Of Remote Sensing* 21.6-7, pp. 1303–1330 (cit. on pp. 4, 8, 37, 38, 43, 64, 105).

Loveland, T. R. & A. S. Belward (2010). "The IGBP-DIS global 1km land cover data set, DISCover: First results". In: *International Journal of Remote Sensing* 18.15, pp. 3289–3295 (cit. on pp. 2, 37, 43).

Lu, H., M. R. Raupach, T. McVicar & D. J. Barrett (2003). "Decomposition of vegetation cover into woody and herbaceous components using AVHRR NDVI time series". In: *Remote Sensing of Environment* 86.1, pp. 1–18 (cit. on pp. 11, 12, 64, 71, 72, 157).

Mahecha, M. D., L. M. Fürst, N. Gobron & H. Lange (2010). "Identifying multiple spatiotemporal patterns: A refined view on terrestrial photosynthetic activity". In: *Pattern Recognition Letters* 31.14, pp. 2309–2317 (cit. on pp. 1, 3, 37, 161).

Masson, V., J.-L. Champeaux, F. Chauvin, C. Meriguet & R. Lacaze (2003). "A Global Database of Land Surface Parameters at 1-km Resolution in Meteorological and Climate Models." In: *Journal of Climate* 16.9, pp. 1261–1282 (cit. on pp. 7, 9).

Matthews, E. (1983). "Global vegetation and land use: New high-resolution data bases for climate studies". In: *Journal of Climate and Applied Meteorology* 22.3, pp. 474–487 (cit. on pp. 1–3, 37, 161).

Meusburger, K, L Mabit, J. H. Park & T Sandor (2013). "Combined use of stable isotopes and fallout radionuclides as soil erosion indicators in a forested mountain site, South Korea." In: *Biogeosciences* 10, pp. 5627–5638 (cit. on pp. 19, 39).

Moody, A & C. E. Woodcock (1995). "The influence of scale and the spatial characteristics of landscapes on land-cover mapping using remote sensing". In: *Landscape Ecology* 10.6, pp. 363–379 (cit. on p. 6).

Mora, B., N.-E. Tsendbazar, M. Herold & O. Arino (2014). "Global Land Cover Mapping: Current Status and Future Trends". In: *Land Use and Land Cover Mapping in Europe*. Dordrecht: Springer Netherlands, pp. 11–30 (cit. on pp. 2, 4–7, 9, 10, 63, 64).

Mountrakis, G., J. Im & C. Ogole (2011). "Support vector machines in remote sensing: A review". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.3, pp. 247–259 (cit. on pp. 10, 107).

Müller, D. & D. K. Munroe (2014). "Current and future challenges in land-use science". In: *Journal of Land Use Science* 9.2, pp. 133–142 (cit. on pp. 1, 3–5).

NASA Land Processes Distributed Active Archive Center (LP DAAC) (2013a). *MOD13A1 Vegetation Indices 16-Day L3 Global 500m*. Tech. rep. 47914 252nd Street, Sioux Falls, South Dakota (cit. on pp. 9, 68, 69, 160).

Nguyen, T. T., M. Ruidisch, T. Koellner & J. Tenhunen (2014). "Synergies and tradeoffs between nitrate leaching and net farm income: The case of nitrogen best management practices in South Korea". In: *Agriculture Ecosystems & Environment* 186, pp. 160–169 (cit. on pp. 3, 157).

Nitze, I., B. Barrett & F. Cawkwell (2015). "Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series". In: *International Journal Of Applied Earth Observation And Geoinformation* 34, pp. 136–146 (cit. on pp. 10, 72, 106).

Obata, K., T. Miura & H. Yoshioka (2012). "Analysis of the Scaling Effects in the Area-Averaged Fraction of Vegetation Cover Retrieved Using an NDVI-Isoline-Based Linear Mixture Model". In: *Remote Sensing* 4.7, pp. 2156–2180 (cit. on pp. 11, 12, 67, 71).

Pal, M (2006). "Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data". In: *International Journal Of Remote Sensing* 27.14, pp. 2877–2894 (cit. on p. 10).

Pittman, K., M. C. Hansen, I. Becker-Reshef, P. V. Potapov & C. O. Justice (2010). "Estimating Global Cropland Extent with Multi-year MODIS Data". In: *Remote Sensing* 2.7, pp. 1844–1863 (cit. on pp. 8, 9, 11, 37, 38, 56, 63, 64, 86, 105).

Poppenborg, P. & T. Koellner (2013). "Do attitudes toward ecosystem services determine agricultural land use practices? An analysis of farmer's decision-making in a South Korean watershed". In: *Land Use Policy* 31.0, pp. 422–429 (cit. on pp. 2, 3, 37, 39, 153, 156, 157).

Potgieter, A. B., A. Apan, P Dunn & G Hammer (2007). "Estimating crop area using seasonal time series of Enhanced Vegetation Index from MODIS satellite imagery". In: *Crop and Pasture Science* 58.4, pp. 316–325 (cit. on pp. 8, 37, 38, 56).

Price, J. (1992). "Estimating vegetation amount from visible and near infrared reflectances". In: *Remote Sensing of Environment* 41.1, pp. 29–34 (cit. on pp. 11, 63).

Rindfuss, R. R., B. Entwisle, S. J. Walsh, L. An, N. Badenoch, D. G. Brown, P. Deadman, T. P. Evans, J. Fox, J. Geoghegan, M. Gutmann, M. Kelly, M. Linderman, J. Liu, G. P. Malanson, C. F. Mena, J. P. Messina, E. F. Moran, D. C. Parker, W. Parton, P. Prasartkul, D. T. Robinson, Y. Sawangdee, L. K. Vanwey & P. H. Verburg (2008). "Land use change: complexity and comparisons". In: *Journal of Land Use Science* 3.1, pp. 1–10 (cit. on p. 3).

Rodriguez-Galiano, V. F., B Ghimire, J Rogan, M Chica-Olmo & J. P. Rigol-Sanchez (2012). "An assessment of the effectiveness of a random forest classifier for land-cover classification". In: *Isprs Journal of Photogrammetry and Remote Sensing* 67, pp. 93–104 (cit. on pp. 10, 72, 74, 106, 117).

Schistad Solberg, A. & R. Solberg (1996). "A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images". In: *Geoscience and Remote Sensing Symposium, 1996. IGARSS'96.'Remote Sensing for a Sustainable Future.', International.* Vol. 3. IEEE, pp. 1484–1486 (cit. on pp. 12, 106).

Schulp, C. & R Alkemade (2011). “Consequences of uncertainty in global-scale land cover maps for mapping ecosystem functions: an analysis of pollination efficiency”. In: *Remote Sensing* 3.9, pp. 2057–2075 (cit. on pp. 1–3, 37, 38, 63, 161).

Schwarz, M. & N. E. Zimmermann (2005). “A new GLM-based method for mapping tree cover continuous fields using regional MODIS reflectance data”. In: *Remote Sensing of Environment* 95.4, pp. 428–443 (cit. on pp. 11, 63, 64, 72, 157).

Schwieder, M., P. Leitão, S. Suess, C. Senf & P. Hostert (2014). “Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques”. In: *Remote Sensing* 6.4, pp. 3427–3445 (cit. on pp. 10–12, 63, 71, 72, 83, 106, 157).

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression.* Tech. rep. Center for Bioinformatics Molecular Biostatistics (cit. on pp. 12, 72).

Senf, C., D. Pflugmacher, S. van der Linden & P. Hostert (2013). “Mapping Rubber Plantations and Natural Forests in Xishuangbanna (Southwest China) Using Multi-Spectral Phenological Metrics from MODIS Time Series”. In: *Remote Sensing* 5.6, pp. 2795–2812 (cit. on pp. 10, 72).

Senf, C., P. J. Leitão, D. Pflugmacher, S. van der Linden & P. Hostert (2015). “Mapping land cover in complex Mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery”. In: *Remote Sensing of Environment* 156, pp. 527–536 (cit. on p. 10).

Seo, B., C. Bogner, P. Poppenborg, E. Martin, M. Hoffmeister, M. Jun, T. Koellner, B. Reineking, C. L. Shope & J. Tenhunen (2014). “Deriving a per-field land use and land cover map in an agricultural mosaic catchment”. In: *Earth System Science Data* submitted (cit. on pp. 8, 16, 66, 109, 141).

Seo, B., P. Poppenborg, E. Martin, M. Hoffmeister, C. Bogner, H. Elsayed Ali, B. Reineking & J. Tenhunen (2014). *Per-field land use and land cover data set of the Haean catchment, South Korea. doi:10.1594/PANGAEA.823677.* data set (cit. on pp. 15, 19, 21, 67, 96, 109).

Sexton, J. O., P. Noojipady, A. Anand, X.-P. Song, S. McMahon, C. Huang, M. Feng, S. Channan & J. R. Townshend (2015). “A model for the propagation of uncertainty from continuous estimates of tree cover to categorical forest cover and change”. In: *Remote Sensing of Environment* 156, pp. 418–425 (cit. on p. 7).

Shope, C. L., G. R. Maharjan, J Tenhunen, B Seo, K Kim, J Riley, S Arnhold, T Koellner, Y. S. Ok, S Peiffer, B Kim, J. H. Park & B Huwe (2014). “Using the SWAT model to improve

process descriptions and define hydrologic partitioning in South Korea". In: *Hydrology And Earth System Sciences* 18.2, pp. 539–557 (cit. on pp. 19, 39, 47, 56, 153, 157).

Smith, M. O., S. L. Ustin, J. B. Adams & A. R. Gillespie (1990). "Vegetation in deserts: I. A regional measure of abundance from multispectral images". In: *Remote Sensing of Environment* 31.1, pp. 1 –26 (cit. on pp. 11, 63, 67).

Sterling, S. M., A. Ducharne & J. Polcher (2012). "The impact of global land-cover change on the terrestrial water cycle". In: *Nature Climate Change* 3.4, pp. 385–390 (cit. on pp. 2, 105).

Sulla-Menashe, D., M. A. Friedl, O. N. Krankina, A. Baccini, C. E. Woodcock, A. Sibley, G. Sun, V. Kharuk & V. Elsakov (2011). "Hierarchical mapping of Northern Eurasian land cover using MODIS data". In: *Remote Sensing of Environment* 115.2, pp. 392–403 (cit. on pp. 9, 158, 162).

Sun, Y., M. S. Kamel, A. K. Wong & Y. Wang (2007). "Cost-sensitive boosting for classification of imbalanced data". In: *Pattern Recognition* 40.12, pp. 3358–3378 (cit. on pp. 12, 106).

Thackway, R., L. Lymburner & J. P. Guerschman (2013). "Dynamic land cover information: bridging the gap between remote sensing and natural resource management". In: *Ecology And Society* 18.1 (cit. on pp. 5, 10, 65).

Thenkabail, P. S., M. Schull & H. Turral (2005). "Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data". In: *Remote Sensing of Environment* 95.3, pp. 317–341 (cit. on pp. 9, 10, 64, 72, 86, 105).

Tolvanen, H., M. Rönkä, P. Vihervaara, M. Kamppinen, C. Arzel, N. Aarras & S. Thessler (2014). "Spatial information in ecosystem service assessment: data applicability in the cascade model context". In: *Journal of Land Use Science*, pp. 1–18 (cit. on p. 1).

Tomaselli, V., P. Dimopoulos, C. Marangi, A. S. Kallimanis, M. Adamo, C. Tarantino, M. Panitsa, M. Terzi, G. Veronico, F. Lovergine, H. Nagendra, R. Lucas, P. Mairota, C. A. Mücher & P. Blonda (2013). "Translating land cover/land use classifications to habitat taxonomies for landscape monitoring: a Mediterranean assessment". In: *Landscape Ecology* 28.5, pp. 905–930 (cit. on p. 3).

Turner, B. L., E. F. Lambin & A. Reenberg (2007). "The emergence of land change science for global environmental change and sustainability". In: *Proceedings of the National Academy of Sciences* 104.52, pp. 20666–20671 (cit. on pp. 2, 105, 161).

U.S. Geological Survey (2012). *Global Land Cover Characteristics Data Base Version 2.0*. Tech. rep. U.S. Geological Survey (cit. on pp. 4, 8, 38, 53, 64, 105).

Vitousek, P. M. (1994). "Beyond global warming: ecology and global change". In: *Ecology* 75.7, pp. 1861–1876 (cit. on p. 2).

Vitousek, P. M., H. A. Mooney, J. Lubchenco & J. M. Melillo (1997). "Human domination of Earth's ecosystems". In: *Science* 277.5325, pp. 494–499 (cit. on p. 2).

Vittek, M., A. Brink, F. Donnay, D. Simonetti & B. Desclée (2014). "Land Cover Change Monitoring Using Landsat MSS/TM Satellite Image Data over West Africa between 1975 and 1990". In: *Remote Sensing* 6.1, pp. 658–676 (cit. on pp. 9, 68, 160).

Vuolo, F. & C. Atzberger (2012). "Exploiting the Classification Performance of Support Vector Machines with Multi-Temporal Moderate-Resolution Imaging Spectroradiometer (MODIS) Data in Areas of Agreement and Disagreement of Existing Land Cover Products". In: *Remote Sensing* 4.12, pp. 3143–3167 (cit. on pp. 10, 107).

Wardlow, Egbert & Kastens (2007). "Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains". In: *Remote Sensing of Environment* 108.3, pp. 290–310 (cit. on pp. 8, 56).

Wardlow, B. D. & S. L. Egbert (2008). "Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the US Central Great Plains". In: *Remote Sensing of Environment* 112.3, pp. 1096–1116 (cit. on pp. 8, 56).

Waske, B., J. A. Benediktsson & J. R. Sveinsson (2009). "Classifying remote sensing data with support vector machines and imbalanced training data". In: *Multiple Classifier Systems*. Springer, pp. 375–384 (cit. on pp. 12, 106).

Watts, J. D., S. L. Powell, R. L. Lawrence & T. Hilker (2010). "Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery". In: *Remote Sensing of Environment* 115.1, pp. 66–75 (cit. on pp. 9, 68, 160).

Williams, D. P., V. Myers & M. S. Silvious (2009). "Mine classification with imbalanced data". In: *Geoscience and Remote Sensing Letters, IEEE* 6.3, pp. 528–532 (cit. on pp. 12, 106).

Xiao, J. & A. Moody (2005). "A comparison of methods for estimating fractional green vegetation cover within a desert-to-upland transition zone in central New Mexico, USA". In: *Remote Sensing of Environment* 98.2-3, pp. 237–250 (cit. on pp. 10, 11, 84, 85).

Yihui, D. & J. C. L. Chan (2005). "The East Asian summer monsoon: an overview". In: *Meteorology and Atmospheric Physics* 89.1-4, pp. 117–142 (cit. on pp. 19, 65, 66).

Zhang, X, C Liao, J Li & Q Sun (2013). "Fractional vegetation cover estimation in arid and semi-arid environments using HJ-1 satellite hyperspectral data". In: *International Journal Of Applied Earth Observation And Geoinformation* 21, pp. 506–512 (cit. on p. 11).

Chapter 2

Deriving a per-field land use and land cover map in an agricultural mosaic catchment

2.1 Introduction

Agricultural land use affects ecosystem services, such as the provision of drinking water or the control of soil erosion. Inappropriate agricultural practice can lead to serious soil loss and pollution of surface water and groundwater by agrochemicals. Detailed data on land use and land cover (LULC) in an agricultural landscape constitute basic information for environmental monitoring and pollution control (Conrad et al., 2010; Pittman et al., 2010; Potgieter et al., 2007).

In general, precise information on land cover is required for running Earth system models (Ottlé et al., 2013) because land use change directly affects numerous climate parameters such as albedo, CO_2 cycling and hydrologic cycles (Mahecha et al., 2010; Matthews, 1983). Additionally, LULC information is crucial for ecosystem services research, decision making and studies on global change in general (Hansen et al., 2013; Poppenborg et al., 2013; Reineking et al., 2013; Schulp et al., 2011).

Remote sensing has been increasingly used to derive better LULC data for the past few decades (Bartholomé et al., 2005; Friedl et al., 2010; Loveland et al., 2000, 2010). Nevertheless, because available global land cover products are still limited thematically, continuous efforts to improve the LULC products are being made (Blanco et al., 2010; Colditz et al., 2011; Fernandes et al., 2004).

Particularly for agricultural landscapes, detailed land cover information is often lacking (Fritz et al., 2013; Pittman et al., 2010; Potgieter et al., 2007). In fact, the most widely used land cover databases such as GlobCover or Moderate Resolution Imaging Spectroradiometer (MODIS) land cover only have a few crop-related classes (Bontemps et al., 2011; Loveland et al., 2000; U.S. Geological Survey, 2012). Especially for heterogeneous arable zones, such as irrigated fields (e.g. Conrad et al., 2010), land cover products based on remote sensing are underdeveloped (Colditz et al., 2011).

Furthermore, spatial resolution of LULC data is often restricted. This limitation is particularly pronounced in heterogeneous landscapes, such as mixed-farming areas, due to the complex mosaic of crop/non-crop land use and land cover types (Schulp et al., 2011). Unlike a homogeneous landscape (e.g. plantation farm), this type of agricultural mosaic needs a comprehensive number of LULC classes in a relatively small area. Therefore, spatial resolution up to several hundred metres might be too coarse for this type of landscape. Longitudinal land cover data also constitute an important element when agricultural land use changes rapidly. MODIS Land Cover Type (MCD12Q1) is the only product that provides annual information. It has been widely used for analysing land cover changes (Loveland et al., 2000).

As a consequence, for an agricultural mosaic landscape with frequently changing land use, the only way to obtain detailed land cover information is surveying the study area.

In our study we address some of the above-mentioned problems and provide thematically and spatially rich land use and land cover data. We censused a small agroecosystem with complex agricultural land use. We recorded field-by-field land use and land cover type; hence, the unit entity of our data set is a single land parcel and we call it “per-field”. We followed Conrad et al. (2010), who defined “per-field” data as a data set based on actual agricultural fields.

In this paper we introduce the data and their collection and post-processing protocol. Additionally we compared our data with MCD12Q1. The data are now available at the public repository Pangaea (10.1594/PANGAEA.823677).

2.2 Material and methods

2.2.1 Study area

The study area, Haean catchment, is located at the border between North and South Korea (128°1′33.101″ E, 38°28′6.231″ N). It is a small agricultural catchment (64.4 km^2) with rice paddies, annual and perennial dry fields and orchard farms. Approximately 1200 inhabitants live in

Haean, mostly commercial farmers running their own small farms in the catchment. Agricultural fields in this area are typically smaller than 40 ha, and agricultural practice is intensive in terms of fertilisation and tillage.

The altitudes in the Haean catchment range from approximately 500 to 1200 m. Due to its characteristic bowl shape, land use changes, consisting predominantly of rice paddies at the valley bottom and dry-field farming on moderate slopes. The higher altitudes are covered by deciduous and mixed forests.

The average annual air temperature is approximately 8 °C, and the mean annual precipitation ranges from 1200 to 1300 mm, with more than 60 % of rainfall occurring during the summer monsoon between June and August (Korean Meteorological Administration, http://web.kma.go.kr/eng). Between 1999 and 2010 the maximum daily rainfall during summer reached up to 223 mm.

This area has been studied intensively as it shows a typical conflict between agriculture and environmental protection (Nguyen et al., 2012; Poppenborg et al., 2013; Reineking et al., 2013). The downstream water quality was heavily degraded by the agricultural activity occurring in the catchment (Meusburger et al., 2013; Shope et al., 2014). The local government pursued different policy measures concerning this conflict, such as subsidising perennial crops, which caused rapid LULC changes in land use and land cover.

2.2.2 Preparation of data collection

Prior to the field campaign, we collected pre-existing information to create an initial "base map". It served as a field template and was particularly useful for gaining access to isolated patches. We used a SPOTMaps image (Astrium Services, http://www.astrium-geo.com), a mosaic of multiple SPOT 5 images, with a ground resolution of 2.5 m. Furthermore, we worked with aerial photographs and a land cover map from the Korean Ministry of Environment (KME) (http://egis.me.go.kr) to complement the SPOTMaps image. From the KME land cover map, we extracted vector-based linear elements such as road and stream networks. An additional land use map by the Research Institute For Gangwon (http://gdri.re.kr) from 2007 provided information on previously surveyed agricultural land use. The data sources are summarised in Table 2.1.

The images selected for the base map were only moderately well georeferenced. The SPOTMaps image, for example, had an approximated location error of 10–15 m according to the specification, and the other spatial data also revealed a substantial location error. Therefore, we georeferenced

Table 2.1. Data used for the base map and gap filling. SPOTMaps served as the main background information for data collection. Maps by the Korean Ministry of Environment (KME) and by the Research Institute For Gangwon (RIG) provided previously recorded land use information and were also used for gap filling.

Name	Format	Temporal coverage	Description	Source
SPOTMaps	Raster (Geographic Tagged Image File Format, GeoTIFF)	2009	Mosaic of multiple SPOT 5 images, resolution 2.5 m, natural colours (three bands)	Astrium Services, http://www.astrium-geo.com/spotmaps
KME land cover map	Vector (polygon)	2000	21 classes based on Landsat Thematic Mapper (TM) images combined with large-scale vector maps	Ministry of Environment, Republic of Korea (KME), http://me.go.kr
RIG land use map	Vector (polygon)	2007	70 classes based on field observation	Research Institute For Gangwon, Republic of Korea (RIG), http://gdri.re.kr

them again using 14 ground control points (GCPs) distributed over the entire catchment. They were established along linear elements, such as roads, and defined by the Global Positioning System (GPS) coordinates averaged over several measurements. After georeferencing by the first-order polynomial (affine) transformation, the horizontal root mean squared error (RMSE) of the final base map image equalled 9.62 m.

2.2.3 Data collection

The main goal of the data collection campaign was to survey LULC information in the entire catchment. We carried out annual campaigns in 2009–2011 to census the entire landscape. The term "census" is adopted here in contrast to the term "sampling" because we recorded LULC information from the whole study area and not from a subset of land parcels. Accordingly, we mapped the complete set of land parcels and documented land cover type together with additional information on data quality and spatial and temporal mixture of land use types (e.g. double dry-field cropping per year or mixed dry fields). In contrast to 2009 and 2010, we were only able to map the northern half of the catchment in 2011 due to time and budget limitations. Therefore, we did not consider these data when calculating descriptive statistics or analysing land use change and only compared the years 2009 and 2010.

We divided landscape elements into two categories, namely patches and linear elements. The former included agricultural and non-agricultural fields, forest, waterbodies and all other areal land cover types best represented by a polygon. In general, we visited patches once per year. However, patches with a spatially or temporally mixed land use type were inspected multiple times. Linear elements comprised roads, stream networks, field edges or any other element that can be represented by a polyline. They were investigated during the whole project period from 2009 to 2011 because of their large extent and relative temporal stability.

To record a landscape element, we marked vertices and edges for each spatial entity as GPS waypoints (WPs) and tracks. The WP IDs were written on the printed base map and corresponding information in the field data book. GPS tracks were continuously stored in the device as we moved around and gave us complementary data for drawing polygon edges and polylines.

We used several GPS devices (Garmin CSX60, Garmin Colorado 300 and Garmin eTrex 30) simultaneously to retrieve location information. The use of multiple devices as a back-up secured the data against sudden power loss. For devices capable of loading custom maps, we loaded the base map in order to simultaneously review newly recorded WPs.

2.2.4 Post-processing

2.2.4.1 Digitising the field records

We digitised the field records into polygons and polylines with LULC type labels. The base map served as background information to complement the field records. In addition to LULC classes, we stored other descriptive information in an attribute table. In the corresponding columns, quality assurance (QA) was recorded as: "?" (questionable), "*" (unknown) and "/" (not valid). For instance, a question mark was assigned if we could not identify the crop reliably. Gap-filled data were also marked by a question mark. A forward slash indicates that the data were collected but was unreliable (e.g. incorrect identification). For further information, the reader is referred to the readme file of the data set at the Pangaea repository.

2.2.4.2 Gap filling

After digitising the field records, some gaps remained between polygons. They occurred mostly around patches that were irregularly shaped and therefore difficult to map. We filled these gaps using the KME land cover map (Table 2.1) and our own data on linear elements.

First, we added the main road and stream networks extracted from the KME land cover map. Subsequently, we created two major linear elements, namely seminatural field edges and a stream network from our GPS track data. For this purpose, we converted the GPS tracks of field edges and non-paved agricultural pathways, which were initially polylines, into polygons by creating 6 m wide buffers encompassing the tracks. Similarly, based on the GPS tracks recorded along small streams, we created the stream network buffers of 5 m width and assigned them to the existing "inland water" polygons.

Finally, we used the KME land cover map to fill the remaining gaps. Forest areas that were inaccessible due to military restrictions made up the major part of the transferred land cover information.

We updated the QA information during the gap-filling procedure. If only original observations without any extrapolated information are of interest, the QA flag can be used to filter out transferred land cover information. Because the LULC data were recorded yearly, the gaps differed from year to year. Therefore, we filled them separately for each year.

2.2.4.3 Definition of LULC classes

We defined a LULC classification scheme with 67 land use and land cover classes to adequately represent the agriculture mosaic in the catchment. If several LULC types coexisted in one

polygon, we assigned it to the LULC type that made up the largest portion and recorded mixture information in the attribute table. The scheme incorporates a large number of regional crop types as well as several natural and seminatural land cover classes found in the area. In the following we call this detailed classification scheme S1.

For vegetative classes, we also recorded information on life form, life cycle and crop type following the land cover classification system (LCCS) developed by the FAO (Food and Agricultural Organization of the United Nations) (Di Gregorio, 2005). We categorised the life cycle of a class as "perennial", "annual" or their mixture "annual/perennial" based on the life cycle of the plant species and the local cultivation practice. In other words, if a perennial crop was harvested after one growing season we classified it as "annual". We distinguished between the life forms "woody", "herbaceous" and "lichens/mosses", or a combination of them. Crop type patches were further subdivided into 12 different crop types (Supplement Table S1 at Pangaea repository). We assigned mixed crop type values for patches where various crop/non-crop vegetation coexisted.

In addition to the S1 scheme containing 67 classes, we reclassified the LULC information according to three simpler schemes. First, we generated a locally optimised scheme with 10 classes (called S2 in the following) that reflects the edaphic and socio-economic conditions in the study area. It consists of the classes "barren", "dry field", "forest", "greenhouse", "inland water", "orchard field", "paddy field", "seminatural" and "urban". Then, based on the FAO-LCCS we regrouped the S1 classes into eight major types (Supplement Table S2 at Pangaea repository). Two of the eight classes are relevant for crop distinction. Finally, we classified our data according to the International Geosphere–Biosphere Programme (IGBP) Discover land cover system which contains 17 classes, two of which are crop classes (Friedl et al., 2010; Loveland et al., 2000, 2010). Thus, the schemes S1, S2, FAO-LCCS and IGBP differ in the total number of classes and the number of crop classes (Table 2.2).

Table 2.2. Characteristics of the different land use and land cover classification schemes.

Name	Description	Total classes	Classes related to agriculture
S1	LULC types observed	67	Individual crops recorded
S2	Locally defined grouping	10	"Dry field", "paddy field" and "orchard field"
FAO-LCCS	FAO-LCCS major land cover classes	8	"Cultivated terrestrial" and "cultivated aquatic"
IGBP	IGBP Discover system	17	"Croplands" and "cropland/ natural vegetation mosaics"

These reclassified LULC data can be used together with global products such as MCD12Q1 or GlobCover that follow the IGBP and FAO-LCCS schemes, respectively. For the IGBP classes, we reclassified some of the perennial crops as non-crop types (forest or shrub) to be consistent with the IGBP system (e.g. "orchard field" coded as "open shrub") (Friedl et al., 2002). We also reclassified rice paddies as "croplands", unlike in S2, which distinguishes "paddy field" from other agricultural types.

2.2.4.4 Comparison with MODIS land cover

We compared the proportions of different classes in our data set with those provided in MCD12Q1 Land Cover Type 1 (IGBP). Additionally, we compared maps derived from our data with those provided in MCD12Q1 for 2009 and 2010. Therefore, we rasterised our survey data at the same spatial resolution (MODIS 500 m sinusoidal grid). We determined the LULC class label of a grid cell covered by multiple polygons based on the exact area size. Therefore, we calculated the fraction of the occupied area in the projected (Euclidean) space and assigned the LULC class labels based on the highest proportion.

To measure the agreement between maps, we derived confusion matrices and calculated Cohen's non-weighted κ (Cohen, 1960):

$$\kappa = \frac{p_\mathrm{o} - p_\mathrm{c}}{1 - p_\mathrm{c}}\,, \tag{2.1}$$

where p_o is the proportion of pixels in which the two data sets agreed and p_c is the proportion of pixels for which agreement is expected by chance.

Recently, κ has been criticised because of its limited use in remote sensing (Pontius Jr et al., 2011). Therefore, we also provide Pontius's quantity disagreement Q and allocation disagreement A. They are defined as

$$Q = \frac{\sum_{g=1}^{J} q_g}{2} \tag{2.2}$$

and

$$A = \frac{\sum_{g=1}^{J} a_g}{2}\,, \tag{2.3}$$

where q_g and a_g are quantity disagreement and allocation disagreement in the LULC class g. They are calculated as

$$q_g = \left| \left(\sum_{i=1}^{J} p_{ig} \right) - \left(\sum_{j=1}^{J} p_{gj} \right) \right| \tag{2.4}$$

and

$$a_g = 2 \min \left[\left(\sum_{i=1}^{J} p_{ig} \right) - p_{gg}\,, \left(\sum_{j=1}^{J} p_{gj} \right) - p_{gg} \right]\,, \tag{2.5}$$

where p_{ig} is the proportion of the area of class g in the reference map, p_{gj} is its proportion in the comparison map and p_{gg} is the proportion classified correctly.

The overall quantity disagreement Q indicates the difference between a reference map and a comparison map due to the less than perfect match in the proportions of the categories. The overall allocation disagreement A shows the difference between a reference map and a comparison map caused by the less than optimal match in the allocation of the categories. Finally, the total disagreement D is the sum of Q and A.

2.2.4.5 Software

We processed the data in GNU R v3.0.2 (R Core Team, 2013) and provide the R code along with the data set in the repository Pangaea (10.1594/PANGAEA.823677). For the reclassification, we used the package raster (Bivand et al., 2014). For the rasterisation, we used the geometry engine GEOS (Geometry Engine - Open Source) (GEOS Development Team, 2014) through the package rgeos (Bivand et al., 2014).

2.3 Results and discussion

2.3.1 Local classification scheme S1

The field survey resulted in vector geographic information system (GIS) data with 67 LULC classes (S1). Overall, the study site can be characterised as an extremely heterogeneous agricultural landscape with a large number of LULC types in its central part (Fig. 2.1; proportions in the Supplement Table S3 at Pangaea repository). We provide more details on the LULC types in the meta information of the data set (cf. Supplement to the data set at Pangaea repository).

The data have 3377 polygons with an average size of 0.019 km^2. Because in 2011 we only surveyed the northern half of the catchment, 12.3 % of the values were lacking for this year.

"Deciduous forest" at the steep hill slopes was stable during the studied period. It occupied more than half of the study area and was therefore the most dominant type (55.6 %, 2-year average). The moderate slopes from the forest edges to the flat centre were dominated by dry-field farms which occupied 16.3 % (2-year average) of the total catchment. The major dry-field crops among the total of the 42 we recorded were soybean, ginseng, potato, radish, European and Chinese cabbages and maize. Rice paddies (8.3 %, 2-year average) were prevalent in the central part and surrounded the small urban core (0.86 %, 2-year average).

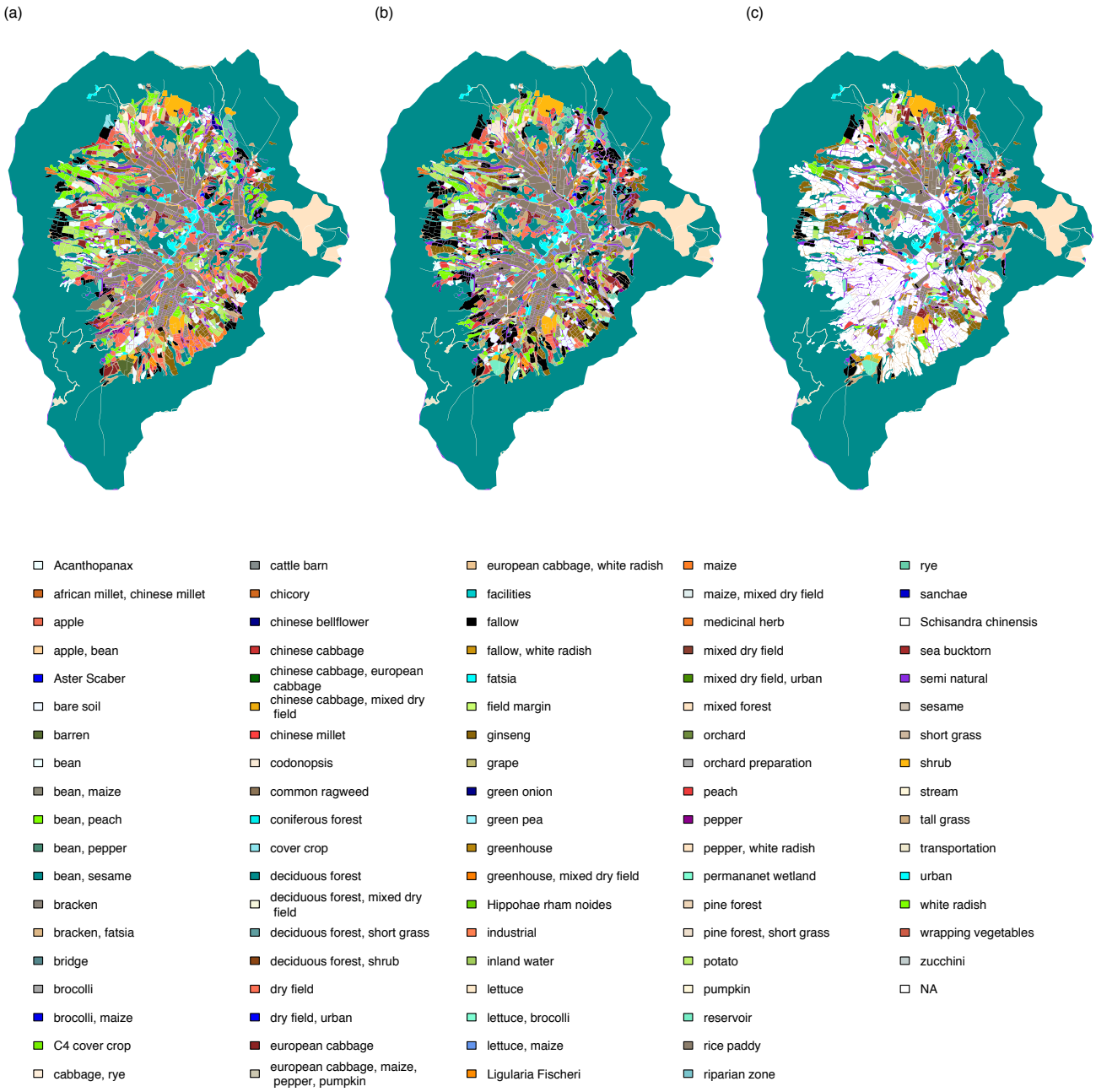

Fig. 2.1 Land use and land cover in the Haean catchment in **(a)** 2009, **(b)** 2010 and **(c)** 2011 according to the classification scheme S1 containing 67 classes.

2.3.1.1 Major changes in land use

During the study period, dry fields and rice farming decreased and orchards and ginseng cultivation increased (Table 2.6 and Supplement Table S3 at Pangaea repository). In fact, "Ginseng" almost doubled from 2009 to 2010 (1.26 to 2.48 %). It is consistent with the rapid ginseng expansion reported by Jun et al. (2010), who suggested replacing annual dry crops by perennial crops to stabilise soils and thus prevent erosion. An expected reduction of soil erosion due to this land use change was discussed in Arnhold et al. (2013), Kettering et al. (2012), and Ruidisch et al. (2013) and Shope et al. (2014).

Additionally, fallow fields increased in 2010 (4.8 %) compared to 2009 (1.9 %) and replaced a large number of dry fields. We attribute these changes partially to the subsidy for fallow fields and partially to corporal regulations requiring at least 3 years of fallow or organic farming before ginseng farming could start. The ginseng company Korea Ginseng Corporation only signs a contract with farmers when those regulations are fulfilled.

Compared to the patches, linear elements such as "seminatural" (6.0 %), "transportation" (0.78 %) and "inland water" (0.32 %) made up a small proportion in 2009 and 2010. Nevertheless, they covered the whole catchment (Fig. 2.1).

Field-level land use change was more pronounced than the change of the proportions due to crop rotation, which is common for the annual crops in the region. The annual crops are rarely cultivated in successive years and the dry-field crops commonly have a 3-year portfolio (e.g. potato–cabbage–soybean). This pattern is most distinctive in the northern part of the arable zone where the colours (LULC types) are displaced between 2009 and 2010 (Fig. 2.1). However this displacement is not reflected in the proportions.

2.3.1.2 Life form and life cycle

For vegetated patches, "herbaceous" vegetation dominated the central agricultural area in contrast to the surrounding forest which was entirely "woody" (Fig. 2.2). "lichens/mosses" type vegetation was not recorded. The life form did not change over the period studied (Table 2.3), possibly because land use changes mainly occurred within the "herbaceous" category (i.e. in the agricultural area).

The distribution of life cycles changed from 2009 to 2010 (Table 2.4). "Annual"-type vegetation dropped from 19.87 to 17.45 % due to decreasing rice paddies and dry fields. In contrast, natural "perennial" vegetation expanded over a larger area (61.53 % in 2009 to 62.78 % in 2010). These changes are clearly visible in the mid-western part of the area (Fig. 2.3) and are probably

Table 2.3. Changes in the FAO-LCCS category life form. Note that the survey data of 2011 are incomplete.

Life form	Survey (%)		
	2009	2010	2011
Herbaceous	30.17	29.64	19.13
Herbaceous/woody	6.82	6.91	7.03
Non-vegetated	2.85	2.81	2.56
Woody	60.15	60.60	59.01
Missing data	0.02	0.03	12.27

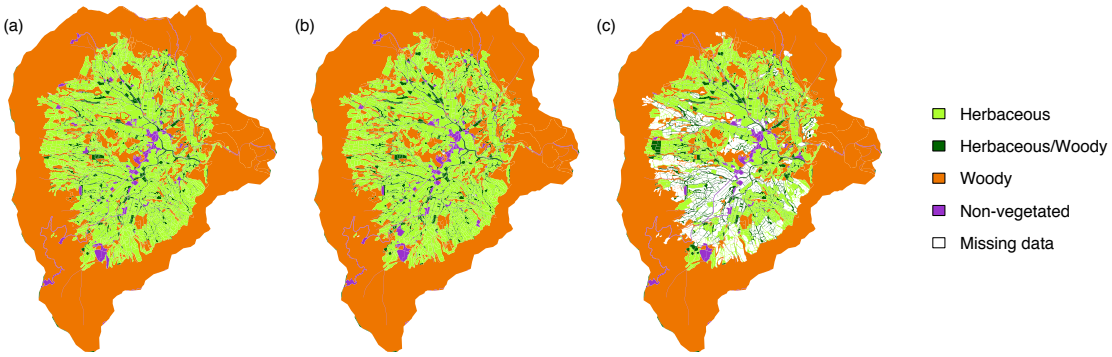

Fig. 2.2 Life form of the vegetation cover according to the FAO-LCCS in **(a)** 2009, **(b)** 2010 and **(c)** 2011.

due to the governmental policy of replacing dry fields by perennial crops such as ginseng and orchards.

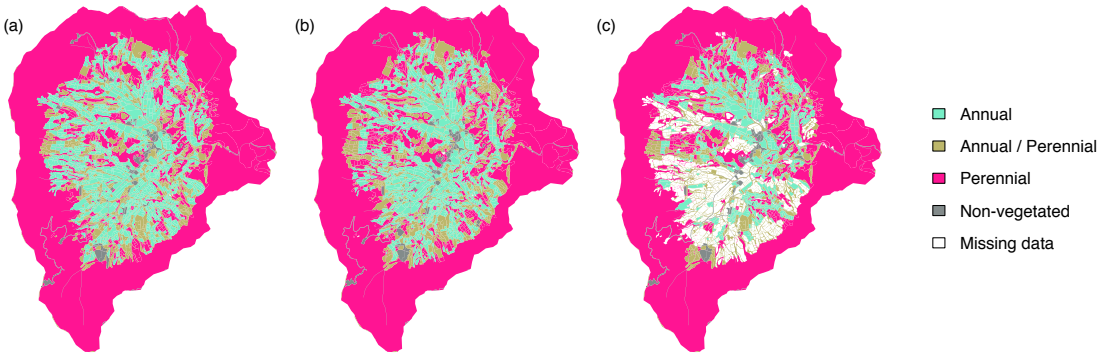

Fig. 2.3 Life cycle of the vegetation cover according to the FAO-LCCS in **(a)** 2009, **(b)** 2010 and **(c)** 2011.

2.3.1.3 Crop types

We found 6 of the 12 FAO-LCCS crop types in the study area, namely "cereals and pseudocereals", "roots and tubers", "pulses and vegetables", "fruits and nuts", "fodder crops" and "industrial crops" (Supplement Table S1 at Pangaea repository). We used combinations of them if multiple

Table 2.4. Changes of the FAO-LCCS category life cycle. Note that the survey data of 2011 are incomplete.

Life cycle	Survey (%)		
	2009	2010	2011
Annual	19.87	17.45	10.58
Annual/perennial	15.85	16.93	13.14
Non-vegetated	2.73	2.81	2.54
Perennial	61.53	62.78	61.46
Missing data	0.02	0.03	12.27

crop types were identified on the same patch. Occasionally, the class "mixed crops" was assigned when the combination was not precisely recorded.

For some crops, the most suitable type was difficult to find. Indeed, the LCCS manual classifies "soybean" as an industrial crop, while in the region it is often used as a vegetable because the green part is popular in local cuisine. "Wild sesame" is another example of a crop with multiple purposes, namely "pulses and vegetables" and "industrial crops". In our study we defined "soybean" and "wild sesame" as "industrial crops".

The 3 years of crop type information are shown in Fig. 2.4 and summarised in Table 2.5. "Cereals and pseudocereals" and "roots and tubers" diminished as "rice paddy", "white radish" and "potato" cultivation decreased. In contrast, "fruit and nuts" and "industrial crops" increased because the orchards and a few other industrial crops such as "ginseng" expanded due to the governmental promotion of perennial crops. Additionally, "non-crop vegetation" rose from 2009 to 2010 (69.1 to 72.0 %) as a consequence of an increased number of fallow fields in preparation for future ginseng farming.

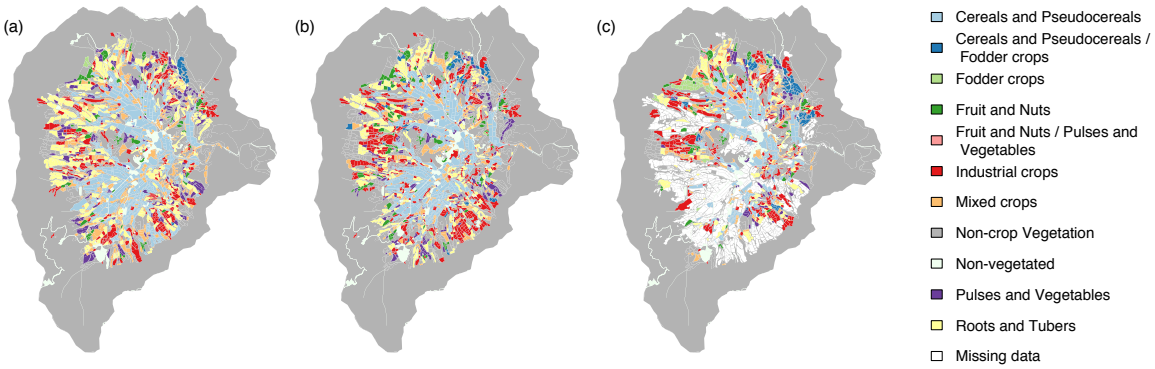

Fig. 2.4 Crop types according to the FAO-LCCS in **(a)** 2009, **(b)** 2010 and **(c)** 2011.

Table 2.5. Proportions of crop types defined according to the FAO-LCCS crop types. Note that the survey data of 2011 are incomplete.

Crop types	Survey (%)		
	2009	2010	2011
Cereals and pseudocereals	9.25	8.34	4.77
Cereals and pseudocereals/fodder crops	0.26	0.77	0.93
Fodder crops	0.07	0.09	0.59
Fruit and nuts	1.07	1.48	0.91
Fruit and nuts/pulses and vegetables	0.00	0.01	0.04
Industrial crops	3.76	5.35	4.27
Mixed crops	4.50	2.74	2.26
Non-crop vegetation	69.08	71.96	67.97
Non-vegetated	2.73	2.81	2.54
Pulses and vegetables	2.57	1.70	0.95
Roots and tubers	6.69	4.71	2.50
Missing data	0.02	0.03	12.27

2.3.2 Classification schemes S2 and FAO-LCCS

The coarser classification scheme S2 summarises the main changes in land use in the study area (Fig. 2.5 and Table 2.6). Actually, "dry field" dropped from 2009 (17.83 %) to 2010 (14.83 %) and the "seminatural" type increased from 11.35 % to 14.18 %. We attribute the latter change to the spread of fallow fields.

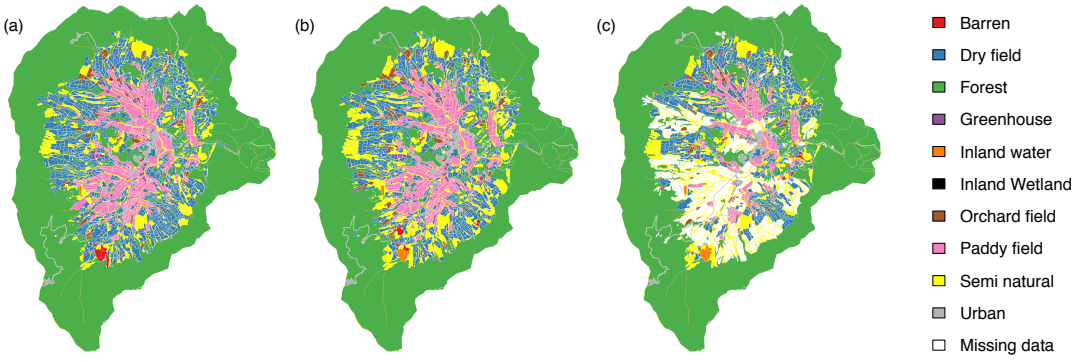

Fig. 2.5 Land use and land cover in the Haean catchment in **(a)** 2009, **(b)** 2010 and **(c)** 2011 according to the classification scheme S2.

The three dominant FAO-LCCS types, namely "natural and seminatural terrestrial vegetation", "cultivated and managed terrestrial area" and "cultivated aquatic or regularly flooded areas" covered 97.2 % (2-year average) of the total area. The "natural and seminatural terrestrial vegetation" prevailed (70.6 %, 2-year average) and increased from 2009 to 2010 (Table 2.7). In contrast, "cultivated and managed terrestrial area" and "cultivated aquatic or regularly flooded areas" decreased, probably due to reduced dry-field and rice farming, respectively.

When applying the FAO-LCCS scheme to our data, the classification of "rice paddy" was

Table 2.6. Changes in land use and land cover based on the classification scheme S2. Note that the survey data of 2011 are incomplete.

Class	Survey (%)		
	2009	2010	2011
Barren	0.31	0.22	0.08
Dry field	17.83	14.83	11.07
Forest	57.74	57.79	57.11
Greenhouse	0.77	0.84	0.58
Inland water	0.69	0.86	0.89
Inland wetland	0.00	0.00	0.00
Orchard field	1.07	1.48	0.91
Paddy field	8.50	8.04	4.65
Seminatural	11.35	14.18	10.86
Urban	1.72	1.72	1.57
Missing data	0.02	0.03	12.27

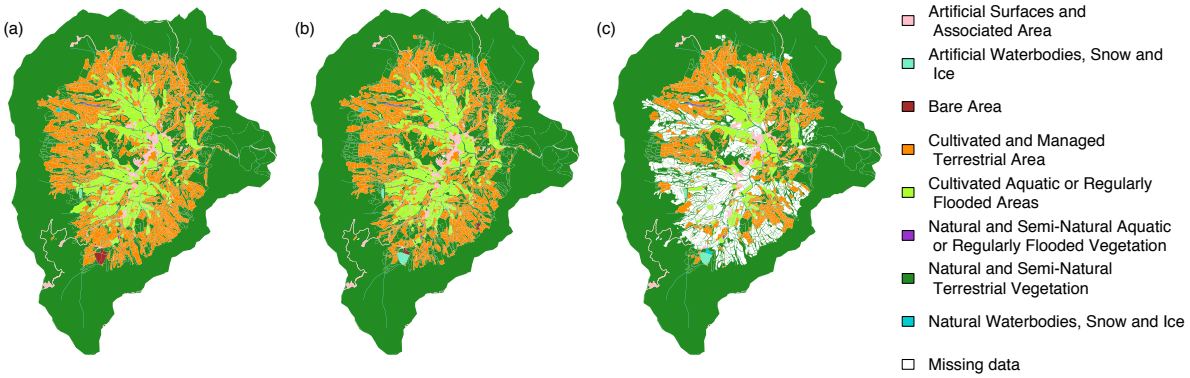

Fig. 2.6 Reclassified land use and land cover in **(a)** 2009, **(b)** 2010 and **(c)** 2011 according to the FAO-LCCS eight major land cover classes. The annual proportions are shown in Supplement Table S2 at Pangaea repository. These classes are defined by the stratified structure with three dichotomous levels: presence of vegetation, edaphic condition and artificiality of cover.

Table 2.7. Annual proportions of the reclassified land use and land cover data according to the FAO-LCCS eight major land cover classes. Note that the survey data of 2011 are incomplete.

LCCS eight major classes	Survey (%)		
	2009	2010	2011
Artificial surfaces and associated area	1.72	1.72	1.57
Artificial waterbodies, snow and ice	0.07	0.23	0.24
Bare area	0.22	0.08	0.05
Cultivated and managed terrestrial area	19.68	17.15	12.56
Cultivated aquatic or regularly flooded areas	8.50	8.04	4.65
Natural and seminatural aquatic or regularly flooded vegetation	0.09	0.09	0.07
Natural and seminatural terrestrial vegetation	69.09	72.02	67.94
Natural waterbodies, snow and ice	0.62	0.63	0.65
Missing data	0.02	0.03	12.27

challenging. In actual fact, in Haean, rice is sometimes irrigated with water from deep wells. However, although the "cultivated aquatic or regularly flooded areas" class excludes irrigated cultivated areas (Di Gregorio, 2005), we assigned rice to this type as it is mostly rainfed.

2.3.3 IGBP classification scheme

2.3.3.1 Comparison between MODIS land cover and the original survey data

We found 10 IGBP classes in our study area, namely "waterbodies", "evergreen needleleaf forests", "deciduous broadleaf forests", "mixed forests", "closed shrublands", "open shrublands", "grasslands", "croplands", "urban and built-up lands" and "barren or sparsely vegetated". In contrast, MCD12Q1 contained only five classes: "deciduous broadleaf forests", "mixed forests", "grasslands", "croplands" and "cropland/natural vegetation mosaics". The first row of Fig. 2.7 shows the original survey data and the third shows MCD12Q1. In addition, Table 2.8 summarises area proportions in both data sets.

Table 2.8. Changes of land use and land cover according to the IGBP 17-class system. The columns under "survey" refer to the survey data and those under "MODIS" to MODIS Land Cover Type (MCD12Q1) following the same classification system. Note that the "waterbodies" and "urban" classes were not detected by MODIS, presumably as a result of coarse resolution (500 m). Note that the survey data of 2011 are incomplete.

IGBP 17 classes	Survey (%)			MODIS (%)		
	2009	2010	2011	2009	2010	2011
Waterbodies	0.69	0.86	0.89	0.00	0.00	0.00
Evergreen needleleaf forests	0.29	0.29	0.30	0.00	0.00	0.00
Evergreen broadleaf forests	0.00	0.00	0.00	0.00	0.00	0.00
Deciduous needleleaf forests	0.00	0.00	0.00	0.00	0.00	0.00
Deciduous broadleaf forests	55.39	55.41	54.73	34.73	27.73	27.41
Mixed forests	2.06	2.08	2.08	12.45	24.58	25.57
Closed shrublands	3.60	3.67	3.14	0.00	0.00	0.00
Open shrublands	1.06	1.48	0.93	0.00	0.00	0.00
Woody savannas	0.00	0.00	0.00	0.00	0.00	0.00
Savannas	0.00	0.00	0.00	0.00	0.00	0.00
Grasslands	7.89	10.80	7.82	10.67	15.67	17.01
Permanent wetlands	0.00	0.00	0.00	0.00	0.00	0.00
Croplands	26.09	22.54	15.58	31.19	26.68	26.35
Urban and built-up lands	2.49	2.57	2.15	0.00	0.00	0.00
Cropland/natural vegetation mosaics	0.00	0.00	0.00	10.97	5.34	3.67
Snow and ice	0.00	0.00	0.00	0.00	0.00	0.00
Barren or sparsely vegetated	0.44	0.22	0.09	0.00	0.00	0.00
Interrupted areas	0.00	0.00	0.00	0.00	0.00	0.00
Missing data	0.02	0.03	12.27	0.00	0.00	0.00

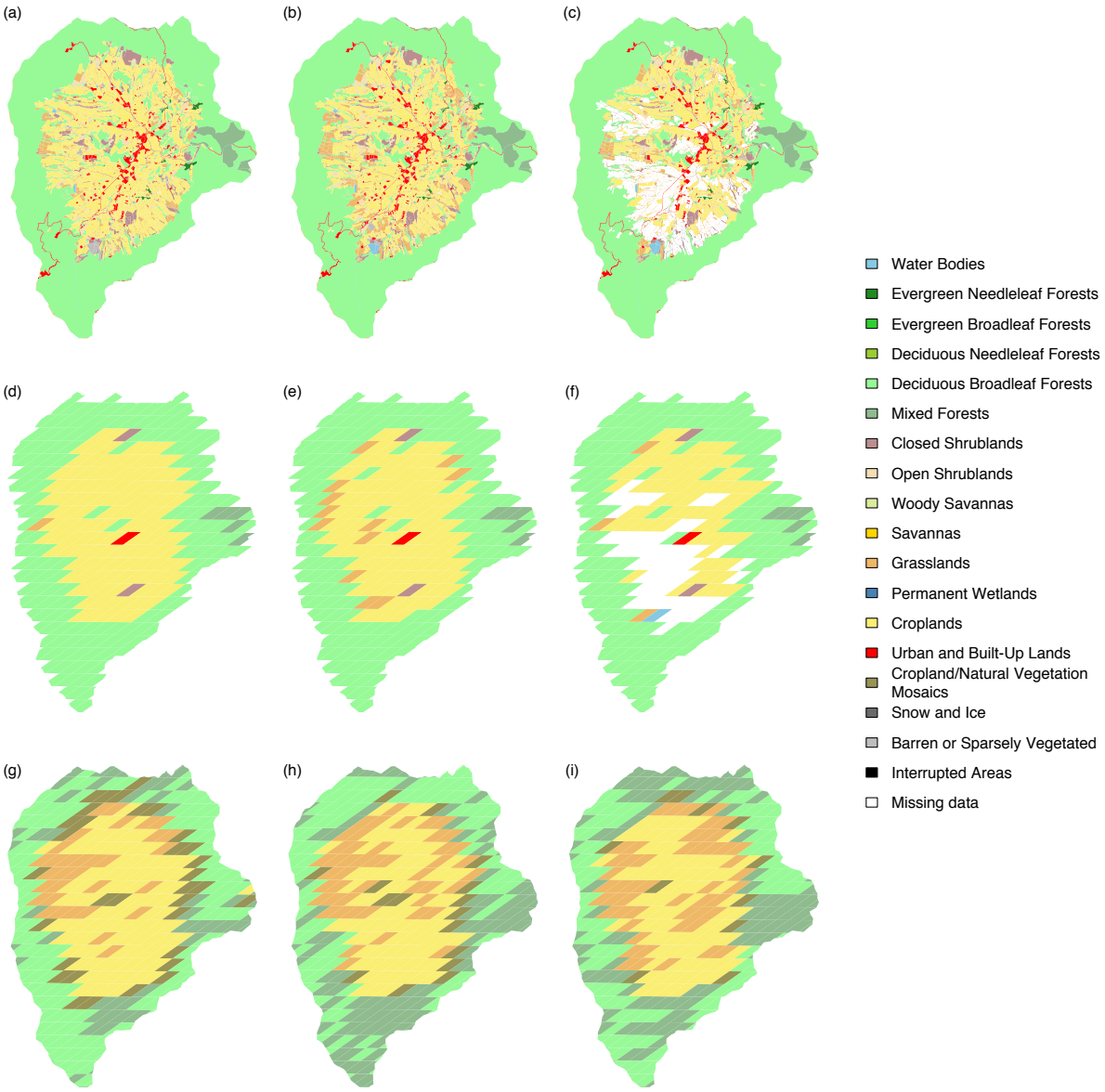

Fig. 2.7 Land use and land cover reclassified according to the IGBP 17-class system: the original survey data in **(a)** 2009, **(b)** 2010 and **(c)** 2011; the rasterised survey data in **(d)** 2009, **(e)** 2010 and **(f)** 2011; MODIS Land Cover Type product (MCD12Q1) in **(g)** 2009, **(h)** 2010 and **(i)** 2011. Note that the IGBP system does not distinguish the paddy field from a general cultivated zone. Note that "interrupted areas" is a special mask for Goode's interrupted area (U.S. Geological Survey, 2012).

For "croplands" the MODIS product shows a moderate agreement with the survey data (29.0 % vs. 24.3 %, 2-year averages). The mosaic class "cropland/natural vegetation mosaics" type was not found in our survey data while in the MODIS data set it comprises 10.97 % in 2009 and 5.34 % in 2010. MODIS assigns this class to pixels containing a mixed of croplands, forests, shrubland and grasslands as long as no single component comprises more than 60 % of the area (Friedl et al., 2002). By definition, this mixture class is ambiguous (Friedl et al., 2002; Friedl et al., 2010). In contrast, we explicitly recorded the individual classes for smaller patches instead of assigning the mosaic class for a larger area.

The shrubland classes as well as the cropland classes are relevant to agriculture as some of the perennial crop types were classified as "closed shrublands" and "open shrublands". We have more than 5 % of shrubland classes in the survey data which are not found in the MODIS product for the 2-year period.

There is an overrepresentation of the agricultural area in MCD12Q1 compared to our ground observations. If we combine all the agriculturally relevant classes, namely "croplands", "cropland/natural vegetation mosaics", "closed shrublands" and "open shrublands", these add up to 37.1 % in the MODIS land cover while they represent only 29.2 % in our survey data (2-year averages).

In contrast, the forested area is underrepresented by MODIS as "deciduous broadleaf forests" and "mixed forests" add up to 49.7 % in the MODIS land cover while they cover 57.5% in our survey, averaged over 2 years. Individually, in our survey, the area of "deciduous broadleaf forests" is larger (55.4 % vs. 31.2 %) and the area of "mixed forests" is substantially smaller (2.08 % vs. 18.5 %) compared to MCD12Q1 (averaged over 2 years).

The disagreement in the agricultural and the forest types may be due to the coarser resolution of the MODIS product (500 m). This becomes more problematic for land cover types smaller than the MODIS pixel in its typical dimension. Indeed, linear elements such as "waterbodies" and "urban and built-up lands" were not found in the product.

We note that, for the forest classes, our limited access to the surrounding forest may have caused inaccuracies in our data. Moreover, the agreement between the two data sets could be higher if we used the mosaic class "cropland/natural vegetation mosaics" for our data. There may have been patches that are better described as mixtures of cropland and natural vegetation than by reclassifying them either as pure cropland or pure natural vegetation. However, analysing this effect is beyond the scope of this work.

2.3.3.2 Comparison between MODIS land cover and the rasterised survey data

After rasterisation, six IGBP classes were found in the survey data, namely "deciduous broadleaf forests", "mixed forests", "closed shrublands", "grasslands", "croplands" and "urban and built-up lands". "Urban and built-up lands" were missing in the MODIS data while "cropland/natural vegetation mosaics" does not exist in our data. Figure 2.7 shows the rasterised ground observations (in the middle row) and MCD12Q1 (in the bottom row).

To compare the two maps, we derived confusion matrices, Cohen's κ and Pontius's Q and A for 2009 (Supplement Table S3 at Pangaea repository) and 2010 (Supplement Table S4 at Pangaea repository). We excluded the year 2011 due to a lack of ground observations. The mean κ for the 2 years equals 0.41, which indicates a fair but not substantial agreement.

For the 2-year average, the total disagreement D is 0.42, the quantity disagreement Q is 0.36 and the allocation disagreement A is 0.053. Thus, quantity disagreement accounts for 87 % of the overall disagreement. This suggests that MCD12Q1 may fail to evaluate the quantity of different LULC classes in complex agricultural landscapes.

2.4 Data structure and data access

The data set and its description are available at the Pangaea repository under the Creative Commons Attribution-NonCommercial 3.0 Unported license. The data contain LULC observations and ancillary information in a single ESRI polygon shape file (ESRI Inc., http://esri.com). The LULC type, QA, management and double-cropping and mixed-use information are provided on an annual basis. The definition of classes and the reclassification table are given separately in a legend table. For each polygon, LULC information for 3 years is given in separate columns (e.g. LULC2009, LULC2010 and LULC2011). Note that multiple entries in a LULC type column occur in cases when the polygon exhibited mixed land uses spatially or temporally.

2.5 Summary and conclusions

We provide an annual per-field land use and land cover data set for the agricultural mosaic catchment Haean (South Korea). During the study period many dry fields were converted to perennial crops such as ginseng and orchards, probably due to governmental policy measures. The comparison between our survey data and the MODIS land cover revealed that the limitation of MODIS cover in identifying irrigated fields could be a substantial source of error. Moreover,

MCD12Q1 overrepresents agricultural types and underrepresents forest types compared to our ground observations. Linear elements such as "waterbodies" were missing in the remote-sensing product due to its coarse spatial resolution. We measured the agreement between the rasterised ground truth and the MODIS land cover. The agreement was fair but not substantial for the primary land cover type.

Global Earth system models are major information sources for global environmental discussions and decision making. These models commonly use satellite-borne land use and land cover data sets as input. These land databases are equipped with generalised agricultural types. However the use of general cropland classes may be inappropriate in complex agricultural landscapes. For example, Berger et al. (2013a) pointeds out the lack of paddy soil and subsoil studies despite their potential impact on global carbon and nitrogen cycles. Recent studies in the same area repeatedly suggested that complex agricultural landscapes needed greater attention (Arnhold et al., 2013; Berger et al., 2013a,b; Kettering et al., 2012; Kim et al., 2014; Ruidisch et al., 2013; Shope et al., 2014). Thus, thematic improvement of global land cover databases is of great importance.

There have been ongoing efforts to extend MODIS land cover databases (Biggs et al., 2006; Gumma et al., 2011; He et al., 2011; Pittman et al., 2010; Potgieter et al., 2007; Wardlow et al., 2007; Wardlow et al., 2008). For natural vegetation, global high-resolution databases are becoming available (e.g. Hansen et al., 2013). Our vector-form data can be useful in developing/validating high-resolution data sets for complex agricultural landscapes because the data include detailed crop type information with a consistent and complete description established by the FAO (Di Gregorio, 2005). Additionally, our data contains different classification systems and can be transformed to any raster grid. Due to this detailed information, our data could be used for regional environmental modelling as well as for ecosystem services research and decision making analysis.

2.6 Acknowledgements

We thank Hamada Elsayed Ali, Sebastian Arnhold, Jaesung Eum and Ralf Geyer for their help in the laboratory and during data collection. This research was supported by the International Research Training Group of Germany and South Korea (DFG/KOSEF, Complex TERRain and ECOlogical Heterogeneity (TERRECO), GRK 1565/1).

MODIS data were obtained from https://lpdaac.usgs.gov, maintained by the NASA Land Processes Distributed Active Archive Center (LP DAAC) at the USGS/Earth Resources Obser-

vation and Science (EROS) Center.

References

Arnhold, S, M Ruidisch, S Bartsch & C. L. Shope (2013). "Simulation of runoff patterns and soil erosion on mountainous farmland with and without plastic-covered ridge-furrow cultivation in South Korea". In: *Transactions of the ASABE* 56, pp. 667–679 (cit. on pp. 47, 56).

Bartholomé, E & A. S. Belward (2005). "GLC2000: a new approach to global land cover mapping from Earth observation data". In: *International Journal Of Remote Sensing* 26.9, pp. 1959–1977 (cit. on pp. 4, 6, 7, 9, 37).

Berger, S., I. Jang, J. Seo, H. Kang & G. Gebauer (2013a). "A record of N2O and CH4 emissions and underlying soil processes of Korean rice paddies as affected by different water management practices". In: *Biogeochemistry* 115.1-3, pp. 317–332 (cit. on p. 56).

Berger, S., Y. Kim, J. Kettering & G. Gebauer (2013b). "Plastic mulching in agriculture—Friend or foe of N2O emissions?" In: *Agriculture Ecosystems & Environment* 167, pp. 43–51 (cit. on p. 56).

Biggs, T. W., P. S. Thenkabail, M. K. Gumma, C. A. Scott, G. R. Parthasaradhi & H. N. Turral (2006). "Irrigated area mapping in heterogeneous landscapes with MODIS time series, ground truth and census data, Krishna Basin, India". In: *International Journal Of Remote Sensing* 27.19, pp. 4245–4266 (cit. on pp. 8, 56, 64).

Bivand, R. & C. Rundel (2014). *rgeos: Interface to Geometry Engine - Open Source (GEOS)* (cit. on pp. 45, 77, 109).

Blanco, P. D., R. R. Colditz, G. L. Saldaña, L. A. Hardtke, R. M. Llamas, N. A. Mari, A. Fischer, C. Caride, P. G. Aceñolaza, H. F. del Valle, M. Lillo-Saavedra, F. Coronato, S. A. Opazo, F. Morelli, J. A. Anaya, W. F. Sione, P. Zamboni & V. B. Arroyo (2010). "A land cover map of Latin America and the Caribbean in the framework of the SERENA project". In: *Remote Sensing of Environment* 132.0, pp. 13–31 (cit. on p. 37).

Bontemps, S., P. Defourny, E. Bogaert, O. Arino, V. Kalogirou & J. Perez (2011). *GLOBCOVER 2009 - Products Description and Validation Report*. Tech. rep. European Space Agency (cit. on pp. 4, 5, 7–9, 38, 63, 64, 105).

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46 (cit. on p. 44).

Colditz, R. R., M Schmidt, C Conrad, M. C. Hansen & S Dech (2011). "Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions". In: *Remote Sensing of Environment* 115.12, pp. 3264–3275 (cit. on pp. 11, 12, 37, 38, 64, 71, 84, 85).

Conrad, C., S. Fritsch, J. Zeidler, G. Rücker & S. Dech (2010). "Per-Field Irrigated Crop Classification in Arid Central Asia Using SPOT and ASTER Data". In: *Remote Sensing* 2.4, pp. 1035–1056 (cit. on pp. 37, 38).

Di Gregorio, A (2005). *Land Cover Classification System: Classification Concepts and User Manual: LCCS*. Rome (Italy). Food and Agriculture Organization of the United Nations (FAO) (cit. on pp. 2, 3, 7, 43, 52, 56, 154).

Fernandes, R., R. Fraser, R. Latifovic, J. Cihlar, J. Beaubien & Y. Du (2004). "Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region". In: *Remote Sensing of Environment* 89.2, pp. 234–251 (cit. on pp. 11, 37, 63, 68, 72, 75, 85, 157).

Friedl, M. A., D. K. McIver, J. Hodges & X. Y. Zhang (2002). "Global land cover mapping from MODIS: algorithms and early results". In: *Remote Sensing of Environment* 83.1-2, pp. 287–302 (cit. on pp. 6, 10, 44, 54).

Friedl, M. A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley & X. Huang (2010). "MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets". In: *Remote Sensing of Environment* 114.1, pp. 168–182 (cit. on pp. 9, 37, 43, 54, 158).

Fritz, S., L. See, L. You, C. Justice, I. Becker Reshef, L. Bydekerke, R. Cumani, P. Defourny, K. Erb, J. Foley, S. Gilliams, P. Gong, M. Hansen, T. Hertel, M. Herold, M. Herrero, F. Kayitakire, J. Latham, O. Leo, I. McCallum, M. Obersteiner, N. Ramankutty, J. Rocha, H. Tang, P. Thornton, C. Vancutsem, M. Velde, S. Wood & C. Woodcock (2013). "The Need for Improved Maps of Global Cropland". In: *Eos, Transactions American Geophysical Union* 94.3, pp. 31–32 (cit. on pp. 4, 7, 8, 38).

GEOS Development Team (2014). *GEOS - Geometry Engine, Open Source*. Open Source Geospatial Foundation (cit. on pp. 45, 77, 109).

Gumma, M. K., P. S. Thenkabail & A. Nelson (2011). "Mapping Irrigated Areas Using MODIS 250 Meter Time-Series Data: A Study on Krishna River Basin (India)". In: *Water* 3.1, pp. 113–131 (cit. on pp. 8, 56, 64, 86).

Hansen, M. C., P. V. Potapov, R Moore, M Hancher, S. A. Turubanova, A Tyukavina, D Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A Kommareddy, A Egorov, L Chini, C. O. Justice & J. R. G. Townshend (2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change". In: *Science* 342.6160, pp. 850–853 (cit. on pp. 1–3, 7, 37, 56, 161).

He, Y. & Y. Bo (2011). "A consistency analysis of MODIS MCD12Q1 and MERIS Globcover land cover datasets over China". In: *Geoinformatics, 2011 19th International Conference on*. IEEE, pp. 1–6 (cit. on pp. 8, 56).

Jun, M. & J. Kang (2010). *Muddy Water Management and Agricultural Development Measures in the Watershed of Soyang Dam: Focused on Haean-myeon, Yanggu-gun*. Tech. rep. Chuncheon (cit. on pp. 19, 47).

Kettering, J., J.-H. Park, S. Lindner, B. Lee, J. Tenhunen & Y. Kuzyakov (2012). "N fluxes in an agricultural catchment under monsoon climate: A budget approach at different scales". In: *Agriculture Ecosystems & Environment* 161, pp. 101–111 (cit. on pp. 47, 56).

Kim, Y., S. Berger, J. Kettering, J. Tenhunen, E. Haas & R. Kiese (2014). "Simulation of N2O emissions and nitrate leaching from plastic mulch radish cultivation with LandscapeDNDC". In: *Ecological Research* 29.3, pp. 441–454 (cit. on p. 56).

Loveland, T. R., B. C. Reed, J. F. Brown, D. O. Ohlen, Z Zhu, L Yang & J. W. Merchant (2000). "Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data". In: *International Journal Of Remote Sensing* 21.6-7, pp. 1303–1330 (cit. on pp. 4, 8, 37, 38, 43, 64, 105).

Loveland, T. R. & A. S. Belward (2010). "The IGBP-DIS global 1km land cover data set, DISCover: First results". In: *International Journal of Remote Sensing* 18.15, pp. 3289–3295 (cit. on pp. 2, 37, 43).

Mahecha, M. D., L. M. Fürst, N. Gobron & H. Lange (2010). "Identifying multiple spatiotemporal patterns: A refined view on terrestrial photosynthetic activity". In: *Pattern Recognition Letters* 31.14, pp. 2309–2317 (cit. on pp. 1, 3, 37, 161).

Matthews, E. (1983). "Global vegetation and land use: New high-resolution data bases for climate studies". In: *Journal of Climate and Applied Meteorology* 22.3, pp. 474–487 (cit. on pp. 1–3, 37, 161).

Meusburger, K, L Mabit, J. H. Park & T Sandor (2013). "Combined use of stable isotopes and fallout radionuclides as soil erosion indicators in a forested mountain site, South Korea." In: *Biogeosciences* 10, pp. 5627–5638 (cit. on pp. 19, 39).

Nguyen, T. T., H. V. Ngu & B. Seo (2012). "Cost and environmental efficiency of rice farms in South Korea". In: *Agricultural Economics* 43.4, pp. 369–378 (cit. on pp. 39, 153, 156).

Ottlé, C, J Lescure, F Maignan, B Poulter, T Wang & N Delbart (2013). "Use of various remote sensing land cover products for plant functional type mapping over Siberia". In: *Earth System Science Data* 5.2, pp. 331–348 (cit. on p. 37).

Pittman, K., M. C. Hansen, I. Becker-Reshef, P. V. Potapov & C. O. Justice (2010). "Estimating Global Cropland Extent with Multi-year MODIS Data". In: *Remote Sensing* 2.7, pp. 1844–1863 (cit. on pp. 8, 9, 11, 37, 38, 56, 63, 64, 86, 105).

Pontius Jr, R. G. & M. Millones (2011). "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment". In: *International Journal Of Remote Sensing* 32.15, pp. 4407–4429 (cit. on p. 44).

Poppenborg, P. & T. Koellner (2013). "Do attitudes toward ecosystem services determine agricultural land use practices? An analysis of farmer's decision-making in a South Korean watershed". In: *Land Use Policy* 31.0, pp. 422–429 (cit. on pp. 2, 3, 37, 39, 153, 156, 157).

Potgieter, A. B., A. Apan, P Dunn & G Hammer (2007). "Estimating crop area using seasonal time series of Enhanced Vegetation Index from MODIS satellite imagery". In: *Crop and Pasture Science* 58.4, pp. 316–325 (cit. on pp. 8, 37, 38, 56).

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on p. 45).

Reineking, B. & B. Seo (2013). "Natural enemy interactions constrain pest control in complex agricultural landscapes." In: *Proceedings of the National Academy of Sciences* 110.14, pp. 5534–5539 (cit. on pp. 37, 39, 153, 157).

Ruidisch, M., S. Arnhold, B. Huwe & C. Bogner (2013). "Is Ridge Cultivation Sustainable? A Case Study from the Haean Catchment, South Korea". In: *Applied and Environmental Soil Science* 2013, Article ID 679467, 11 pages (cit. on pp. 47, 56).

Schulp, C. & R Alkemade (2011). "Consequences of uncertainty in global-scale land cover maps for mapping ecosystem functions: an analysis of pollination efficiency". In: *Remote Sensing* 3.9, pp. 2057–2075 (cit. on pp. 1–3, 37, 38, 63, 161).

Shope, C. L., G. R. Maharjan, J Tenhunen, B Seo, K Kim, J Riley, S Arnhold, T Koellner, Y. S. Ok, S Peiffer, B Kim, J. H. Park & B Huwe (2014). "Using the SWAT model to improve process descriptions and define hydrologic partitioning in South Korea". In: *Hydrology And Earth System Sciences* 18.2, pp. 539–557 (cit. on pp. 19, 39, 47, 56, 153, 157).

U.S. Geological Survey (2012). *Global Land Cover Characteristics Data Base Version 2.0*. Tech. rep. U.S. Geological Survey (cit. on pp. 4, 8, 38, 53, 64, 105).

Wardlow, Egbert & Kastens (2007). "Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains". In: *Remote Sensing of Environment* 108.3, pp. 290–310 (cit. on pp. 8, 56).

Wardlow, B. D. & S. L. Egbert (2008). “Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the US Central Great Plains”. In: *Remote Sensing of Environment* 112.3, pp. 1096–1116 (cit. on pp. 8, 56).

Chapter 3

Mapping Fractional Land Use and Land Cover in a Monsoon Region: The Effects of Data Processing Options

3.1 Introduction

Conventional global land use/land cover (LULC) maps are discrete raster maps assigning land cover types to each pixel. Recent techniques allow continuous mapping of land use such as fractional cover. Fractional land cover consists of proportions of non-overlapping land cover types in pixels of a given raster grid (Defries et al., 2000; Price, 1992; Smith et al., 1990). It is often called sub-pixel land cover as it can be conceived as an interpretation of land cover types at the sub-pixel level (Fernandes et al., 2004). It is also called 'continuous fields' Defries et al., 2000; Schwarz et al., 2005. Fractional land cover is increasingly used as a key descriptor of ecosystems and their functions (e.g. Bevanda et al., 2014; Fernandes et al., 2004; Guerschman et al., 2009; Pittman et al., 2010; Schwieder et al., 2014).

In heterogeneous landscapes such as mixed agricultural areas, a substantial number of LULC types often occur in a relatively small area. Therefore, a few general cropland types with spatial resolution up to several hundred meters are insufficient to represent this type of landscape (Schulp et al., 2011) and appropriate land cover information is restricted or often unavailable (Mora et al., 2014; Pittman et al., 2010). Moreover, Herold et al. (2008) showed that the global land cover products are generally limited in representing agriculture-related mixture classes.

Yet, currently available global land cover databases such as GlobCover or Moderate Resolution Imaging Spectroradiometer (MODIS) land cover have only a few crop-related types (Bontemps et

al., 2011; Loveland et al., 2000; U.S. Geological Survey, 2012) and generally lack high resolution maps or fractional land cover data (Colditz et al., 2011; Pittman et al., 2010). For instance, GlobCover 2000 is provided at 300 m resolution and has four cropland or relevant mixture types, and MODIS Land Cover Type (MCD12Q2) product provides five raster land cover layers at 500 m (Bontemps et al., 2011; Loveland et al., 2000; U.S. Geological Survey, 2012), each of which with only one or two cropland types. MODIS Vegetation Continuous Fields product (MOD44B) is the only product that provides fractional cover data. However, in the current version (V005) it is limited to tree-related types, namely "tree", "non-tree", and "bare soil". The limitations of the global land cover databases are particularly pronounced in heterogeneous agricultural landscapes due to the mosaic of crop/non-crop LULC types (Mora et al., 2014).

To retrieve thematically and spatially rich land cover data, one can attempt to extract additional information from existing multi-spectral medium-resolution sensors. Deriving fractional land cover from existing satellite products can enrich the information contents with little additional cost. Furthermore it can be applied to the past-time data. Accordingly, there have been continuous efforts to derive fractional land cover information from existing raster data (Defries et al., 2000; Schwarz et al., 2005). Among various existing sensors, NASA's MODIS (MODerate Resolution Imaging Spectroradiometer) sensor possesses temporal continuity and global coverage. While it has not been commonly used for cropland mapping due to its coarse spatial resolution, it may be able to identify detailed information aided by a methodological elaboration (Pittman et al., 2010).

In this regard, multi-type fractional land cover data can be a valuable information source about agricultural landscapes, especially if such information can be obtained from globally available multi-spectral products. Despite their limited spectral and spatial resolution, MODIS multi-spectral products provide good temporal resolution and can be useful to map agricultural areas (Pittman et al., 2010; Verbeiren et al., 2008). Indeed, MODIS time series contain the complete seasonal dynamics and therefore potentially useful information to distinguish land cover types (e.g. Hüttich et al., 2009; Thenkabail et al., 2005) and has been used to map agricultural LULC types (e.g. Biggs et al., 2006; Brown et al., 2013a; Gumma et al., 2011). Regarding fractional cover, Lu et al. (2003) showed that MODIS time series are suitable to map fractional woody and herbaceous covers.

To develop a fractional land cover model, a number of decisions at the model formulation stage need to be made. First, one needs appropriate predictor data – a difficult choice due to an increasing number of satellite products (e.g. Clark et al., 2010). Second, a suitable algorithm and training parameters should be chosen to avoid sub-optimal performance. Third, pre- and

post-processing strategies should be determined (e.g. Guerschman et al., 2009). We will denote all these decisions 'data-processing options' hereafter.

Improperly selected data-processing options can degrade the model performance by reducing information contained in the data. Optimal data-processing options are case-specific (i.e. dependent on the purpose, cost and processing capacities) (Thackway et al., 2013) thus cannot be universally evaluated. Therefore, in the course of model building, the modeller should select proper data-processing options.

In monsoonal areas, there is a specific problem undermining model performance. In these areas, acquisition of cloud-free data during monsoon is generally difficult due to long-lasting rainfalls (Guerschman et al., 2009; Yihui et al., 2005). For example, South Korean summer shows typical East Asian monsoon weather with persistent and intensive raining period from June to September. This period is called "Changma" (long lasting rain) in Korean literature (Kang et al., 2009). Due to the long-lasting rainfall, cloud-free spectral data are often lacking in the region.

In a heterogeneous agricultural landscape in South Korea, we aim to derive fractional LULC from multi-spectral satellite data using a data mining algorithm. It is challenging because the study area is a complex heterogeneous agricultural landscape. Spectral datasets are supposedly cloud-contaminated because the study area is situated in a monsoon region. In this context, we set up the main objectives as 1) to develop a fractional LULC modelling framework with globally available data (i.e. multi-spectral data) and 2) to evaluate relevant data-processing options, namely selection of spectral predictor sets, time intervals, and smoothing options.

The study is based on the following hypotheses: (1) the full information of a spectral data product (e.g. all available reflectance bands) perform better than a subset of it (e.g. a single reflectance band) or an index function (e.g. NDVI), (2) multi-day composited data with a narrow (e.g. 8-day) composite window (Huete et al., 1999) produce a better regression performance due to more details in the data, and (3) smoothing of input data improves the regression performance because it reduces possible cloud contamination. These hypotheses were chosen in accordance with the characteristics of the study area.

In addition to the main analysis, we assess the relative importance of the spectral bands and the data acquisition dates. Based on the result, we discuss the current capacity and potential of the multi-type fractional cover model in heterogeneous agricultural landscapes.

3.2 Materials and Methods

3.2.1 Study area

The study area Haean-myeon is located at the border between North and South Korea (128°1′33.101″E, 38°28′6.231″N) (Figure 3.1). It is a small agricultural catchment (64.4 km^2) with elevations ranging between 500 m and 1200 m above see level. The catchment is a heterogeneous agricultural landscape comprised of various natural and artificial LULC types. Seo et al. (2014) reported 67 LULC types from a three-year field-level LULC census.

The average air temperature of the study area is 8.5° C at the central plateau. The annual average rainfall equals 1599 mm and the maximum daily rainfall was 223 mm between 1999 and 2010 (Korean Meteorological Administration, http://web.kma.go.kr/eng). The study site belongs to the East Asian summer monsoon (EASM) region (Yihui et al., 2005). More than 60% of annual precipitation is concentrated during the monsoon period from June to August and extreme rainfall events occur frequently.

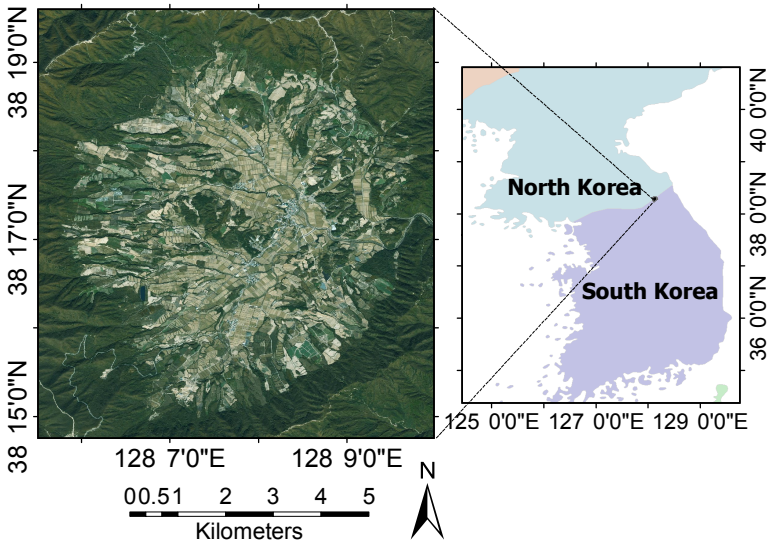

Fig. 3.1 Map and the location of the study site 'Haean' on the Korean peninsula. The catchment is an agricultural hotspot located in the protected temperate forest. The satellite image is a SPOTMaps mosaic product (Astrium Services, http://www.astrium-geo.com) acquired in 2009.

3.2.2 Data

3.2.2.1 Land use/land cover and fractional cover data

For the analysis, we used the LULC polygon data censused in 2010 for the site. The reference data consists of spatial polygons with the observed LULC information and is archived at the public

repository *Pangaea* (Seo et al., 2014). Additionally to the raw LULC type labels, it provides reclassified type labels based on four classification schemes. We used a reclassified LULC labels in a 10-class scheme, which was designed to describe the edaphic and socio-economic conditions of the area. The scheme includes "Barren", "Dry field", "Forest", "Greenhouse", "Inland water", "Inland wetland", "Orchard field", "Paddy field", "Semi natural", and "Urban". This scheme was selected as it distinguishes paddy field from other agricultural types. More details about the LULC data is provided in the meta information of the dataset (Seo et al., 2014).

Due to the bowl-shaped topography of the catchment, LULC types are unevenly distributed (Figure 3.2). The steep slopes and the encompassing mountain ridges are covered by "Forest" consisting of a variety of species from the genus Oak (*Quercus* spp.).

The lower area is dominated by the managed land use types. "Paddy field" occurs at the central plateau while "Dry field" and "Semi natural" dominate on the surrounding lower slopes. The aforementioned four LULC types are large or moderately large in area proportions ($> 8\%$) and cover 95.0% of the total area (Table 3.1). We will denote these types as 'major types'. The rest of the LULC types are smaller in area proportions ($< 2\%$). We denote the next five types "Urban", "Orchard field", "Inland water", "Greenhouse" and "Barren" as 'minor types'. The smallest type by area "Inland wetland" was excluded from the analysis due to its extreme rarity. Note that the selected 9 types make up 99.9% of the study area.

Table 3.1. The land use/land cover types in the Haean catchment in 2010. "Inland wetland" was excluded from the analysis due to its extreme rarity.

Type	Area (km^2)	Area (%)	Category
Forest	37.195	57.805	Major types
Dry field	9.543	14.831	
Semi natural	9.124	14.180	
Paddy field	5.178	8.047	
Urban	1.108	1.723	Minor types
Orchard field	0.952	1.480	
Inland water	0.556	0.864	
Greenhouse	0.544	0.845	
Barren	0.144	0.224	
Inland wetland	0.0004	0.0007	-

Fractional vegetation cover is defined as the sum of the vegetated patch area divided by the total area (Asner et al., 2000; Smith et al., 1990). In a satellite image, it is calculated per pixel and ranges from 0 (0% cover) to 1 (100% cover) (Obata et al., 2012). Similarly fractional LULC can be defined as the sum of the LULC patch area divided by the total area in each pixel of a given

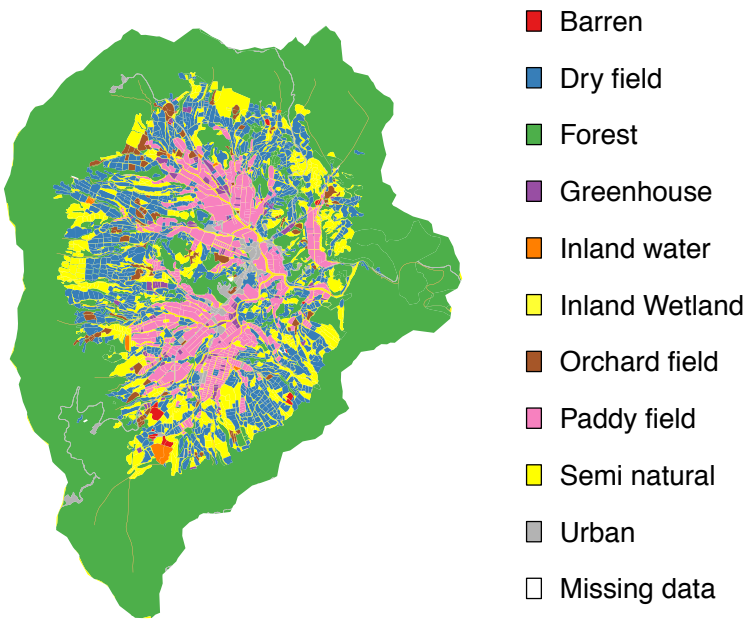

Fig. 3.2 The reference land use/land cover in the Haean catchment in 2010. The reference LULC in cover fraction is shown in Supplementary Figure 3.9.

raster grid (Fernandes et al., 2004). The study site is located in the MODIS tile H28V5 and covered 299 pixels of the 500 m sinusoidal grid (SR-ORG:6842). We chose the 500 m grid as the base grid and derived a per pixel fractional cover data from the observed LULC data. To derive per pixel LULC fractions, we first converted the MODIS raster grid to polygons by pixel (i.e. one polygon per pixel). Then we projected the grid polygons into the WGS84/UTM52N space (EPSG:32652) and overlaid the observed LULC polygons. In the projected space, we calculated the area fractions of the LULC types in all grid polygons (Supplementary Figure 3.9).

3.2.2.2 MODIS spectral data

We used multi-spectral data products as predictors of the fractional LULC model. We chose MODIS collection 5 MOD13A1/MYD13A1 products. Other satellite products such as Landsat Thematic Mapper (https://lta.cr.usgs.gov/TM) are also often used for land monitoring (Vittek et al., 2014; Watts et al., 2010). However, due to its 16-day repeating interval, Landsat products are severely cloud contaminated. For the study area, the Landsat 5 collection at NASA EOSDIS system (http://reverb.echo.nasa.gov) provides only a few cloud free images in 2010. In contrast, the MODIS 16-day products are less cloud contaminated due to daily acquisition and its composition procedure (NASA Land Processes Distributed Active Archive Center (LP DAAC), 2013a).

MOD13A1/MYD13A1 products supply 23 scenes/year at 500 m resolution each. Note that a

time series of MOD13A1 starts from the first day of a year but MYD13A1 from the 9th day. Hence there is an 8-day difference in acquisition date (Didan et al., 2006). Each product contains 12 Science Data Sets (SDS) (NASA Land Processes Distributed Active Archive Center (LP DAAC), 2013a). Among the SDSs, we chose four surface reflectance bands (B1–3, B7), Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and vegetation index quality assurance (QA). The six biophysical SDSs were used as predictors for regression. The QA SDS was used as data weight only for smoothing. For simplicity, we will denote all the biophysical SDSs as spectral bands in the following.

Each spectral band delivers specific information about land cover (Didan et al., 2006). The red band 1 (B1) is sensitive to vegetation chlorophyll and its wavelength is 620–670 nm. The near-infrared (NIR) band 2 (B2) covers 841–876 nm and has been widely used to evaluate ground vegetation viability together with B1. Band 3 (B3) is commonly called the blue band as it is sensitive to water vapour; its wavelength ranges between 459 and 479 nm. The mid-infrared band 7 (B7) with wavelengths between 2105 nm and 2155 nm contains information about land and cloud properties.

NDVI and EVI are vegetation indices designed to capture above ground vegetation properties and biophysical processes (Didan et al., 2006; Huete et al., 1999). The indices are calculated from the reflectance values. NDVI is defined as a function of the red (B1) and the NIR (B2) bands:

$$NDVI = \frac{B2 - B1}{B2 + B1}. \tag{3.1}$$

EVI is designed to remove soil and atmospheric contamination by incorporating additional terms and makes use of the blue band (B3) (Huete et al., 1999). The MODIS EVI is derived as

$$EVI = G \cdot \frac{B2 - B1}{B2 + C_1 \cdot B1 - C_2 \cdot B3 + L}, \tag{3.2}$$

where L is the canopy background adjustment; C1 and C2 are the coefficients to correct for aerosol influences; and G is a scaling factor. The coefficients used in the MODIS EVI algorithm are $L = 1$, $C_1 = 6$, $C_2 = 7.5$, and $G = 2.5$ (Huete et al., 1999).

We acquired the MODIS products from NASA Land Processes Distributed Active Archive Center (LP DAAC) at the USGS/Earth Resources Observation and Science (EROS) Center (https://lpdaac.usgs.gov).

3.2.3 Scenarios

We considered three key data-processing options: predictor set, time interval, and smoothing. Each option comprises several choices. From all combinations of the three options, we formulated 16 scenarios (Table 3.2) and evaluated them using an 16-fold cross-validation (CV) (Section 3.2.4.2). The efficacy of a data-processing option was estimated by the average performance of the associated scenarios. The overall research procedure is illustrated in Figure 3.3.

Predictor set We prepared four predictor sets to compare model performance based on different spectral data. The predictor sets 'NDVI' and 'EVI' contained a corresponding vegetation index data. 'SR' predictor set contained the four surface reflectance bands (B1–B3, and B7). The predictor set 'Full' incorporates all the six available data bands.

Time interval Spectral input data was prepared in 8-day and 16-day intervals. For 16-day input, we simply used MOD13A1 data. For 8-day input, we merged MOD13A1 and MYD13A1 products to produce a quasi 8-day MODIS 13A1 data using the 8-day difference in acquisition date described in Section 3.2.2.2. This results in 46 (8-day) or 23 (16-day) data points per band for each MODIS pixel.
Note that we used the quasi 8-day data instead of the 8-day MODIS products (MOD/MYD09A1). This is because we want to use the 8-day data most similar to the 16-day data. Additionally, the 09A1 products lack NDVI and EVI data sets.

Smoothing We prepared spectral input data with and without smoothing. By comparing the two input data sets, we evaluated the efficacy of data smoothing in a monsoonal catchment. We chose the 'Savitzky-Golay' (SG) filter (Savitzky et al., 1964), which is widely used for smoothing time series data in remote sensing (e.g. Fontana et al., 2008). The filter is designed to retrieve the upper envelope of a time series by using a local polynomial regression iteratively to fit the time series (Hird et al., 2009). It can filter out negatively biased noises (e.g. NDVI decreases due to cloud contamination), which can be useful in monsoonal regions. Fontana et al., 2008
We used the adaptive SG filter provided by the software TIMESAT 3.1 (Eklundh et al., 2012; Jonsson et al., 2004). The seasonal course of the spectral data was smoothed per pixel independently. The MODIS QA data layer was used to weight the values. By the weighting, a data point acquired under a non-optimal condition (e.g. cloudy weather) had only 10% of influence during the smoothing process, compared to data acquired under optimal conditions (e.g. sunny weather). The TIMESAT smoothing parameters were determined according to the software

manual (Eklundh et al., 2012). The size of the fitting window was 3 for 16-day data and 5 for 8-day data. The adaptation strength was 1.5 and the number of envelope iterations was 3. Note that the 3-year data (2009–2011) was processed concurrently as the software encourages to use a longer time series than the target data.

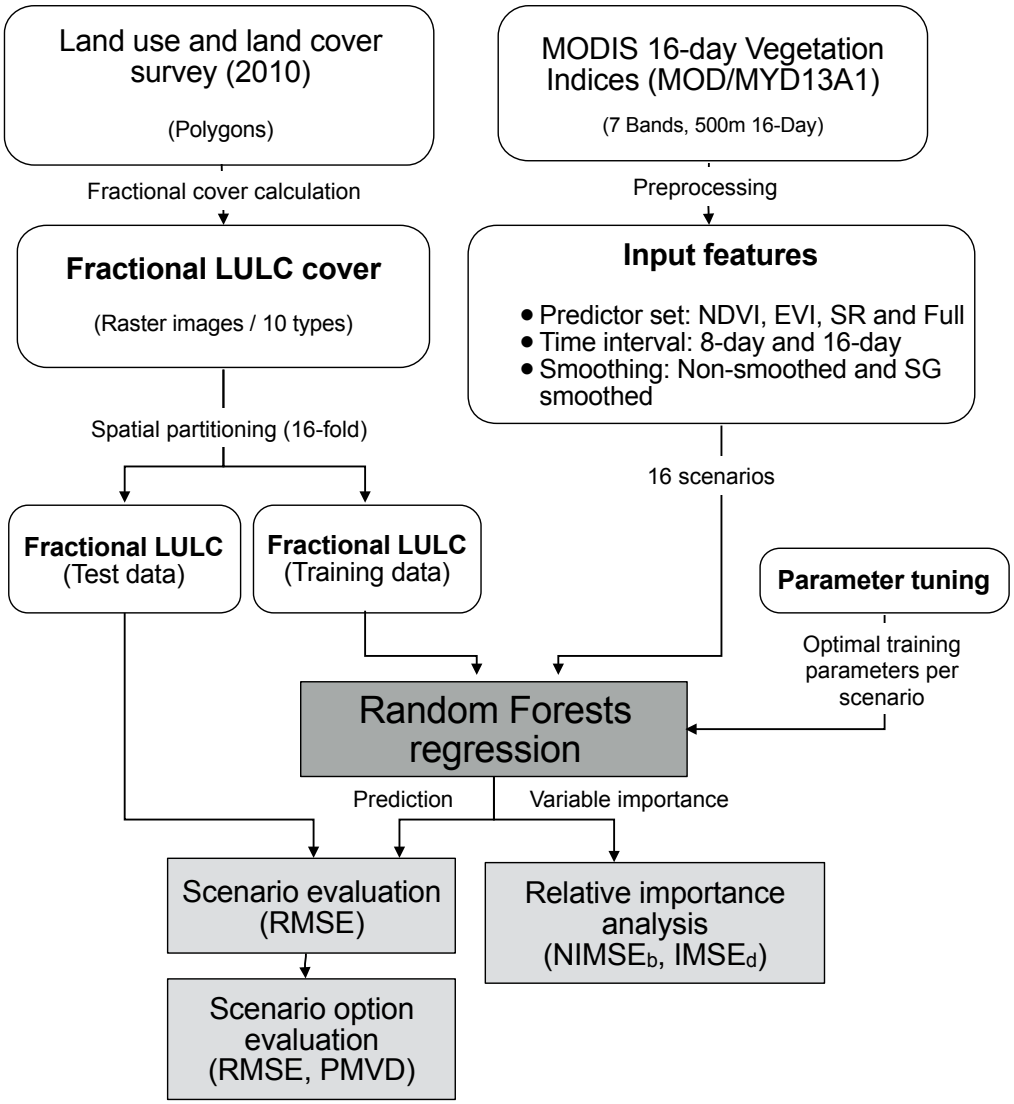

Fig. 3.3 Overview of the fractional cover regression model building and evaluation procedure.

3.2.4 Model construction

3.2.4.1 Random Forest regression

We hypothesised that per pixel LULC fractions can be retrieved from spectral data in line with the previous studies (e.g. Colditz et al., 2011; Guerschman et al., 2009; Lu et al., 2003; Obata et al., 2012; Schwieder et al., 2014). Modelling multi-type LULC fractions can be conceived as a multi-output regression task. This task can be accomplished either by simultaneously modelling a multi-output response, or by separately modelling single-output responses and aggregating

Table 3.2. Specification of the scenarios in combinations of the predictor set, time interval, and smoothing options.

Smoothing		No smoothing		Savitzky-Golay (SG) smoothing	
Time interval		8-day	16-day	8-day	16-day
Predictor set	NDVI	S1	S5	S9	S13
	EVI	S2	S6	S10	S14
	SR	S3	S7	S11	S15
	Full	S4	S8	S12	S16

the outcomes (Hothorn et al., 2006; Segal, 2004). In this study, we decomposed the multi-type fractional cover regression task into a set of single-type regression tasks. Accordingly, we built a fractional cover model for each LULC type and aggregated the model outcomes.

Fractional cover regression can be implemented via various techniques. The techniques include the fuzzy classifier (Foody et al., 1996), the time series model (Lu et al., 2003), linear models (DeFries et al., 1995; Schwarz et al., 2005), data mining algorithms (Fernandes et al., 2004; Schwieder et al., 2014), and spectral mixture analysis (SMA) (Asner et al., 2000; Guerschman et al., 2009). Among various techniques, we used the regression mode of Random Forest (RF). RF is a decision-tree based ensembling algorithm that uses bootstrap aggregation (bagging) and the random sub-space method (Breiman, 2001; Prasad et al., 2006). It is suitable for modelling non-linear relationships and can handle a large number of covariates as it tends not to overfit the data (Breiman, 2001; Prasad et al., 2006; Segal, 2004). Its performance is comparable to the other state-of-the-art learning algorithms such as support vector machine or neural networks (Attarchi et al., 2014; Gislason et al., 2006; Prasad et al., 2006; Schwieder et al., 2014). Moreover it is convenient to set up compared to other data mining algorithms as it has a small number of training parameters Liaw et al., 2002.

In land cover modelling, Random Forest (RF) has been used to classify land cover (Clark et al., 2010; Gislason et al., 2006; Hüttich et al., 2009; Nitze et al., 2015; Rodriguez-Galiano et al., 2012; Thenkabail et al., 2005), vegetation type (Hüttich et al., 2009; Immitzer et al., 2012; Senf et al., 2013), and also crop type (Ghimire et al., 2010; Nitze et al., 2012). In fractional land cover regression, Schwieder et al. (2014) used RF to estimate shrub cover fractions in which RF showed comparable performance with support vector machine and partial least squares regression.

3.2.4.2 Spatial cross-validation

Due to the bagging and the random sub-spacing of RF (Breiman, 2001), the bootstrap samples for training (in-bag data) can be correlated with the test samples (out-of-bag data), especially for spatial models (Brenning, 2005). To avoid dependencies between training and test data, we externally partitioned training and test data by a spatial partitioning scheme utilised by Reineking et al. (2010). The spatial partitioning was implemented in our study as follows. First, we binary split the whole area six-times recursively. The recursive split divides the catchment into 64 sub-clusters. Second, we form 16 clusters by randomly sampling four sub-clusters for each; one cluster is comprised of four spatially disjointed sub-clusters as distinguished by different colours in Supplementary Figure 3.10.

3.2.4.3 Fractional cover estimation

Let T be the number of LULC types such that each type i has a set $F_i = \{f_{i,1}, ..., f_{i,n}\}$ of n observed LULC fractions, where $f_{i,j}$ is the fractional area of the pixel j covered by the LULC type i, and n is the total number of pixels belonging to the study area.
A LULC fraction $f_{i,j} \in [0, 1]$ and all fractions of one pixel sum up to one

$$\sum_{i=1}^{T} f_{i,j} = 1 \tag{3.3}$$

for all $j = \{1, ..., n\}$.
First we built a RF regression model per type. Given a type i, we used the observed fraction $F_i = \{f_{i,1}, ..., f_{i,n}\}$ as response and a set of feature vectors $P = \{p_1, ..., p_n\}$ as predictor. Each feature vector contained $n_{feature}$ features varied by the spectral data used (Supplementary Table 3.4).
The regression model was trained/tested with a 16-fold cross validation (c.f. Section 3.2.4.2 for details). By accumulating test pixels of all CV folds, we obtained the predicted fractions $\hat{F}_i = \{\hat{f}_{i,1}, ..., \hat{f}_{i,n}\}$ of the type i over the entire study area. Note that RF produces predictions from all regression trees (Breiman, 2001), therefore for each pixel n_{tree} fractions were predicted, where n_{tree} is the total number of regression trees. We took the mean of the n_{tree} predictions. This generated a set of LULC fractional cover for the study area.
Then we normalised the type-wise predictions by Eq. 3.3. The normalised prediction $\hat{F}_i^*$ was calculated as

$$\hat{F}_i^* = \frac{\hat{F}_i}{\sum_{j=1}^{T} \hat{f}_{i,j}} \tag{3.4}$$

where $\hat{F}_{i,j}$ is the type-wise prediction of the type i for the pixel j. Finally we obtain the predicted LULC fractions $\hat{F}^* = \{\hat{f}_1^*, ..., \hat{f}_T^*\}$.

3.2.4.4 Training parameters

RF has three training parameters: the number of trees in the forest (n_{tree}), the number of randomly selected variables on each split (m_{try}), and the number of minimal samples in terminal nodes (*nodesize*). These parameters need to be tuned to avoid sub-optimal model performance Rodriguez-Galiano et al., 2012; Strobl et al., 2008.

To find the optimal n_{tree} and *nodesize* we performed a grid search on the training folds. We used a grid from all combinations of n_{tree}= $\{100, 200, ..., 1000\}$ and *nodesize*= $\{1, 2, 3, 4, 5\}$. Grid searching was implemented using an internal validation. We repartitioned the training data folds into a new training data and a new test data. The new test data contained two spatial clusters, randomly selected without replacement. We trained the model on the new training data with different parameter values and predicted the hold-out data. This was repeated for all 9 types and we calculated the mean root mean square error (*RMSE*) over all types. Overall, the model performance improved with large n_{tree} and small *nodesize* (Supplementary Figure 3.11).

We optimised n_{tree} and *nodesize* separately based on its marginal *RMSE* on the tuning grid. We chose parameters by minimising the marginal error metric unlike Rodriguez-Galiano et al. (2012) or Leutner et al. (2012) who used the joint error metric on the grid. We tried both approaches but opted for the marginal error based selection. Compared to the joint error based selection, the marginal error based selection was less sensitive to the between-partition variations. In consequence, it led to more stable parameter selection between scenarios.

The parameter m_{try} was determined by the square root of $n_{feature}$ without grid searching as in (Clark et al., 2012). Since the scenarios have unequal number of input features, m_{try} varied between scenarios. The chosen parameter values are summarised in Supplementary Table 3.4.

3.2.5 Model evaluation

3.2.5.1 Overall regression performance

We used the cross-validation error metrics instead of the default out-of-bag (*OOB*) error of RF. As discussed in Section 3.2.4.2, the *OOB* error can be biased due to a possible correlation

between in-bag training samples and out-of-bag test samples, especially for spatial models. Instead, we used cross-validation *RMSE* to evaluate regression performance. The *RMSE* of the LULC type i is calculated as

$$RMSE_i = \sqrt{\frac{\sum_{j=1}^{n}(f_{i,j} - \hat{f}^*_{i,j})^2}{n}} \tag{3.5}$$

where $f_{i,j}$ is the observed and $\hat{f}^*_{i,j}$ is the predicted LULC fraction for the type i in pixel j, and n is the total number of pixels.

Furthermore, we used the coefficient of determination (R^2) and Spearman's rank correlation coefficient (ρ) (Gibbons et al., 2003) also based on cross-validation. The R^2 was used to compare our results with the previous studies on fractional cover estimation (e.g. Fernandes et al., 2004). Spearman's ρ was used to estimate the association between observed and predicted fractions (Gibbons et al., 2003).

3.2.5.2 Relative contribution of data-processing options

Additionally to cross-validation error, we examined the relationship between the data-processing options and the performance of the fractional cover regression models. For this analysis, we built a linear model explaining the *RMSE* of the regression model for each LULC type by the different data-processing options:

$$RMSE_i = \beta_0 + \beta_1 O_p + \beta_2 O_t + \beta_3 O_s + \epsilon \tag{3.6}$$

where $RMSE_i$ is the *RMSE* of the type i; O_p is a categorical variable denoting the chosen predictor set option, O_t time interval option, and O_s smoothing option; ϵ is the error term. Note that we did not include interaction terms based on the preliminary model selection using F-statistics (not shown here).

We assumed that the 'relative contribution' ('relative importance' in Grömping (2006)) of an option is that of the corresponding regressor to the linear model. Then we quantified relative contributions of the regressors by decomposing the amount of explained variance of the linear model due to regressors. We used proportional marginal variance decomposition (PMVD) method (Feldman, 2005; Grömping, 2006) which decomposes the explained variance of the linear model into non-negative contributions, which sum to the total variance explained. PMVD is able to deal with correlated regressors by averaging over different orderings. Moreover, it has desirable properties such as 'admissibility'.

Each linear model (per type) was estimated based on the 16 samples from all 16 scenarios. Statistical significance of the type-wise models were tested using F-statistics to validate the model structure.

3.2.5.3 Marginal performance of data-processing options

The efficacy of a data-processing option was estimated by average regression performance of the scenarios using the option. We will call it 'marginal performance' in the following. The marginal performance (M) of a data-processing option k for a performance metric q is calculated as

$$M_{k,q} = \frac{\sum_{x \in s^k} q(x)}{|s^k|} \tag{3.7}$$

where s^k is a set of scenarios using the option k and $|s^k|$ is the number of elements of s^k.

3.2.5.4 Relative importance of spectral bands and acquisition dates

We quantified the relative importance of spectral bands and acquisition dates to identify the most relevant ones for the regression performance. RF provides two importance metrics for quantifying the influence of input features (Breiman, 2001; Segal et al., 2011). Among the metrics, we used the increased mean square error (*IMSE*), which is a permutation-based measure. Another metric namely increased node purity (*INP*) is measured by node purity, in case of regression the residual sum of squares. We avoided to use *INP* because of the possible bias due to the random subspacing (i.e. random selection of features). For classification, the *INP* is known to be biased as the impurity measure (Gini index) favours predictor variables with many categories (Genuer et al., 2010; Strobl et al., 2007). *IMSE* of a feature f is derived as

$$IMSE_f = \frac{\sum_{k=1}^{n_{tree}} (\overline{MSE_k} - MSE_{f,k})}{n_{tree}} \times \frac{1}{\sqrt{s^2/n_{tree}}} \tag{3.8}$$

where n_{tree} is the size of the forest, $\overline{MSE_k}$ is the mean squared *OOB* error of tree k, $MSE_{f,k}$ is the error after permuting the feature f and s^2 is the standard deviation of the differences between the two errors; if s^2 is zero, the division is omitted. Due to the cross-validation scheme, we computed $IMSE_f$ in each cross-validation fold and averaged them. Note that the variable importance metric is calculated from the *OOB* samples.

Our goal was to assess the relative importance of spectral bands and acquisition dates on the regression model. As we used the time series of multiple spectral bands, input features can be grouped either by band or by acquisition date.

We defined the importance of a band as the sum of the importance metrics of the features belonging to the band. Let a predictor set $X = \{x_1, ..., x_l\}$ have l features some of which belong to a spectral band b. We calculated importance of the band b as

$$IMSE_b = \frac{\sum_{x \in b} IMSE_x}{l_b} \tag{3.9}$$

where l_b is the number of the features belonging to the band.

To facilitate comparisons between different bands, we normalized $IMSE_b$ as

$$NIMSE_b = \frac{IMSE_b}{\sum_{b=1}^{n_{band}} IMSE_b} \tag{3.10}$$

where n_{band} is the number of the bands in a predictor set. To derive $NIMSE_b$ we used the two groups of the scenarios: scenarios using 'SR' predictor set (S3, S7, S11, and S15) and scenarios using 'Full' predictor set (S4, S8, S12, and S16). As they are different in the number of spectral bands, we calculated two sets of $NIMSE_b$. For each group individually, we calculated the mean importance measures from the included scenarios.

Likewise the importance of an acquisition date is defined as the sum of the importance metrics of the features acquired at a particular date d:

$$IMSE_d = \frac{\sum_{x \in d} IMSE_x}{l_d} \tag{3.11}$$

where l_d is the number of the features acquired at the date d. To derive $IMSE_d$ we used the 'Full' predictor set based scenarios (S4, S8, S12 and S16). As 8-day and 16-day data differ in the number of data points, we extracted two seasonal $IMSE_d$ curves individually by interval.

3.2.6 Software

We used `GNU R` 3.1.2 (R Core Team, 2014) and the `R` packages `randomForest` version 4.6–7 (Liaw, 2012), `raster` version 2.3–40 (Hijmans, 2014), and `relaimpo` version 2.2–2 (Grömping, 2006). The geometry engine `GEOS` 3.4.2 (GEOS Development Team, 2014) was used via the `R` package `rgeos` 0.3–8 (Bivand et al., 2014) and the software TIMESAT 3.1 (Eklundh et al., 2012) for smoothing the spectral data.

3.3 Results

3.3.1 Overall regression performance

The average performance of all scenarios in *RMSE*, ρ, and R^2 were 0.057, 0.624, and 0.414, respectively (Table 3.3). The best scenario S4 used 'Full' predictor set in '8-day' interval with 'No smoothing'. The worst scenario S14 used 'EVI' predictor set in '16-day' interval with 'SG smoothing'. Maps of the modelled LULC fractions are provided in Supplementary Figures 3.12 and 3.13 for averaged and for the best scenario, respectively.

Table 3.3. Fractional LULC regression performance by scenario. All the performance metrics were averaged over LULC types.

Name	Data-processing options			Model performance		
	Predictor set	Time interval	Smoothing	*RMSE*	ρ	R^2
S1	NDVI	8-day	No smoothing	0.056	0.658	0.428
S2	EVI			0.057	0.639	0.438
S3	SR			0.054	0.657	0.441
S4	Full			0.053	0.663	0.455
S5	NDVI	16-day		0.056	0.630	0.410
S6	EVI			0.060	0.601	0.395
S7	SR			0.056	0.638	0.430
S8	Full			0.055	0.634	0.434
S9	NDVI	8-day	SG smoothing	0.058	0.618	0.399
S10	EVI			0.059	0.601	0.389
S11	SR			0.054	0.633	0.411
S12	Full			0.053	0.634	0.434
S13	NDVI	16-day		0.061	0.588	0.364
S14	EVI			0.064	0.572	0.347
S15	SR			0.057	0.611	0.418
S16	Full			0.056	0.609	0.424
	Avg.			0.057	0.624	0.414

624

3.3.2 Type-wise regression performance

Spearman's rank correlation between the observed and the predicted LULC fractions were high on the average (avg. $\rho = 0.624$; Table 3.3 and Supplementary Figure 3.14). Not only for the major types but also for some of the minor types the rank correlations were rather high (Supplementary Table 3.5). For example, ρ was 0.48 for "Orchard field" and 0.54 for "Inland water", for which predicting absolute fractions were not at all successful ($R^2 < 0.10$). Similarly, for "Greenhouse" rank correlation $\rho(= 0.59)$ indicates better model performance than which suggested by $R^2(= 0.25)$. This implies that the regression model may be useful to detect minor types (i.e. binary classification).

To further investigate the performance degradation of the minor type models, we analysed the relationship between R^2 and the total area proportions of the LULC types (Figure 3.4). R^2 increased with increasing area proportion. Since the minor LULC types occurred only sporadically over the area, a large number of pixels have zero fraction for the minor types. Therefore, the distribution of the observed fractions of minor types was right-skewed (Supplementary Figure 3.15a).

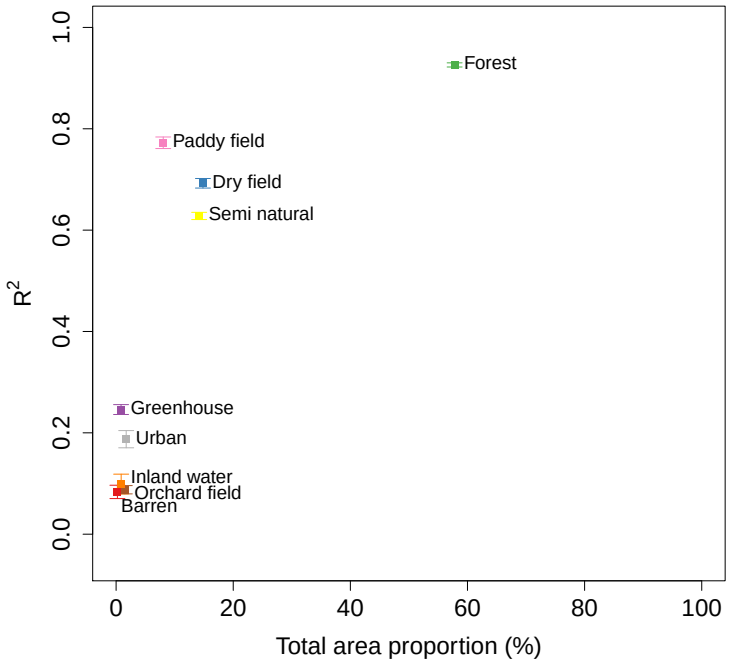

Fig. 3.4 Observed total area proportions of the LULC types are plotted against the mean type-wise R^2 over all scenarios. The area proportions were calculated at the catchment level. The error bars indicate the standard errors of the means over the scenarios.

3.3.3 Relative contribution of data-processing options

Relative contributions of the data-processing options are shown in Figure 3.5 and Supplementary Table 3.8. The linear models explaining type-wise *RMSE* by data-processing options were all significant ($p < 0.05$) except for "Barren".

For the 9 types averaged, 73.2% of the variance of the *RMSE* was explained by predictor set (O_p; 36.3%), time interval (O_t; 19.0%) and smoothing (O_s; 17.9%), respectively.

Among the three options, O_p was of the highest contribution for "Forest", "Dry field", "Paddy field", "Urban", and "Greenhouse". "Semi natural" and "Inland water" were most attributed by O_t and "Orchard field" by O_s.

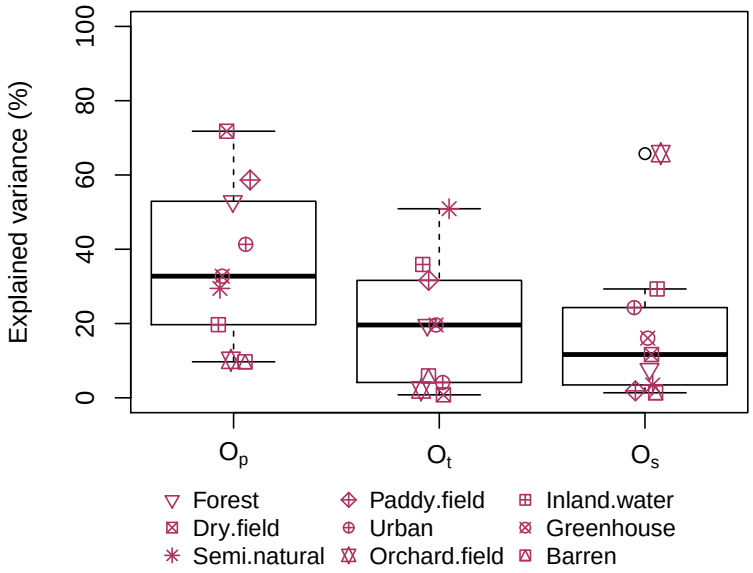

Fig. 3.5 Relative contribution of the data-processing options in explaining *RMSE* in a linear regression model per type. O_p is a categorical variable denoting the chosen predictor set option, O_t time interval option, and O_s smoothing option. The relative contributions were calculated by proportional marginal variance decomposition (PMVD) (Feldman, 2005). The 9 points per option represent the 9 LULC types.

3.3.4 Marginal performance of data-processing options

Among the four predictor set options, 'Full' predictor set based scenarios achieved the best average *RMSE* (0.054) followed by 'SR' predictor set based scenarios (0.055). Between the vegetation indices, the marginal *RMSE* of the predictor set 'NDVI' was smaller (0.058) compared to 'EVI' (0.060).

The ranks of the predictor sets varied between the LULC types (Figure 3.6a). The 'Full' predictor set was the best set for 6 out of 9 types. Although, "Greenhouse" and "Barren" were best

predicted by 'SR' predictor set, the differences between the predictor sets were small. The single vegetation index predictor set 'EVI' was the best predictor set for "Inland Water".

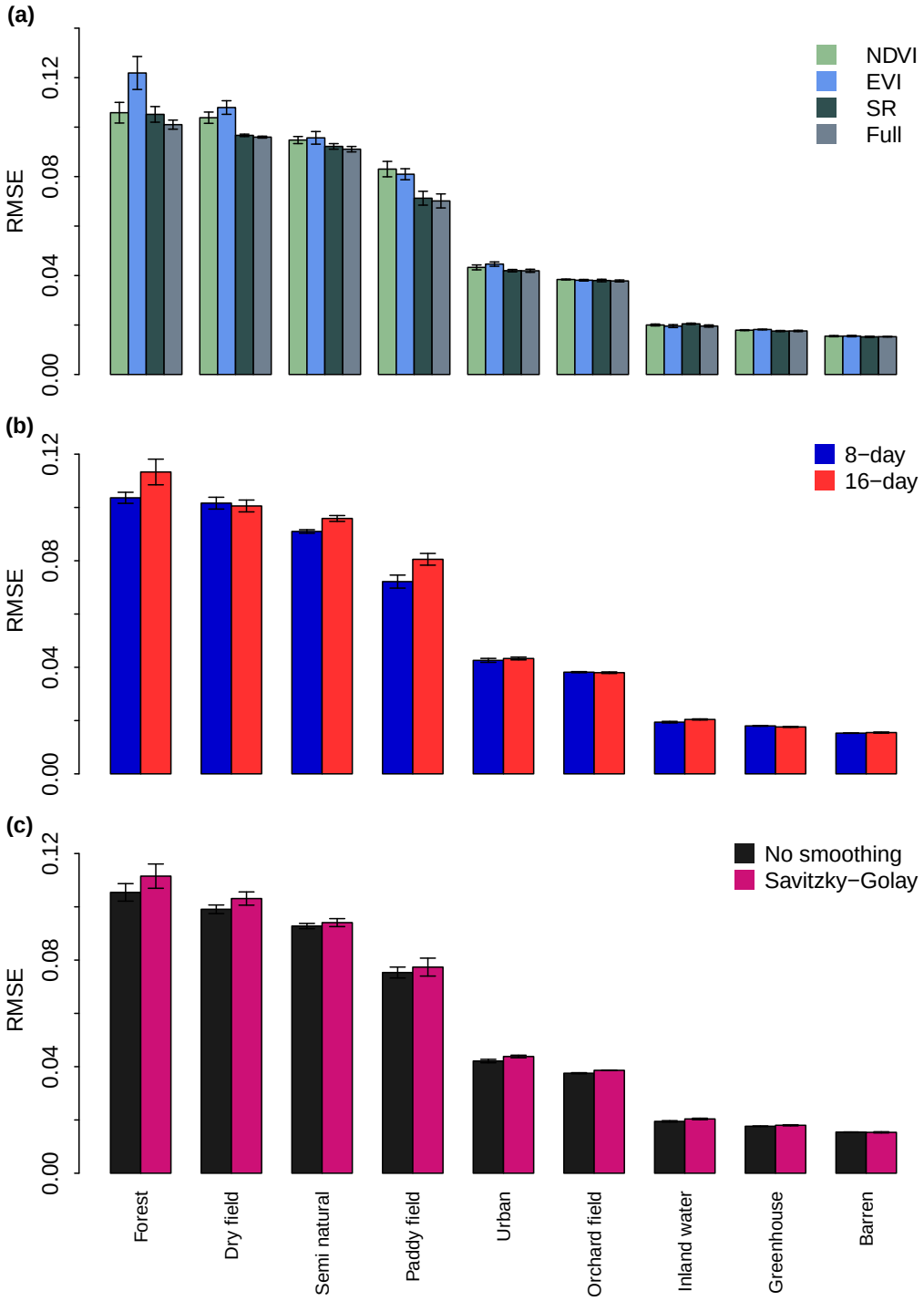

Fig. 3.6 Performance of the data-processing options measured by marginal *RMSE*: (a) predictor set, (b) time interval, and (c) smoothing. The cross-validated regression metrics were averaged over the other data-processing options to derive marginal performance metrics (3.7). The bars indicate standard errors of the mean.

Regarding the time interval, the 8-day scenarios (mean $RMSE = 0.056$) marginally outperformed the 16-day scenarios (mean $RMSE = 0.058$) (Figure 3.6c and Table 3.3). This does not hold for

the LULC types “Dry field”, “Orchard field” and “Greenhouse”. These types are minor types except “Dry field”.

The scenarios with ‘No smoothing’ performed better (mean $RMSE = 0.056$) than the SG smoothed scenarios (mean $RMSE = 0.058$) (Figure 3.6c). For the individual types, the non-smoothed predictors performed better except for “Barren” (Table 3.3).

3.3.5 Relative importance of spectral bands

The mean relative importance of the spectral bands were calculated with the ‘Full’ predictor set based scenarios (Figure 3.7a) and ‘SR’ predictor set based scenarios (Figure 3.7b).

Using the variable importance metric from ‘SR’ predictor set based scenarios, we assessed the relative importance of the four reflectance bands when used with no vegetation index (Figure 3.7b). On the average, the $NIMSE_b$ of B1 (48.6%) and B2 (46.9%) were substantially higher than that of B3 (2.2%) and B7 (2.3%) and made up 95.5% of the total $IMSE$ (Supplementary Table 3.6). The two bands were almost equally important among all LULC types.

For the most dominant type “Forest”, $NIMSE_b$ of B3 (11.0%) and B7 (12.3%) were relatively large compared to that of the rarer types. However, especially for the five rarest types, B3 and B7 were negligible with less than 0.5% of $NIMSE_b$.

In ‘Full’ predictor set based scenarios, NDVI, EVI and B1 bands were similar in $NIMSE_b$ (31–33%) and made up 96.5% of the total $IMSE$ (Figure 3.7a and Supplementary Table 3.7). After including NDVI and EVI, B2 became negligible (1.3%), while B1 remained important (31.8%). The contribution of B3 and B7 stayed small with an $NIMSE_b$ equalled to 1.0% and 1.1% respectively.

Only the major types such as “Forest” or “Dry field” benefitted from the bands B2, B3 and B7. The $NIMSE_b$ of these three bands were smaller than 0.2% for the minor types.

3.3.6 Seasonal variation of relative importance

Figure 3.8 shows seasonal variation of $IMSE_d$ by type. Both in 8-day and 16-day intervals, we observed large variable importances in the off-monsoon periods like the start and the end of the growing season. The $IMSE_d$ during the summer monsoon season around day of a year (DOY) 200 were rather low for most of the LULC types, suggesting that the features representing this period were less influential on the regression performance.

In a large portion of the types, peaks are found in March (around DOY 90), which is the sowing season in the study area. Other peaks commonly occurred in September, which is the harvest

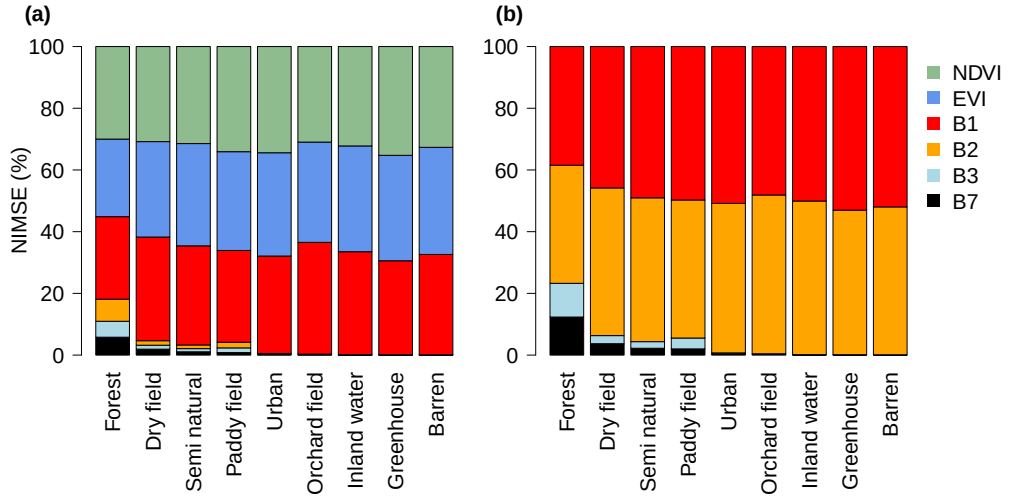

Fig. 3.7 Normalised increased mean square error ($NIMSE_b$) of spectral bands from (a) 'Full' predictor set based scenarios (S4, S8, S12, and S16) and (b) 'SR' predictor set based scenarios (S3, S7, S11, and S15).

season for most of the local crops (e.g. paddy rice and annual dry field crops) as well as the senescence of natural vegetation types.

The shapes of the seasonal $IMSE_d$ curves differed between the LULC types. For instance, the seasonal $IMSE_d$ of "Paddy field" showed the highest peak in September (around DOY 260) (Figure 3.8d), which shows that the model is most sensitive to the harvest season. In contrast, "Forest" exhibits the highest peak in late February (around DOY 80) (Figure 3.8a).

The number of major peaks of relative importance was different between types. The $IMSE_d$ of "Dry field" and "Semi natural" can be characterised as bimodal because of the two peaks around the sowing season (around DOY 60) and the harvest season (around DOY 260). However, for rarer types such as "Inland water" or "Greenhouse", relative importance curves display multiple peaks both in 8-day and 16-day $IMSE_d$ curves.

3.4 Discussion

3.4.1 Regression performance

The regression performance of the major type models was comparable to previously published studies. Schwieder et al. (2014), for example, reported a mean $R^2 = 0.60$ for a fractional shrub cover model using three machine learning algorithms including RF. Verbeiren et al. (2008) confirmed that, at sub-pixel level, land cover estimation with multiple types is challenging; a mean R^2 of the fractional cover estimation with 8 types was 0.41 using a neural network model and 0.29 using a linear mixture model. It is similar to the R^2 of the major type models

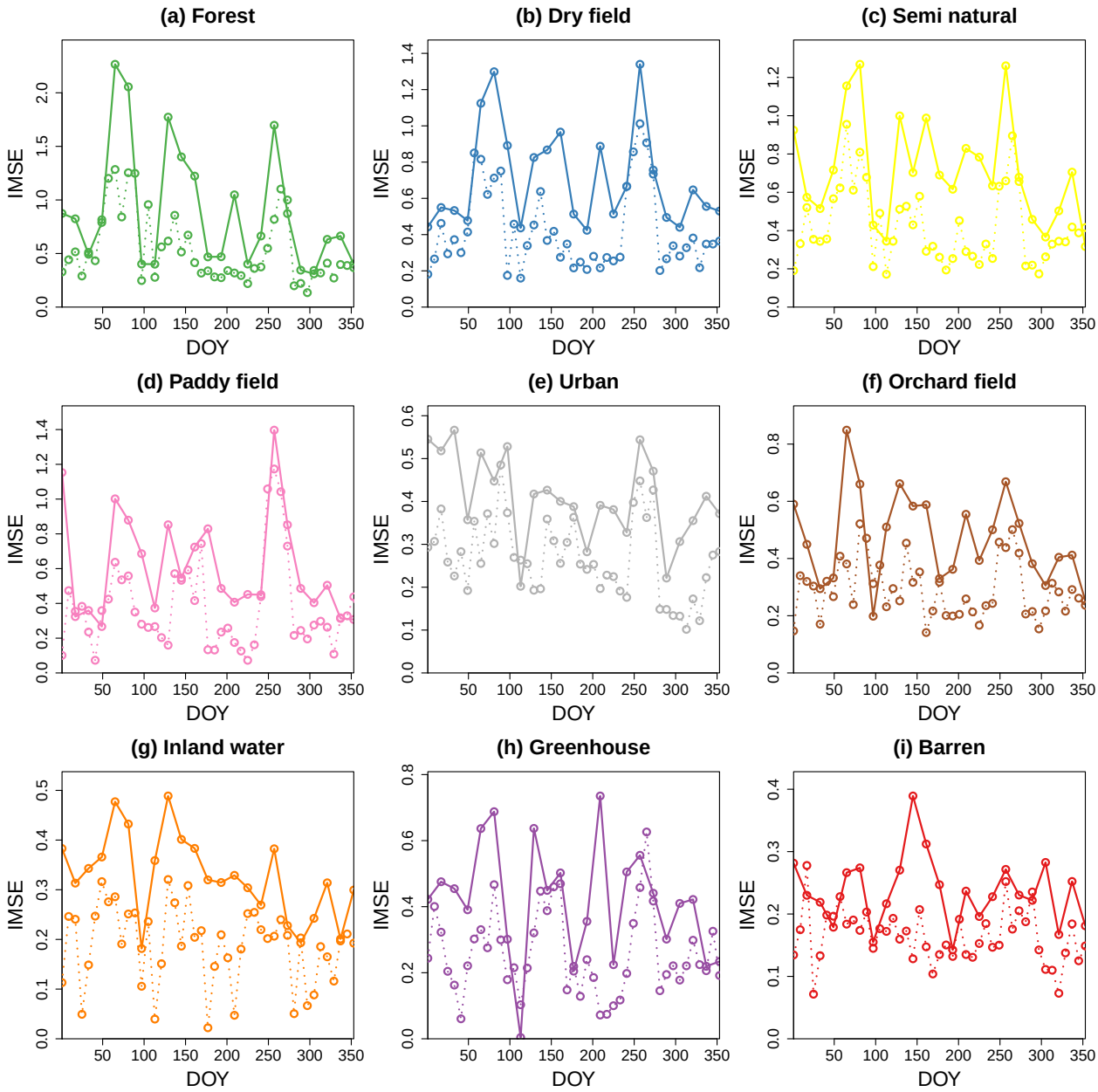

Fig. 3.8 Seasonal variations of increased mean square error ($IMSE_d$) are displayed to visualise relative importance of the acquisition dates; dotted line indicates the $IMSE_d$ from the 8-day data based scenarios and solid line from the 16-day data based scenarios. Note that we used only 'Full' predictor set based scenarios (S4, S8, S12 and S16).

(> 0.6) (Supplementary Figure 3.14 and Supplementary Table 3.5). Xiao et al. (2005) reported higher $R^2 (> 0.75)$ for fractional green vegetation cover estimations, however their model was not validated against ground observation and/or with cross-validation.

A regression task with multiple responses is inherently more difficult than a single-response regression. Our results are comparable to the work by Colditz et al. (2011), for example, who used a 14-class land cover system in South Africa (total accuracy = 55.0%) and Germany (total accuracy = 51.6%). Note that type-wise regression performance was missing in their study.

Fernandes et al. (2004) reported that the regression of fractional covers of minor types was more difficult; the average predictive R^2 was 0.57 for the two dominant types ("Conifer forest" and "Shrub") while 0.33 for the three minor types ("Deciduous forest", "Barren" and"Water"). Dennison et al. (2003) reported comparable overall accuracy (55.9%) from a fractional cover model with 6 LULC types. Note that their models were evaluated without cross-validation.

We attribute the low performance to the right-skewed distributions of LULC fractions in the training data (Supplementary Figure 3.15a). Since the minor LULC types occurred only sporadically over the area, a large number of pixels have zero fraction for the minor types.

When training data are skewed, a RF regression model has a limitation in prediction due to the way how regression trees are constructed. If the training data is right-skewed or even zero-inflated, the model is insufficiently trained on the high response values (e.g. high LULC fractions). As discussed in Section 3.2.4.3, the prediction is the average response of all trees. Thus, RF does not search for the best tree but averages all trees. When trained with the skewed data, it can cause an underestimation bias in prediction. O'Leary et al. (2009) noted the same issue in RF classification when training data is imbalanced.

Our result confirm that minor types are difficult to estimate in fractional cover studies and thus need more attention. It is even more important to resolve the issues related to minor types in agricultural areas. Due to fragmented land use patterns and heterogeneities embedded in land cover classification systems (e.g. lumped cropland types), minor types are inevitably occurring in this type of landscape. To our knowledge, there were only few studies dealing with multiple LULC types in continuous land cover studies and the case studies generally suffer from poor performance regarding agricultural types (e.g. Dennison et al., 2003; Verbeiren et al., 2008) and often lack appropriate model validation (e.g. Colditz et al., 2011; Johnson et al., 2012; Xiao et al., 2005).

The two-step modelling approach such as the Hurdle model (Mullahy, 1986) may alleviate the issue of minor types in fractional cover estimation. In the Hurdle model approach, first occurrence of a desired response (e.g. LULC type) is modelled and the degree of the response is estimated for the instances passed the first 'hurdle'. This approach may alleviate the issue of the right-skewed training data. However, the issue of the missing high response values in training data needs to be resolved independently.

The Hurdle model can be used in combination with machine learning (e.g. Lieske et al., 2014; Povak et al., 2013). Fractional LULC regression with the Hurdle formulation would be an interesting future work.

3.4.2 Relative importance

In our case, the information contained in the red channel (B1) was not perfectly encapsulated in the vegetation indices. This implies that we will lose some information if we use only the vegetation indices. The blue (B3) and MIR (B7) channels influenced only subtly on the regression performance especially for the minor types. This contradicts our initial assumption that these bands could be useful to distinguish LULC types due to extra information.

MODIS EVI utilises an extra band B3 compared to NDVI. However, 'EVI' predictor set based scenarios were outperformed by 'NDVI' predictor set based scenarios as if B3 did not supply any incremental information about the vegetation activity or land cover status. It may be due to the way MODIS EVI is parametrised. In principle, the parameters in the EVI formula should be determined on-site. However, fixed parameter values are used for the MODIS EVI product for convenience. EVI could be a better predictor with site-specific calibration.

In agricultural fields, land use can be altered in a short time period by which spectral signals can be abruptly changed. Therefore it appears natural that the 8-day scenarios outperformed the 16-day scenarios. Vegetative LULC types are continuously changing within a single year. Therefore it is difficult to capture its characteristics using satellite images from a small number of overpasses (Hüttich et al., 2009; Thenkabail et al., 2005). Moreover, crops have a relatively short life-cycles as well as frequent human interventions, thus may not be fully characterised by a small number of images (Gumma et al., 2011; Li et al., 2014). We therefore recommend using full time series of satellite data to model multi-type LULC cover.

Additional features may further improve regression performance. For example, phenology metrics such as green-on or green-off dates are used to identify vegetation and land cover types (e.g. Lu et al., 2014; Pittman et al., 2010). However, costs of adding features (i.e. computing time) should be carefully considered.

3.5 Conclusions

Existing global land use/land cover (LULC) raster maps have limited spatial and thematic resolution particularly unfavourable to complex agricultural landscapes. As a contribution to resolving this issue, we developed a fractional cover regression model and a strategy to set up the model with globally available satellite products. When properly chosen and processed, coarse satellite products can yield useful information at the sub-pixel level such as fractional land cover. Among the data processing options, choice of predictor sets was the most important.

In estimating absolute fraction, the model performance differed among LULC types depending on the distributions of the observed fraction data. For the minor types, predicting absolute fractions remained difficult. The monsoon period was not the most important period on the regression performance but the critical periods varied by land cover type.

Estimating fractional land cover is a useful strategy for obtaining continuous representation of LULC. It may also alleviate computational burden related to the use of high-resolution raster images. However, fractional cover estimation especially with multiple land cover types is still underdeveloped. With possible elaborations such as the Hurdle formulation, it may be possible to extract useful land cover information from coarse multi-spectral satellite products.

Our study demonstrated how to build a reliable fractional cover regression model by choosing optimal data-processing options. Our evaluation framework and findings can be a useful guide to make informed decisions in similar studies.

3.6 Acknowledgments

This research was supported by International Research Training Group between Germany and South Korea (DFG/KOSEF, Complex TERRain and ECOlogical Heterogeneity - TERRECO, GRK 1565/1). We are grateful to David Harter, Timothy Thrippleton, Andreas Schweiger, and Wanmo Kang for providing advice and technical help.

MODIS information obtained from https://lpdaac.usgs.gov, maintained by the NASA Land Processes Distributed Active Archive Center (LP DAAC) at the USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota. 2003. Data for the image were provided by NASA.

References

Asner, G. & D. B. Lobell (2000). "A biogeophysical approach for automated SWIR unmixing of soils and vegetation". In: *Remote Sensing of Environment* 74.1, pp. 99–112 (cit. on pp. 11, 67, 72, 157).

Attarchi, S. & R. Gloaguen (2014). "Classifying complex mountainous forests with L-Band SAR and Landsat data integration: A comparison among different machine learning methods in the hyrcanian forest". In: *Remote Sensing* 6.5, pp. 3624–3647 (cit. on pp. 10, 72, 107).

Bevanda, M., N. Horning, B. Reineking, M. Heurich, M. Wegmann & J. Mueller (2014). "Adding structure to land cover - using fractional cover to study animal habitat use". In: *Movement Ecology* 2.1, p. 26 (cit. on pp. 11, 63).

Biggs, T. W., P. S. Thenkabail, M. K. Gumma, C. A. Scott, G. R. Parthasaradhi & H. N. Turral (2006). "Irrigated area mapping in heterogeneous landscapes with MODIS time series, ground truth and census data, Krishna Basin, India". In: *International Journal Of Remote Sensing* 27.19, pp. 4245–4266 (cit. on pp. 8, 56, 64).

Bivand, R. & C. Rundel (2014). *rgeos: Interface to Geometry Engine - Open Source (GEOS)* (cit. on pp. 45, 77, 109).

Bontemps, S., P. Defourny, E. Bogaert, O. Arino, V. Kalogirou & J. Perez (2011). *GLOBCOVER 2009 - Products Description and Validation Report*. Tech. rep. European Space Agency (cit. on pp. 4, 5, 7–9, 38, 63, 64, 105).

Breiman, L (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32 (cit. on pp. 10, 12, 72, 73, 76, 107).

Brenning, A (2005). "Spatial prediction models for landslide hazards: review, comparison and evaluation". In: *Natural Hazards and Earth System Sciences* 5.6, pp. 853–862 (cit. on p. 73).

Brown, J. C., J. H. Kastens, A. C. Coutinho, D. d. C. Victoria & C. R. Bishop (2013a). "Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data". In: *Remote Sensing of Environment* 130, pp. 39–50 (cit. on p. 64).

Clark, M. L., T. M. Aide, H. R. Grau & G. Riner (2010). "A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America". In: *Remote Sensing of Environment* 114.11, pp. 2816–2832 (cit. on pp. 10, 64, 72, 106).

Clark, M. L. & D. A. Roberts (2012). "Species-Level Differences in Hyperspectral Metrics among Tropical Rainforest Trees as Determined by a Tree-Based Classifier". In: *Remote Sensing* 4.12, pp. 1820–1855 (cit. on pp. 74, 102, 117).

Colditz, R. R., M Schmidt, C Conrad, M. C. Hansen & S Dech (2011). "Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions". In: *Remote Sensing of Environment* 115.12, pp. 3264–3275 (cit. on pp. 11, 12, 37, 38, 64, 71, 84, 85).

Defries, R. S., M. C. Hansen & J. Townshend (2000). "Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8 km AVHRR data". In: *International Journal Of Remote Sensing* 21.6-7, pp. 1389–1414 (cit. on pp. 11, 63, 64, 157).

DeFries, R. S., C. B. Field, I. Fung, C. O. Justice, S. Los, P. A. Matson, E. Matthews, H. A. Mooney, C. S. Potter, K. Prentice, et al. (1995). "Mapping the land surface for global atmosphere-biosphere models: Toward continuous distributions of vegetation's functional properties". In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 100.D10, pp. 20867–20882 (cit. on pp. 11, 72, 157).

Dennison, P. & D. Roberts (2003). "Endmember selection for multiple endmember spectral mixture analysis using endmember average RMSE". In: *Remote Sensing of Environment* 87, pp. 123–135 (cit. on pp. 10, 85).

Didan, K. & A. Huete (2006). "MODIS vegetation index product series collection 5 change summary". In: *Terrestrial Biophysics and Remote Sensing (TBRS) laboratory, The University of Arizona June* 29, p. 2006 (cit. on p. 69).

Eklundh, L. & P. Jönsson (2012). *TIMESAT 3.1 Software Manual*. Tech. rep. Lund University and Malmö University (cit. on pp. 70, 71, 77).

Feldman, B. E. (2005). "Relative Importance and Value". In: *SSRN Electronic Journal* (cit. on pp. 75, 80, 104).

Fernandes, R., R. Fraser, R. Latifovic, J. Cihlar, J. Beaubien & Y. Du (2004). "Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region". In: *Remote Sensing of Environment* 89.2, pp. 234–251 (cit. on pp. 11, 37, 63, 68, 72, 75, 85, 157).

Fontana, F., C. Rixen, T. Jonas, G. Aberegg & S. Wunderle (2008). "Alpine grassland phenology as seen in AVHRR, VEGETATION, and MODIS NDVI time series - a comparison with in situ measurements". In: *Sensors* 8.4, pp. 2833–2853 (cit. on p. 70).

Foody, G. M. & M. K. Arora (1996). "Incorporating mixed pixels in the training, allocation and testing stages of supervised classifications". In: *Pattern Recognition Letters* 17.13, pp. 1389–1398 (cit. on pp. 11, 13, 72, 157).

Genuer, R., J.-M. Poggi & C. Tuleau-Malot (2010). "Variable selection using random forests". In: *Pattern Recognition Letters* 31.14, pp. 2225–2236 (cit. on p. 76).

GEOS Development Team (2014). *GEOS - Geometry Engine, Open Source*. Open Source Geospatial Foundation (cit. on pp. 45, 77, 109).

Ghimire, B, J Rogan & J Miller (2010). "Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic". In: *Remote Sensing Letters* 1.1, pp. 45–54 (cit. on pp. 10, 72).

Gibbons, J. D. & S. Chakraborti (2003). *Nonparametric Statistical Inference, Fourth Edition*. Revised and Expanded. Marcel Dekker (cit. on p. 75).

Gislason, P. O., J. A. Benediktsson & J. R. Sveinsson (2006). "Random Forests for land cover classification". In: *Pattern Recognition Letters* 27.4, pp. 294–300 (cit. on pp. 10, 72).

Grömping, U. (2006). "Relative Importance for Linear Regression in R: The Package relaimpo". In: *Journal of Statistical Software* 17.1, pp. 1–27 (cit. on pp. 75, 77).

Guerschman, J. P., M. J. Hill, L. J. Renzullo, D. J. Barrett, A. S. Marks & E. J. Botha (2009). "Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors". In: *Remote Sensing of Environment* 113.5, pp. 928–945 (cit. on pp. 11, 12, 19, 63, 65, 71, 72, 157).

Gumma, M. K., P. S. Thenkabail & A. Nelson (2011). "Mapping Irrigated Areas Using MODIS 250 Meter Time-Series Data: A Study on Krishna River Basin (India)". In: *Water* 3.1, pp. 113–131 (cit. on pp. 8, 56, 64, 86).

Herold, M, P Mayaux, C. E. Woodcock, A Baccini & C Schmullius (2008). "Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets". In: *Remote Sensing of Environment* 112.5, pp. 2538–2556 (cit. on pp. 4, 5, 8, 9, 63).

Hijmans, R. J. (2014). *raster: raster: Geographic data analysis and modeling* (cit. on p. 77).

Hird, J. N. & G. J. McDermid (2009). "Noise reduction of NDVI time series: An empirical comparison of selected techniques". In: *Remote Sensing of Environment* 113.1, pp. 248–258 (cit. on p. 70).

Hothorn, T., K. Hornik & A. Zeileis (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework". In: *Journal of Computational and Graphical Statistics* 15.3, pp. 1–21 (cit. on p. 72).

Huete, A., C. Justice & W. Van Leeuwen (1999). *MODIS Vegetation Index (MOD 13): Algorithm Theoretical Basis Document*. Tech. rep. (cit. on pp. 65, 69, 108).

Hüttich, C., U. Gessner, M. Herold, B. J. Strohbach, M. Schmidt, M. Keil & S. Dech (2009). "On the Suitability of MODIS Time Series Metrics to Map Vegetation Types in Dry Savanna

Ecosystems: A Case Study in the Kalahari of NE Namibia". In: *Remote Sensing* 1.4, pp. 620–643 (cit. on pp. 10, 64, 72, 86, 106).

Immitzer, M., C. Atzberger & T. Koukal (2012). "Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data". In: *Remote Sensing* 4.12, pp. 2661–2693 (cit. on pp. 10, 12, 72, 107).

Johnson, B., R. Tateishi & T. Kobayashi (2012). "Remote Sensing of Fractional Green Vegetation Cover Using Spatially-Interpolated Endmembers". In: *Remote Sensing* 4.12, pp. 2619–2634 (cit. on pp. 11, 85).

Jonsson, P & L Eklundh (2004). "TIMESAT-a program for analyzing time-series of satellite sensor data". In: *Computers & Geosciences* 30.8, pp. 833–845 (cit. on p. 70).

Kang, M., S Park, H Kwon, H. T. Choi, Y. J. Choi & J Kim (2009). "Evapotranspiration from a deciduous forest in a complex terrain and a heterogeneous farmland under monsoon climate". In: *Asia-Pacific Journal of Atmospheric Sciences* 45.2, pp. 175–191 (cit. on pp. 19, 65).

Leutner, B. F., B. Reineking, J. Müller, M. Bachmann, C. Beierkuhnlein, S. Dech & M. Wegmann (2012). "Modelling Forest α-Diversity and Floristic Composition — On the Added Value of LiDAR plus Hyperspectral Remote Sensing". In: *Remote Sensing* 4.12, pp. 2818–2845 (cit. on pp. 74, 117).

Li, L., M. Friedl, Q. Xin, J. Gray, Y. Pan & S. Frolking (2014). "Mapping Crop Cycles in China Using MODIS-EVI Time Series". In: *Remote Sensing* 6.3, pp. 2473–2493 (cit. on p. 86).

Liaw, A. (2012). *randomForest: Breiman and Cutler's random forests for classification and regression*. 4.6-7 (cit. on p. 77).

Liaw, A. & M. Wiener (2002). "Classification and Regression by randomForest". In: *R news* 2.3, pp. 18–22 (cit. on pp. 12, 72, 117).

Lieske, D. J., D. A. Fifield & C. Gjerdrum (2014). "Maps, models, and marine vulnerability: Assessing the community distribution of seabirds at-sea". In: *Biological Conservation* 172.C, pp. 15–28 (cit. on p. 85).

Loveland, T. R., B. C. Reed, J. F. Brown, D. O. Ohlen, Z Zhu, L Yang & J. W. Merchant (2000). "Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data". In: *International Journal Of Remote Sensing* 21.6-7, pp. 1303–1330 (cit. on pp. 4, 8, 37, 38, 43, 64, 105).

Lu, H., M. R. Raupach, T. McVicar & D. J. Barrett (2003). "Decomposition of vegetation cover into woody and herbaceous components using AVHRR NDVI time series". In: *Remote Sensing of Environment* 86.1, pp. 1–18 (cit. on pp. 11, 12, 64, 71, 72, 157).

Lu, L., C. Kuenzer, H. Guo, Q. Li, T. Long & X. Li (2014). "A Novel Land Cover Classification Map Based on a MODIS Time-Series in Xinjiang, China". In: *Remote Sensing* 6.4, pp. 3387–3408 (cit. on p. 86).

Mora, B., N.-E. Tsendbazar, M. Herold & O. Arino (2014). "Global Land Cover Mapping: Current Status and Future Trends". In: *Land Use and Land Cover Mapping in Europe*. Dordrecht: Springer Netherlands, pp. 11–30 (cit. on pp. 2, 4–7, 9, 10, 63, 64).

Mullahy, J. (1986). "Specification and testing of some modified count data models". In: *Journal of Econometrics* 33.3, pp. 341–365 (cit. on p. 85).

NASA Land Processes Distributed Active Archive Center (LP DAAC) (2013a). *MOD13A1 Vegetation Indices 16-Day L3 Global 500m*. Tech. rep. 47914 252nd Street, Sioux Falls, South Dakota (cit. on pp. 9, 68, 69, 160).

Nitze, I, U Schulthess & H Asche (2012). "Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification". In: *Proc of the 4th GEOBIA* (cit. on p. 72).

Nitze, I., B. Barrett & F. Cawkwell (2015). "Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series". In: *International Journal Of Applied Earth Observation And Geoinformation* 34, pp. 136–146 (cit. on pp. 10, 72, 106).

Obata, K., T. Miura & H. Yoshioka (2012). "Analysis of the Scaling Effects in the Area-Averaged Fraction of Vegetation Cover Retrieved Using an NDVI-Isoline-Based Linear Mixture Model". In: *Remote Sensing* 4.7, pp. 2156–2180 (cit. on pp. 11, 12, 67, 71).

O'Leary, R. A., R. W. Francis, K. W. Carter, M. J. Firth, U. R. Kees & N. H. de Klerk (2009). "A comparison of Bayesian classification trees and random forest to identify classifiers for childhood leukaemia". In: *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation*, pp. 4276–4282 (cit. on p. 85).

Pittman, K., M. C. Hansen, I. Becker-Reshef, P. V. Potapov & C. O. Justice (2010). "Estimating Global Cropland Extent with Multi-year MODIS Data". In: *Remote Sensing* 2.7, pp. 1844–1863 (cit. on pp. 8, 9, 11, 37, 38, 56, 63, 64, 86, 105).

Povak, N. A., P. F. Hessburg, K. M. Reynolds, T. J. Sullivan, T. C. McDonnell & R. B. Salter (2013). "Machine learning and hurdle models for improving regional predictions of stream water acid neutralizing capacity". In: *Water Resources Research* 49.6, pp. 3531–3546 (cit. on p. 85).

Prasad, A. M., L. R. Iverson & A. Liaw (2006). "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction". In: *Ecosystems* 9.2, pp. 181–199 (cit. on p. 72).

Price, J. (1992). "Estimating vegetation amount from visible and near infrared reflectances". In: *Remote Sensing of Environment* 41.1, pp. 29–34 (cit. on pp. 11, 63).

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on pp. 77, 109, 117).

Reineking, B., P. Weibel, M. Conedera & H. Bugmann (2010). "Environmental determinants of lightning- v.human-induced forest fire ignitions differ in a temperate mountain region of Switzerland". In: *International Journal of Wildland Fire* 19.5, pp. 541–557 (cit. on p. 73).

Rodriguez-Galiano, V. F., B Ghimire, J Rogan, M Chica-Olmo & J. P. Rigol-Sanchez (2012). "An assessment of the effectiveness of a random forest classifier for land-cover classification". In: *Isprs Journal of Photogrammetry and Remote Sensing* 67, pp. 93–104 (cit. on pp. 10, 72, 74, 106, 117).

Savitzky, A & M. Golay (1964). "Smoothing and differentiation of data by simplified least squares procedures." In: *Analytical chemistry* 36.8, pp. 1627–1639 (cit. on p. 70).

Schulp, C. & R Alkemade (2011). "Consequences of uncertainty in global-scale land cover maps for mapping ecosystem functions: an analysis of pollination efficiency". In: *Remote Sensing* 3.9, pp. 2057–2075 (cit. on pp. 1–3, 37, 38, 63, 161).

Schwarz, M. & N. E. Zimmermann (2005). "A new GLM-based method for mapping tree cover continuous fields using regional MODIS reflectance data". In: *Remote Sensing of Environment* 95.4, pp. 428–443 (cit. on pp. 11, 63, 64, 72, 157).

Schwieder, M., P. Leitão, S. Suess, C. Senf & P. Hostert (2014). "Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques". In: *Remote Sensing* 6.4, pp. 3427–3445 (cit. on pp. 10–12, 63, 71, 72, 83, 106, 157).

Segal, M & Y Xiao (2011). "Multivariate random forests". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, pp. 80–87 (cit. on p. 76).

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression*. Tech. rep. Center for Bioinformatics Molecular Biostatistics (cit. on pp. 12, 72).

Senf, C., D. Pflugmacher, S. van der Linden & P. Hostert (2013). "Mapping Rubber Plantations and Natural Forests in Xishuangbanna (Southwest China) Using Multi-Spectral Phenological Metrics from MODIS Time Series". In: *Remote Sensing* 5.6, pp. 2795–2812 (cit. on pp. 10, 72).

Seo, B., C. Bogner, P. Poppenborg, E. Martin, M. Hoffmeister, M. Jun, T. Koellner, B. Reineking, C. L. Shope & J. Tenhunen (2014). "Deriving a per-field land use and land cover map in an agricultural mosaic catchment". In: *Earth System Science Data* submitted (cit. on pp. 8, 16, 66, 109, 141).

Seo, B., P. Poppenborg, E. Martin, M. Hoffmeister, C. Bogner, H. Elsayed Ali, B. Reineking & J. Tenhunen (2014). *Per-field land use and land cover data set of the Haean catchment, South Korea. doi:10.1594/PANGAEA.823677*. data set (cit. on pp. 15, 19, 21, 67, 96, 109).

Smith, M. O., S. L. Ustin, J. B. Adams & A. R. Gillespie (1990). "Vegetation in deserts: I. A regional measure of abundance from multispectral images". In: *Remote Sensing of Environment* 31.1, pp. 1 –26 (cit. on pp. 11, 63, 67).

Strobl, C., A.-L. Boulesteix, A. Zeileis & T. Hothorn (2007). "Bias in random forest variable importance measures: illustrations, sources and a solution." In: *Bmc Bioinformatics* 8.1, p. 25 (cit. on p. 76).

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin & A. Zeileis (2008). "Conditional variable importance for random forests." In: *Bmc Bioinformatics* 9.1, p. 307 (cit. on p. 74).

Thackway, R., L. Lymburner & J. P. Guerschman (2013). "Dynamic land cover information: bridging the gap between remote sensing and natural resource management". In: *Ecology And Society* 18.1 (cit. on pp. 5, 10, 65).

Thenkabail, P. S., M. Schull & H. Turral (2005). "Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data". In: *Remote Sensing of Environment* 95.3, pp. 317–341 (cit. on pp. 9, 10, 64, 72, 86, 105).

U.S. Geological Survey (2012). *Global Land Cover Characteristics Data Base Version 2.0*. Tech. rep. U.S. Geological Survey (cit. on pp. 4, 8, 38, 53, 64, 105).

Verbeiren, S., H. Eerens, I. Piccard, I. Bauwens & J. Van Orshoven (2008). "Sub-pixel classification of SPOT-VEGETATION time series for the assessment of regional crop areas in Belgium". In: *International Journal Of Applied Earth Observation And Geoinformation* 10.4, pp. 486–497 (cit. on pp. 64, 83, 85).

Vittek, M., A. Brink, F. Donnay, D. Simonetti & B. Desclée (2014). "Land Cover Change Monitoring Using Landsat MSS/TM Satellite Image Data over West Africa between 1975 and 1990". In: *Remote Sensing* 6.1, pp. 658–676 (cit. on pp. 9, 68, 160).

Watts, J. D., S. L. Powell, R. L. Lawrence & T. Hilker (2010). "Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery". In: *Remote Sensing of Environment* 115.1, pp. 66–75 (cit. on pp. 9, 68, 160).

Xiao, J. & A. Moody (2005). “A comparison of methods for estimating fractional green vegetation cover within a desert-to-upland transition zone in central New Mexico, USA”. In: *Remote Sensing of Environment* 98.2-3, pp. 237–250 (cit. on pp. 10, 11, 84, 85).

Yihui, D. & J. C. L. Chan (2005). “The East Asian summer monsoon: an overview”. In: *Meteorology and Atmospheric Physics* 89.1-4, pp. 117–142 (cit. on pp. 19, 65, 66).

Supplementary Figures

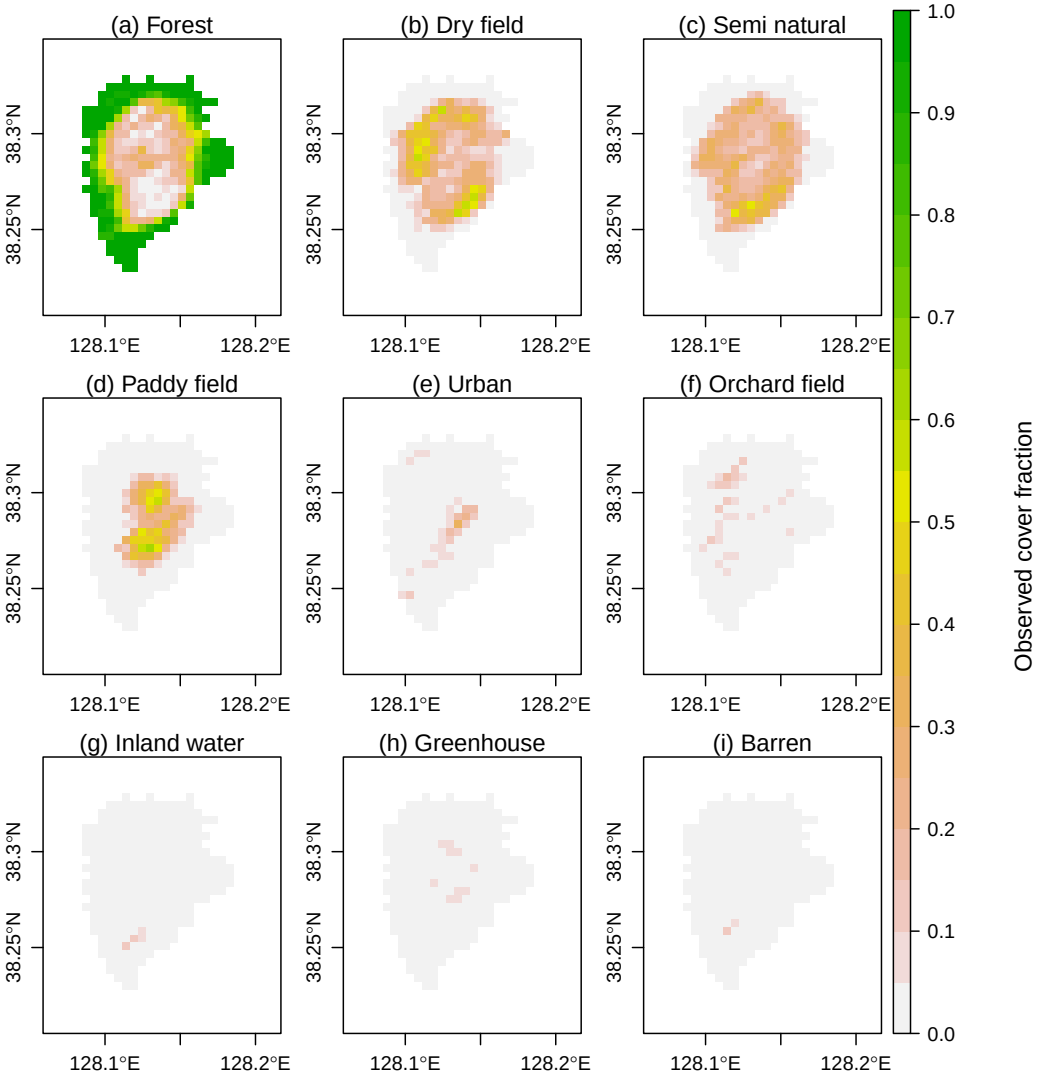

Fig. 3.9 The reference land use/land cover (LULC) fractions of the study site in 2010. LULC fractions were calculated from the original polygon data (Seo et al., 2014) to fit the MODIS 500 m sinusoidal grid (EPSG: 6842) and range from 0 (0% cover) to 1 (100% cover).

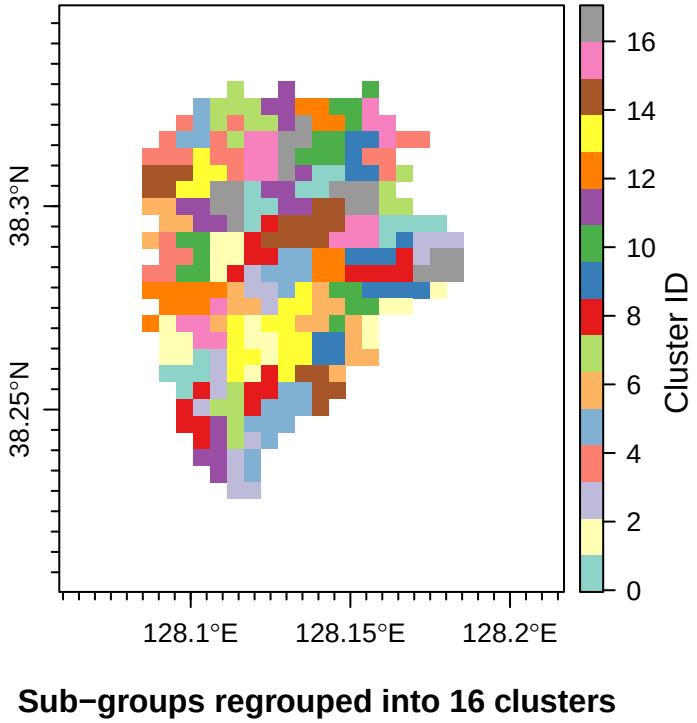

Fig. 3.10 Location of the 16 clusters and the 64 sub-clusters used for spatial cross-validation. Adjacent pixels in the same colour indicate a sub-cluster and four of the sub-clusters comprise a cluster. In each cross-validation fold, one cluster was hold-out as test data and the rest 15 clusters trained a Random Forest regression model. The mean size of the clusters was 4.00 km^2 and the sub-clusters was 1.00 km^2.

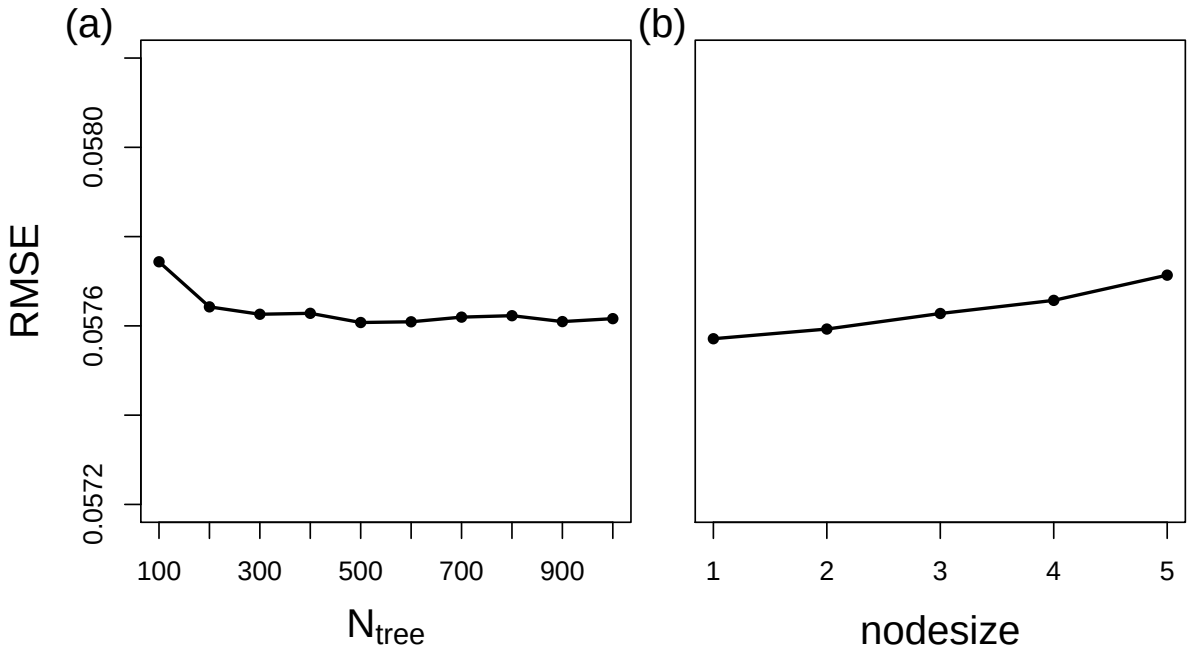

Fig. 3.11 Variations of *RMSE* with changing Random Forest parameters (a) N_{tree} and (b) *nodesize* during the parameter tuning based on the repartitioning of the training data. For illustrating the general response of the model, the mean *RMSE* of all scenarios and the LULC types are displayed. Note that the optimal n_{tree} and *RMSE* were determined individually per scenario.

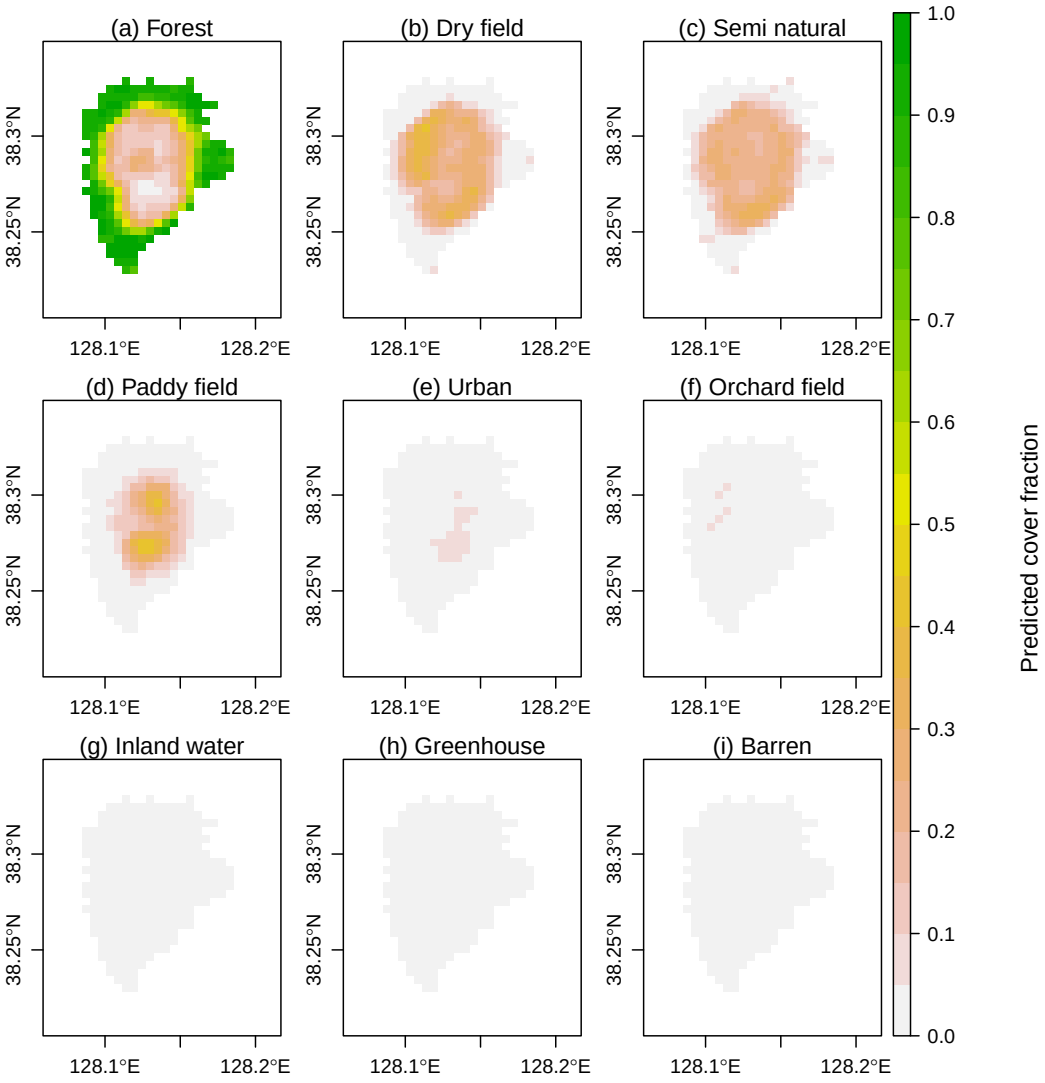

Fig. 3.12 Mean predicted LULC fractions of the study area. Maps from the averaged fractions over the all 16 scenarios.

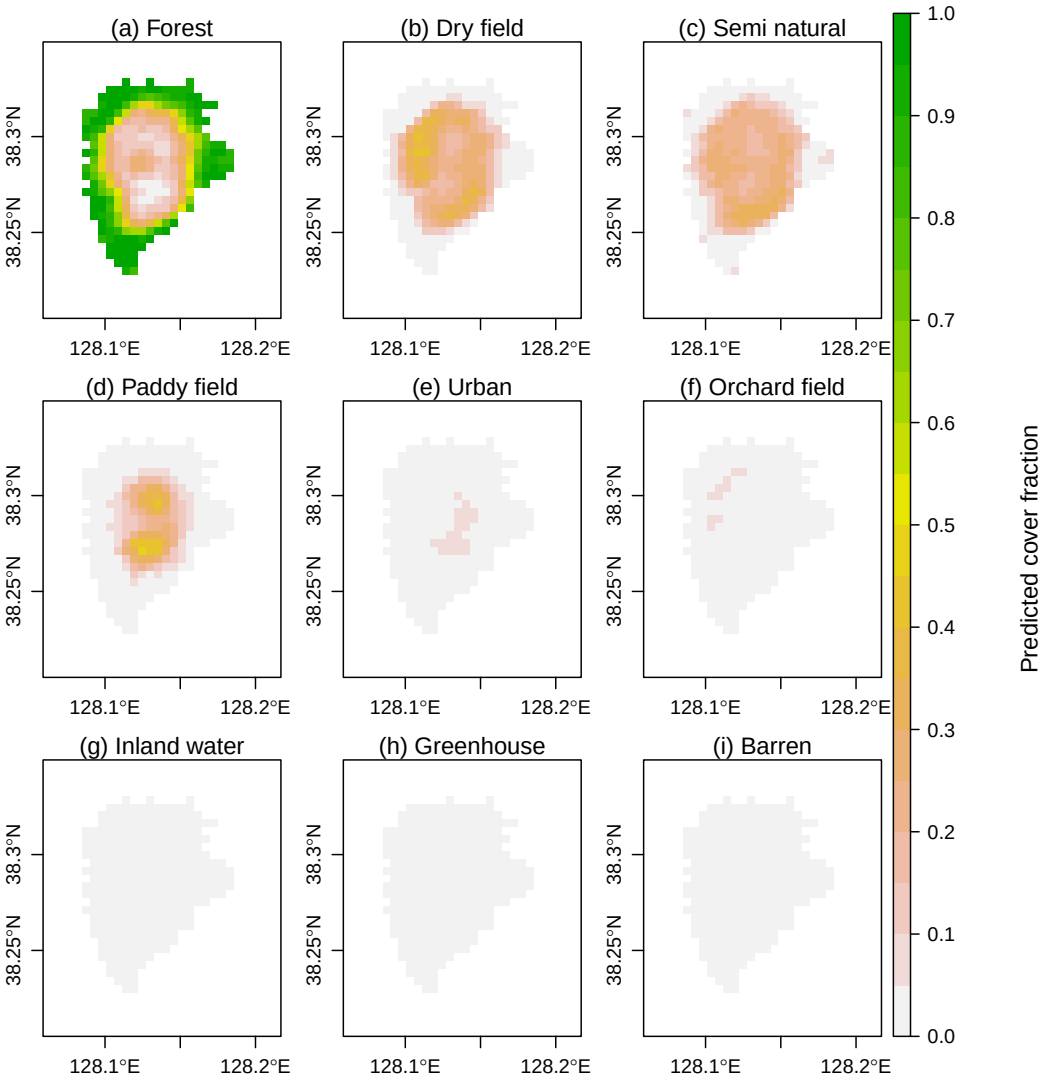

Fig. 3.13 Predicted LULC fractions from the best performed scenario (S4). This scenario used the non-smoothed full features in 8-day interval as predictor.

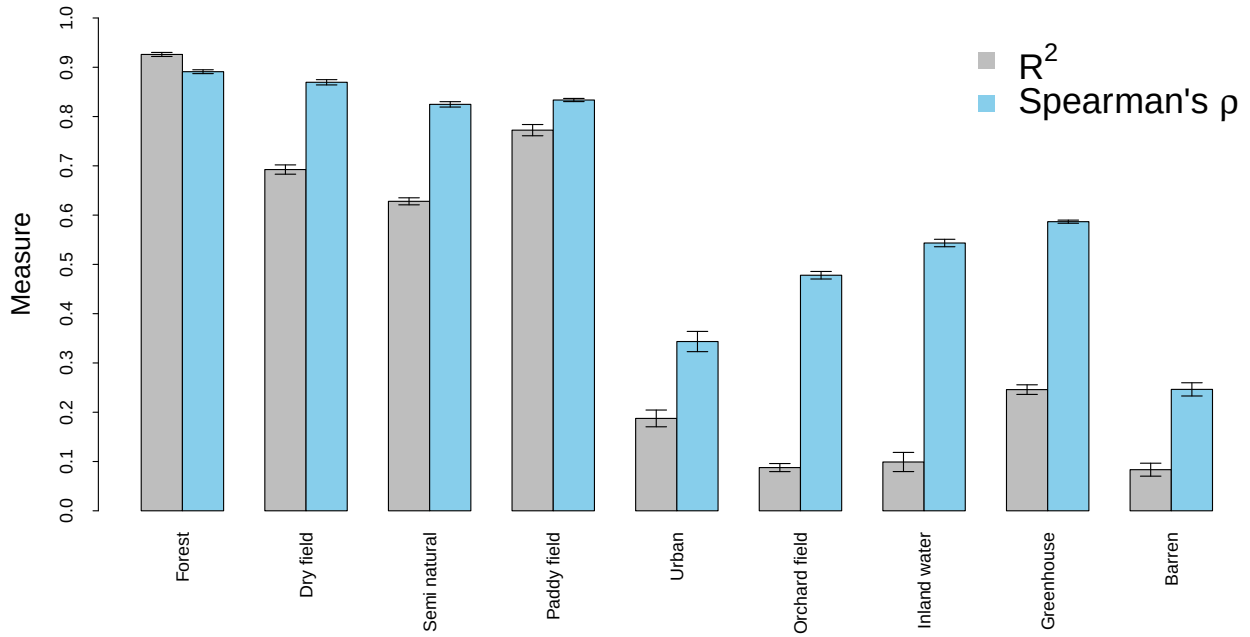

Fig. 3.14 R^2 and Spearman's rank correlation coefficients between observed and predicted fractions. Error bars indicate the standard error of the mean over the scenarios.

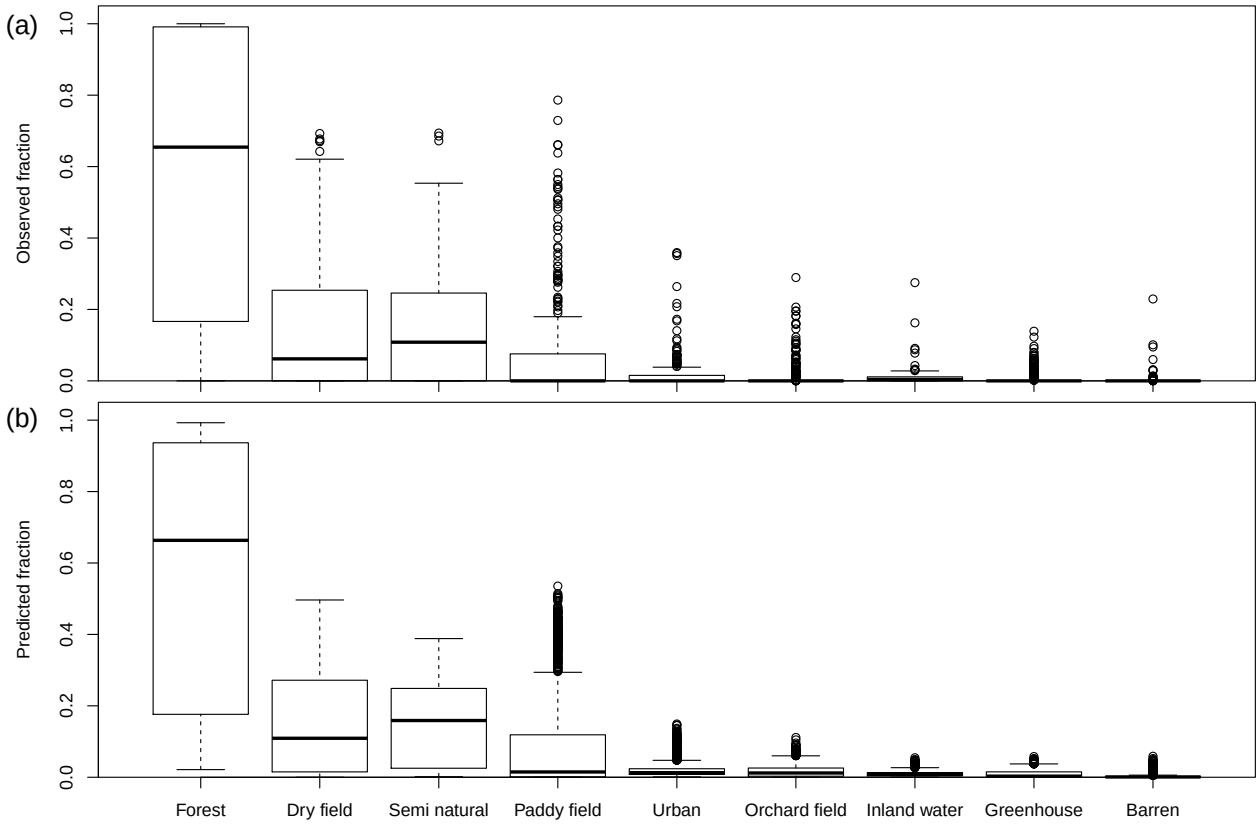

Fig. 3.15 Distributions of cover fractions of (a) the ground LULC observations and (b) the averaged predictions from scenarios S1 through S16.

Supplementary Tables

Table 3.4. Specification of the scenarios and the Random Forest training parameters. The parameters n_{tree} and *nodesize* were tuned and m_{try} was determined by the square root of $n_{feature}$ (Clark et al., 2012; Khalilia et al., 2011).

Name	data-processing options					Parameters		
	Predictor set	Time interval	Smoothing	n_{band}	$n_{feature}$	n_{tree}	m_{try}	*nodesize*
S1	NDVI	8-day	No smoothing	1	46	600	6	1
S2	EVI			1	46	700	6	2
S3	SR			4	184	400	13	3
S4	Full			6	276	700	16	1
S5	NDVI	16-day		1	23	200	4	1
S6	EVI			1	23	500	4	1
S7	SR			4	92	300	9	1
S8	Full			6	138	800	11	2
S9	NDVI	8-day	Savitzky-Golay smoothing	1	46	600	6	4
S10	EVI			1	46	500	6	1
S11	SR			4	184	500	13	1
S12	Full			6	276	400	16	1
S13	NDVI	16-day		1	23	800	4	1
S14	EVI			1	23	300	4	1
S15	SR			4	92	900	9	1
S16	Full			6	138	600	11	1

Table 3.5. Type-wise performance measures between observed and predicted fractions averaged over all scenarios.

Classes	$RMSE$	ρ	R^2
Forest	0.11	0.89	0.93
Dry field	0.10	0.87	0.69
Semi natural	0.09	0.82	0.63
Paddy field	0.08	0.83	0.77
Urban	0.04	0.34	0.19
Orchard field	0.04	0.48	0.09
Inland water	0.02	0.54	0.10
Greenhouse	0.02	0.59	0.25
Barren	0.02	0.25	0.08
Avg.	0.06	0.62	0.41

Table 3.6. Normalised increased mean square error ($NIMSE_b$) of the four spectral bands extracted from the 'SR' predictor set based scenarios (S3, S7, S11, and S15).

	$NIMSE_b$ (%)			
Classes	B1	B2	B3	B7
Forest	38.5	38.3	10.9	12.3
Dry field	45.8	47.9	2.6	3.7
Semi natural	49.0	46.6	2.1	2.2
Paddy field	49.7	44.7	3.5	2.0
Urban	50.8	48.4	0.4	0.4
Orchard field	48.1	51.4	0.2	0.2
Inland water	50.1	49.8	0.1	0.1
Greenhouse	53.0	46.8	0.1	0.1
Barren	52.0	47.9	0.1	0.0
Avg.	48.6	46.9	2.2	2.3

Table 3.7. $NIMSE_b$ of the six bands extracted from the 'Full' predictor set based scenarios (S4, S8, S12, and S16).

	$NIMSE_b$ (%)					
Classes	NDVI	EVI	B1	B2	B3	B7
Forest	30.0	25.1	26.7	7.2	5.2	5.8
Dry field	30.8	30.9	33.6	1.5	1.3	1.9
Semi natural	31.4	33.1	32.1	1.2	1.0	1.1
Paddy field	34.1	32.0	29.7	1.8	1.5	0.8
Urban	34.4	33.5	31.6	0.2	0.2	0.2
Orchard field	31.0	32.5	36.2	0.1	0.1	0.1
Inland water	32.2	34.3	33.4	0.1	0.0	0.0
Greenhouse	35.3	34.2	30.5	0.0	0.0	0.0
Barren	32.7	34.7	32.5	0.0	0.0	0.0
Avg.	32.4	32.3	31.8	1.3	1.0	1.1

Table 3.8. Summary of the linear models explaining the model's $RMSE$ by the three data-processing options: $RMSE \sim O_p+O_t+O_s$, where O_p is a categorical variable denoting the chosen predictor set option, O_t time interval option, and O_s smoothing option. Statistical significance was tested by F-statistics and the relative contribution (i.e. proportion of variance explained) of the options were calculated via proportional marginal variance decomposition (PMVD) method (Feldman, 2005).

Type	Pr(>F)	Explained variance (%)		
		O_p	O_t	O_s
Forest	0.00	52.92	19.61	7.75
Dry field	0.00	71.78	0.80	11.65
Semi natural	0.00	29.45	50.90	3.47
Paddy field	0.00	58.65	31.62	1.86
Urban	0.02	41.32	4.14	24.30
Orchard field	0.00	10.10	2.11	65.74
Inland water	0.00	19.69	35.90	29.31
Greenhouse	0.02	32.75	19.61	16.06
Barren	0.83	9.70	5.90	1.36
Avg.	-	36.26	18.96	17.94

Chapter 4

Improving the classification of rare land use and land cover types using synthetic data

4.1 Introduction

Detailed information on land use and land cover (LULC) is essential in many areas of environmental sciences. A constantly growing body of literature emphasizes the impact that changes in land use may have on Earth's climate (e.g. Chhabra et al., 2006; Foley et al., 2005; Turner et al., 2007), biodiversity (Dawson et al., 2011; Hoffmann et al., 2010) and water cycle (Sterling et al., 2012). Among different human land use forms, cultivated ecosystems (for production of food, feed and fibre) are particularly frequent and occupy 34% of the land surface (Chhabra et al., 2006).

However, detecting LULC changes in cultivated and especially in agricultural areas might be challenging. Agricultural landscapes are frequently made up of a spatial mosaic of different crop types. In contrast, the most frequently used global land cover databases like GlobCover or MODIS Land Cover Type contain only few crop-related classes (Bontemps et al., 2011; Loveland et al., 2000; U.S. Geological Survey, 2012). To derive more detailed LULC maps, data from the Moderate Resolution Imaging Spectroradiometer (MODIS) can be used (e.g. Pittman et al., 2010; Thenkabail et al., 2005).

One advantage of MODIS is global coverage. Additionally, due to daily revisiting times the final products can be temporally aggregated to avoid data gaps. Several different MODIS products exist with varying degrees of preprocessing. We focus here on the Vegetation Indices product

MOD13Q1 that has a moderate spatial resolution of 250 m and provides time series of surface reflectance with a temporal resolution of 16 days.

MODIS time series are particularly suitable to track seasonal variation of vegetation development. Hüttich et al. (2009), for example, used MODIS time series metrics to map vegetation types in dry African Savannah. Brown et al. (2013b) analysed land use changes in an agricultural area in Brazil. Based on a detailed ground reference data set, the authors could distinguish 15 land-use classes related to agriculture. However, some of these classes had to be eliminated from the analysis because they were rare. Others had to be grouped due to their spectral similarity.

Rare or minor classes are often difficult to classify. It is commonly recognized that classifiers perform best on (approximately) equally distributed classes (e.g. Chawla et al., 2004; Fernández et al., 2011). However, because minor classes are ubiquitous in remote sensing, the data sets are often imbalanced. In general, there are three major ways to cope with imbalanced data sets. The first is to adapt the classification algorithm to reinforce learning of the minor classes (e.g. Bruzzone et al., 1997; Williams et al., 2009). The second is to adjust the classifier by assigning different costs to misclassification in rare versus frequent classes (e.g. Alejo et al., 2009; Sun et al., 2007). The third is by re-sampling the data set (e.g. García et al., 2011; He et al., 2009; Waske et al., 2009, and references therein). This last approach has the advantage to be independent from the classifier used.

Oversampling of the rare classes with replacement or undersampling of the major class have been discussed by several authors (Japkowicz et al., 2002; Ling et al., 1998; Schistad Solberg et al., 1996). However, the potential of these approaches to improve the classification accuracy of rare classes seems to be limited. In particular random oversampling with replacement can lead to overfitting (Chawla, 2010).

To overcome the issue of overfitting, Chawla et al. (2002) proposed to generate new minority instances by a synthetic minority oversampling technique (SMOTE) instead of oversampling with replacement. They reported that the synthetic points created by SMOTE forced the classifier to learn larger and less specific regions and thus changed the boundaries between classes. SMOTE performs better than oversampling the minority class by replacement and can be combined with undersampling of the majority class.

In remote sensing, machine learning algorithms have been widely adopted for land cover and land use analysis. For example, Random Forests (RF) have been used in modelling land cover in a variety of landscapes (Clark et al., 2010; Nitze et al., 2015; Schwieder et al., 2014) including a heterogeneous agricultural region (Rodriguez-Galiano et al., 2012). A Random Forest classifier can deal with a large number of highly correlated features (e.g. spectral data) and non-linear

relationships (Breiman, 2001; Immitzer et al., 2012). Similarly, Support Vector Machines (SVM) have also gained increasing attention (Attarchi et al., 2014; Mountrakis et al., 2011; Vuolo et al., 2012). They are suitable for relatively small data sets and a large number of features (Camps-Valls et al., 2009). Therefore, it appears natural to combine these state-of-the are machine learning methods with SMOTE for classification tasks on imbalanced data sets (Akbani et al., 2004; Johnson et al., 2013). Comparing alternative machine learning algorithms helps in identifying general benefits of SMOTE that are not specific to one particular machine learning algorithm.

In our work, we apply SMOTE to a complex real-world data set characterized by a large imbalance ratio. It contains 17 classes (two major and 15 minor classes) – a multi-majority and multi-minority data set (Wang et al., 2012) – and a small number of points in minority classes. The goal of our study is to improve the classification of rare classes in an agricultural mosaic catchment by using SMOTE on the standard MODIS product MOD13Q1. We compare four different classification scenarios and two machine learning algorithms (RF and SVM) on original imbalanced and synthetically oversampled data. We quantify the effect of SMOTE on overall model performance and on different groups of land cover classes. In particular, we show that in the presence of class overlap increasing the number of training points does not guarantee a better classification result. To our knowledge, only few papers in the remote sensing literature address this issue on a complex real-world data set. Finally, we analyse by which mechanism the alternative scenarios affect model performance, looking at the relationship between class labels and surface reflectance (mutual information) and the difficulty of classification (entropy).

4.2 Data and study area

4.2.1 Study area

The studied catchment Haean (128°1′33.101″E, 38°28′6.231″N) is located in the mountainous watershed Soyang in the northeastern part of South Korea (Figure 4.1). This watershed is a protected temperate forest. However, some parts, including our study area, are used intensively for conventional agriculture. The Haean catchment has a total area of 64 km^2 with elevations ranging from 500 m to 1200 m. The agricultural zone is located in the center of the catchment and has a mosaic structure with various LULC types such as annual and perennial crops, seminatural and urban area. The catchment is surrounded by a dense deciduous forest (Figure 4.2a).

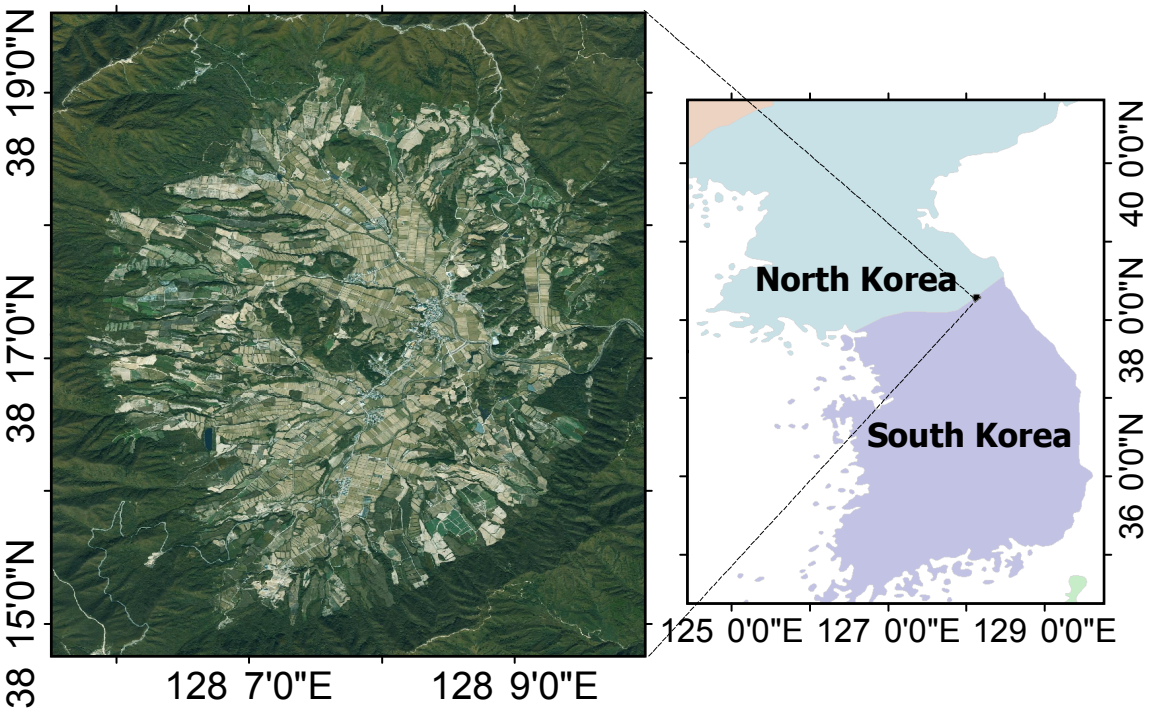

Fig. 4.1 Map of the Haean catchment located at the border between North and South Korea. The satellite image is a SPOTMaps mosaic product (Astrium Services, http://www.astrium-geo.com) acquired in 2009.

4.2.2 MODIS surface reflectance

We used MODIS 16-day reflectance layers from the Collection 5 MOD13Q1 product (NASA Land Processes Distributed Active Archive Center (LP DAAC), 2013b) for the year 2010. It supplies four reflectance bands (bands 1, 2, 3 and 7) at 250 m scale. Every observation in MOD13Q1 is a temporal composite of 16 daily measurements that were filtered to remove cloud contamination (Huete et al., 1999; Solano et al., 2010).

The four reflectance bands have specific information related to their spectral ranges. The red channel band 1 (B1) covers 620–670 nm and is sensitive to chlorophyll in vegetation. The near-infrared channel band 2 (B2) covers 841–876 nm and has been widely used to evaluate ground vegetation viability together with B1. Band 3 (B3) is commonly called the blue channel due to its sensitivity to water vapour. It covers 459–479 nm and is used to filter the cloud covered data or detect water bodies. Additionally, it serves to differentiate soil from vegetation. The range of the mid-infrared band 7 (B7) equals 2105–2155 nm and is also used to examine land and cloud properties.

Every band contained 23 images per year and we obtained 96 data points per pixel of 250 m $\times$ 250 m. The whole catchment covered 1198 pixels of the MODIS tile H28V5. No pretreatment was applied to the retrieved MODIS data.

4.2.3 Reference land use and land cover data

The reference LULC data set was obtained by ground census of the whole study area in 2010. It contains 67 crop/non-crop LULC types and is available online at the public repository Pangaea (Seo et al., 2014; Seo et al., 2014). Originally, the data set consisted of projected geospatial polygons (WGS84 / UTM 52N; EPSG:32652) with LULC classes assigned to each polygon. To increase the sample sizes for locally important LULC types we merged similar ones and obtained 59 classes (Supplementary Table 4.2). Subsequently, for this study we converted the polygons to a raster image on the MODIS 250 m sinusoidal grid (SR-ORG:6842).

We determined the LULC class of a grid cell covered by multiple spatial polygons based on the exact area size: We calculated the fraction of the occupied area in the projected space and assigned the LULC class based on the highest proportion. The rasterisation yielded a data set containing 28 LULC classes (Table 4.1). Figure 4.2 shows the original data containing 59 classes and the rasterized data set with 28 LULC classes. The rasterisation was done in `R` (R Core Team, 2014) using the geometry engine `GEOS` (GEOS Development Team, 2014) and the package `rgeos` (Bivand et al., 2014).

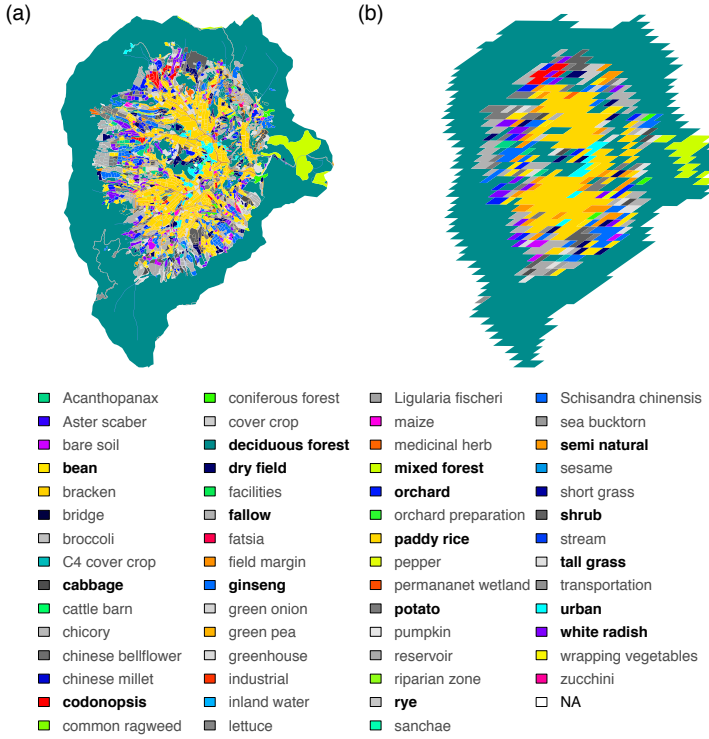

Fig. 4.2 Land use and land cover of the Haean catchment surveyed in 2010. (a) Original polygon data (59 classes) and (b) rasterized sinusoidal grid (28 classes). The names in bold indicate the 17 classes used for classification.

Table 4.1. Distribution of the 28 land use and land cover classes in the rasterized data set. The first 17 classes were used for classification.

LULC	Pixels		Area	
	–	(%)	(km^2)	(%)
deciduous forest	719	60.02	35.67	55.46
paddy rice	148	12.35	5.18	8.05
fallow	62	5.18	3.08	4.78
ginseng	34	2.84	1.59	2.48
semi natural	32	2.67	3.79	5.89
potato	32	2.67	1.57	2.45
mixed forest	22	1.84	1.34	2.08
bean	20	1.67	1.47	2.28
white radish	18	1.50	1.16	1.81
dry field	14	1.17	1.22	1.90
tall grass	13	1.09	1.45	2.26
orchard	13	1.09	0.94	1.46
shrub	13	1.09	0.66	1.03
rye	10	0.83	0.49	0.77
urban	9	0.75	0.56	0.86
codonopsis	7	0.58	0.29	0.46
cabbage	6	0.50	0.69	1.07
greenhouse	4	0.33	0.54	0.85
Acanthopanax	4	0.33	0.19	0.29
reservoir	4	0.33	0.15	0.23
bare soil	4	0.33	0.14	0.22
coniferous forest	3	0.25	0.19	0.29
transportation	2	0.17	0.50	0.77
pepper	1	0.08	0.16	0.24
medicinal herb	1	0.08	0.08	0.12
Ligularia fischeri	1	0.08	0.05	0.08
C4 cover crop	1	0.08	0.05	0.07
Schisandra chinensis	1	0.08	0.01	0.02

4.3 Methods and data analysis

4.3.1 Difficulty of classification

As pointed out by Kononenko et al. (1991) the distribution of the classes is closely related to the difficulty of the classification task. Consider a classification task with M classes C_i, $i = 1, \ldots, M$ with prior probabilities $p(C_i)$. We require $\sum_i p(C_i) = 1$. The amount of information to correctly classify one instance with prior probability $p(C_i)$ into class C_i equals $-\log p(C_i)$ (Shannon, 1948). Correspondingly, the amount of information we need to correctly state that an instance does not belong to class C_i equals $-\log(1 - p(C_i))$. Then, the expected amount of information to classify one instance equals to the entropy

$$H = -\sum_{i}^{M} p(C_i) \log p(C_i) \,. \tag{4.1}$$

Intuitively, a classification task with equal prior probabilities $p(C_i) = 1/M$ is the most difficult one. Additionally, a problem with more classes is in general more difficult to solve. This accords well with the properties of the entropy. Indeed, for a classification task with equal prior probabilities the entropy is greater than for one with unequal probabilities (Kononenko et al., 1991; Shannon, 1948). Moreover, if we split one class, H increases.

A classifier trained on an unbalanced data set (i.e. an easy task with small entropy) has more potential to specialize on the majority class and to neglect minority classes. Actually, by classifying every instance into the majority class it can attain a high accuracy (Valverde-Albacete et al., 2014) (see Section 4.3.4 for the definition of accuracy). Therefore, altering the distribution of data changes the difficulty of the classification task.

4.3.2 Data resampling and preprocessing

4.3.2.1 Generating synthetic data points

The synthetic minority oversampling technique (SMOTE) oversamples a rare class by generating new instances (Chawla et al., 2002). For every existing point P_i in a given rare class, it inserts synthetic points along a line that connects this point to one of its k nearest neighbours. Depending on the oversampling rate N, several k nearest neighbours can be chosen randomly and several points can be generated along one connecting line. To create a new point, SMOTE calculates the difference between the chosen nearest neighbour and the point P_i, weights this difference by a random number between 0 and 1 and adds this difference to the point P_i. Figure 4.3 illustrates this principle in two dimensions.

4.3.2.2 Choice of rare classes

In order to generate synthetic data, the rare class must contain at least some original points. The distribution of the LULC types in the Haean catchment is highly imbalanced (Table 4.1). The imbalance ratio is defined as the number of pixels in the most frequent class (in our case 'deciduous forest') divided by the number of pixels in the rare class. The majority class 'deciduous forest' contains more than four times as many pixels as the second most abundant class 'paddy rice'. The imbalance ratio between the forest and the rare classes is even larger.

In the following, we will call the classes 'deciduous forest' and 'paddy rice' majority classes and

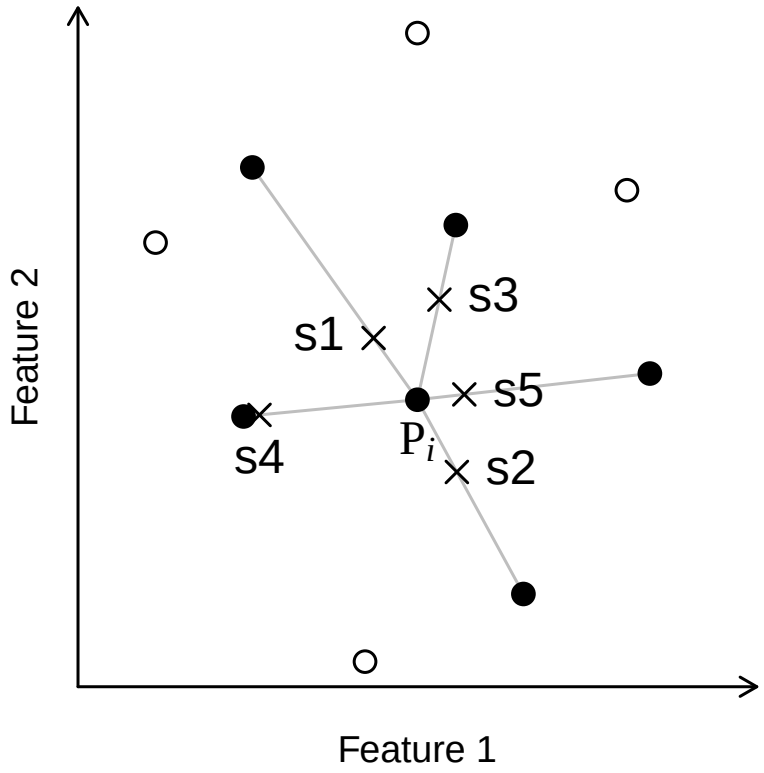

Fig. 4.3 Illustration of the synthetic minority oversampling technique (SMOTE) in two dimensions. SMOTE generates synthetic points (crosses denoted $s1$ through $s5$) along the connection lines between a point P_i (black dot denoted P_i) and its k nearest neighbours (black dots). In this case, the number of nearest neighbours $k = 5$ and the oversampling rate $N = 5$. Circles show other minority samples that are not the k nearest neighbours of P_i.

denote the others as minority classes. For SMOTE we selected only minority classes containing at least six pixels resulting in a data set with 17 LULC classes (Tables 4.1 and 4.2). They make up 97.8% of the total number of pixels and 96.1% of the total area of the catchment.

4.3.2.3 Removing Tomek links

Additionally to oversampling the minority classes, we inspected the neighbourhood relationships in the original data set. We calculated the Euclidean distance between the reflectance data of all pixels in the chosen 17 LULC classes and identified the closest neighbour of every pixel (i.e. its first nearest neighbour). The most frequent nearest neighbours of nine of 17 LULC classes belonged to the same class (Figure 4.4). However, in eight minority classes the most frequent nearest neighbour belonged to a different class, namely 'deciduous forest', 'paddy rice' or 'fallow'.

Direct neighbours belonging to different classes are called Tomek links (Tomek, 1976). More formally, a pair of points P_i and P_j belonging to different classes forms a Tomek link, if there is no third point P_ℓ such that $d(P_i, P_\ell) < d(P_i, P_j)$ or $d(P_j, P_\ell) < d(P_i, P_j)$, where $d(\cdot, \cdot)$ is the

distance between points. Points forming a Tomek link are either borderline instances or one of them is noisy. We removed Tomek links by undersampling the majority class 'deciduous forest' in order to reduce the class overlap with the agricultural classes (Batista et al., 2004). The removal of Tomek links was done using the `R` package `unbalanced` (Pozzolo et al., 2014).

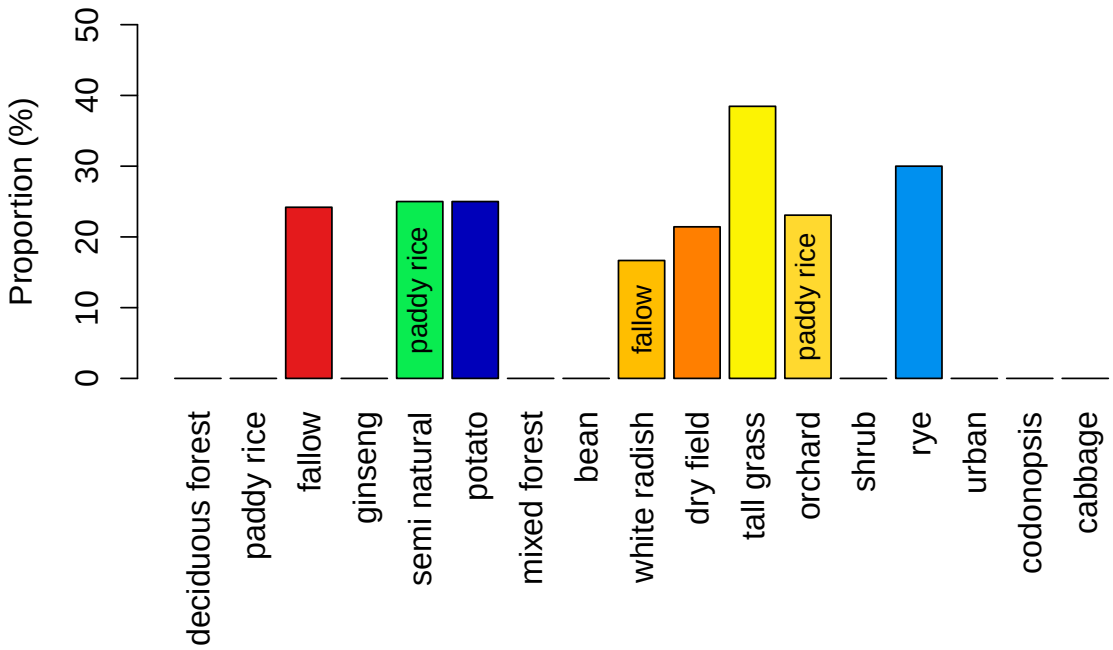

Fig. 4.4 Proportion of the most frequent nearest neighbours belonging to a different class (Tomek links) in the total number of nearest neighbours. The most frequent nearest neighbours in the classes with zero proportion belonged to the same class. All Tomek links were with 'deciduous forest', except in the classes 'semi natural' (with 'paddy rice'), 'white radish' (with 'fallow') and 'orchard' (with 'paddy rice').

4.3.3 Mutual information: relationship between class labels and surface reflectance

Changing the distribution of classes by synthetic oversampling of minority classes and undersampling of the majority classes probably affects the relationship between the class labels and the spectral bands. Spectral bands that are stronger related to minor classes might become more influential. To quantify this change we calculated the mutual information MI between the class labels and the surface reflectance values to see possible changes in their relationships. The mutual information is a general measure of dependency between random variables

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)\,, \tag{4.2}$$

where $H(X)$ and $H(Y)$ are the Shannon entropies of the random variables X and Y, respectively,

and $H(X, Y)$ is the joint Shannon entropy of X and Y (Shannon, 1948).

We estimated the mutual information by the method of k (in our case $k = 3$) nearest neighbours (Kraskov et al., 2004). It avoids the binning (discretization) of the real-valued data that is known to produce biased estimates (Kraskov et al., 2004). The calculations were done using the `R` package `parmigene` (Sales et al., 2012).

The mutual information is non-negative, but has no upper bound. In order to facilitate comparisons between different data sets, we normalized it as proposed by Joe, 1989 and modified by Numata et al., 2008:

$$MI^*(X, Y) = \text{sign}\left(\widehat{MI}(X, Y)\right)\left(1 - \mathrm{e}^{-2|\widehat{MI}(X,Y)|}\right)^{\frac{1}{2}}, \tag{4.3}$$

where $\widehat{MI}(X, Y)$ is the estimate of the mutual information and the function sign$(\cdot)$ evaluates the sign of its argument. This normalization takes possible estimation inaccuracies of $MI(X, Y)$ into account by using the function sign$(\cdot)$ (Numata et al., 2008). Although $MI(X, Y) \geq 0$, the estimate $\widehat{MI}(X, Y)$ might be negative indicating estimation errors and the sign function allows for negative values of $MI^*(X, Y)$ if $\widehat{MI}(X, Y) < 0$. If $\widehat{MI}(X, Y) \geq 0$, then $MI^*(X, Y) \in [0, 1]$. $MI^*(X, Y)$ measures the overall dependency between X and Y and shows how well Y can be predicted by X. It is 0 iff X contains no information about Y (i.e. X and Y are statistically independent), approaches 1 for increasing $\widehat{MI}(X, Y)$ and equals 1 if there is a perfect functional relationship between X and Y. If X and Y are normally distributed, $MI^*(X, Y)$ becomes the absolute value of the linear (i.e. Pearson's) correlation coefficient (Joe, 1989).

4.3.4 Performance measures

Usually the performance of a classifier is assessed via the confusion matrix (Figure 4.5 shows an example for a binary classification). The predicted classes appear in the columns and the actual ones in the rows. The single cells are the number of correctly classified positive (TP: True Positive) and negative examples (TN: True Negative) and misclassified positive (FP: False Positive) and negative examples (FN: False Negative).

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Fig. 4.5 Confusion matrix to evaluate the performance of a binary classifier. TP: true positive, FP: false positive, FN: false negative and TN: true negative.

A classical measure of performance is the (predictive) accuracy $A = (TP + TN)/(TP + FP +$

$TN + FN$). However, for imbalanced data sets A is known to be inappropriate because it masks a poor performance on minority class (Weiss, 2004; Weiss et al., 2003). Therefore, the Receiver Operating Characteristic (ROC) graph was proposed to compare different classification results. In an ROC graph the false positive rate $FPR = FP/(FP + TN)$ is plotted on the x-axis and the true positive rate (also called recall R) $TPR = TP/(TP + FN)$ on the y-axis (e.g. Fawcett, 2006). The closer a classifier approaches the point $(0, 1)$ in this graph, the better its performance.

To summarize the confusion matrix in our multi-class classification task we calculated *G-mean*, *F-score* and the normalized information distance (*NID*) (Kraskov et al., 2005). *G-mean* is the geometric mean of recall per class and is often used in classification with imbalanced data. Its multi-class version is defined as

$$G\text{-}mean = \left(\prod_{i=1}^{M} R_i\right)^{\frac{1}{M}}, \tag{4.4}$$

where $R_i = TP_i/(TP_i + FN_i)$ is the recall in class i and M the number of classes (Kubat et al., 1997; Sun et al., 2006). *F-score* was originally defined for binary classification and we use the macro-averaged extension to multi-class problems because it treats all classes equally (Sokolova et al., 2009)

$$F\text{-}score = \frac{2P_M R_M}{P_M + R_M}, \tag{4.5}$$

where P_M is the macro-averaged precision

$$P_M = \frac{1}{M}\sum_{i=1}^{M}\frac{TP_i}{TP_i + FP_i}, \tag{4.6}$$

and R_M the macro-averaged recall

$$R_M = \frac{1}{M}\sum_{i=1}^{M} R_i \tag{4.7}$$

and M is the number of classes. Precision reports how many of instances recognized as positive are indeed positive and recall shows how many of actually positive instances were found.

NID belongs to the group of information theoretic measures that have a strong mathematical foundation (Vinh et al., 2010). It is a metric and a normalized measure in the range $[0, 1]$ and is used to assess the distance (i.e. dissimilarity) between partitions X and Y

$$NID(X, Y) = 1 - \frac{MI(X, Y)}{\max(H(X), H(Y)}, \tag{4.8}$$

where $MI(X, Y)$ is the mutual information, $H(X)$ and $H(Y)$ are the Shannon entropies of X and Y, respectively. $NID(X, Y)$ is zero iff $X = Y$ (i.e. the partitions are identical) and 1 iff X and Y are independent (i.e. the partitions are maximally dissimilar).

4.3.5 Classification scenarios

In order to evaluate how the classification performance of the classifiers changes when the data are preprocessed, we compared different classification scenarios. In every scenario, we carried out a 6-fold stratified cross validation (SCV). Thus we split the data randomly (each class separately) in 6 folds, used 5 folds to train the classifier and the hold-out fold to test it. Because each fold was used once as a hold-out test fold, we obtained predictions for every pixel. The random splitting was repeated 5 times.
We used the same splitting in training and test data for all scenarios. These were defined as follows:

S1: SCV of original data This is the base-line scenario. Any data resampling procedure and subsequent classification has to perform at least as well as in this scenario.

S2: SCV of original data with Tomek links removed We removed Tomek links in the majority class 'deciduous forest' in the training data.

S3: SCV with SMOTEd training data The goal of this scenario is to obtain an equal distribution of minority classes for training. Therefore, after removing Tomek links as in S2, the minority classes in the training folds were synthetically oversampled approximately up to the number of 'paddy rice' pixels (i.e. 123 pixels). The majority classes 'deciduous forest' and 'paddy rice' were not oversampled.

S4: SCV with SMOTEd training data and undersampling Additionally to the removal of Tomek links and SMOTE as in S3, the majority class 'deciduous forest' was undersampled by randomly selecting 123 pixels for training (corresponding to the number of 'paddy rice' pixels in the training data) to obtain an equal distribution of all classes in the training folds. The majority class 'paddy rice' was not oversampled.

4.3.6 Optimizing the hyperparameters

RF has three hyperparameters, namely the number of trees (n_{tree}), the number of randomly selected variables on each split (m_{try}) and the minimum number of samples in terminal nodes (*nodesize*). A sufficiently large number of trees is necessary for a good model performance. However, the computing time increases with increasing n_{tree}. Similarly, a smaller *nodesize* value generally increases the model performance as well as its complexity.

Prior to the main analysis, we performed a grid search to find the optimal n_{tree} and *nodesize* per scenario. We used a grid from all combinations of $n_{tree} = \{100, 200, ..., 1000\}$ and *nodesize* $= \{1, 2, 3, 4, 5\}$. Rodriguez-Galiano et al. (2012) and Leutner et al. (2012) suggested to optimize n_{tree} and *nodesize* simultaneously. However, we determined them independently of each other because this method was less sensitive to variations between partitions and led to a more stable parameter selection. The hyperparameter m_{try} was set to the square root of the number of features (Clark et al., 2012).

We used an SVM classifier with a Gaussian radial basis function (RBF) kernel. It has two hyperparemters, namely σ and C. Caputo et al. (2002) have shown that optimal values of σ can be determined based on the training data and lie between the inverse of the 10% and 90% quantiles of the distance between the points. We set it to the inverse of the median.

C is a regularization parameter and controls the trade-off between model complexity and misclassification. A large value of C strongly penalizes misclassification and might lead to a winding decision boundary and thus to overfitting. In contrast, smaller values of C tolerate more misclassification and force the boundary to be smoother (e.g. Hastie et al., 2009). Unlike σ, the hyperparameter C has to be tuned. Therefore, we performed a grid search. We first used a coarse grid $C = 10^{-2,-1,...,5}$ and then refined it around the area of good performance of the classifier and tried $C = \{0.01, 0.10, 0.20, 0.50, 0.75, 1.00, 2, 4, 8, 10, 15, 20, 25, 30\}$.

We optimized the hyperparameters in an internal cross validation. First, we split the training folds again into 5 folds, trained the classifier on 4 folds with different parameter values and then predicted the hold-out fold. This internal cross validation ensures that parameters are optimized on the training data only. We compared the optimization based on *F-score* with parameters chosen by the classification error (the proportion of misclassified points). In other words we determined n_{tree} and *nodesize* leading to the maximum *F-score* or to the minimum classification error and compared their performance in the scenarios.

The calculations were done in `GNU R` (R Core Team, 2014) using the R packages `randomForest` version 4.6–7 (Liaw et al., 2002) and `kernlab` (Karatzoglou et al., 2004).

4.4 Results

4.4.1 Data distribution and oversampling rate

The distribution of the training data in different scenarios either equalled that of the original data (S1) or was altered by removal of Tomek links (S2), additional oversampling (S3) or additional over- and undersampling (S4) to approach a balance between the LULC classes (Supplementary Figure 4.10).

The scenario S1 was characterized by a high imbalance ratio of the original data set of approximately 100 : 1. Even the imbalance of the minority classes was still approximately 10 : 1 (the ratio between 'fallow' and 'cabbage'). Because only 48 points (on average) were removed in each training fold of the 5 repetitions, the imbalance ratio in S2 remained comparable to S1. In contrast, beside the dominance of 'deciduous forest', the classes became evenly distributed in S3 after synthetic oversampling. Finally, the combination of synthetic oversampling of minority classes and random undersampling of the majority class 'deciduous forest' generated a nearly equal distribution of classes in S4. Note that Tomek links were also removed in S3 and S4.

The synthetic oversampling rate N ranged between 100% for 'fallow' and 2300% for 'cabbage' (Supplementary Table 4.3). In other words, for every existing point of class 'fallow' one new sample was generated, while 23 new instances were created for every pixel in the class 'cabbage'. In their work Chawla et al. (2002) recommended to choose the oversampling rate not larger than the number of nearest neighbours k. For $k = 5$, for example, the oversampling rate should not exceed 500%. However, Maciejewski et al. (2011) have shown that best classification results on data sets with a high imbalance ratio could be achieved with N four to five times larger than k. In our study, we have chosen $k = 5$, thus oversampling rates up to 2300% are large but likely not excessive.

4.4.2 Optimized hyperparameters

We compared the optimization of the RF (n_{tree} and *nodesize*) and SVM (C) hyperparameters based on *F-score* and the classification error (Supplementary Figures 4.11 and 4.12). *F-score* responded sensitively to variations of n_{tree} and *nodesize* and varied stronger between different folds compared to the classification error. It occasionally preferred simpler parameters (i.e. smaller n_{tree} and larger *nodesize*) while the classification error generally pointed to more complex parameters (i.e. larger n_{tree} and smaller *nodesize*).

F-score and the classification error both lacked a sharp extreme (maximum for *F-score* and min-

imum for classification error) and were rather flat for a large range of C values (Supplementary Figure 4.13). To avoid selecting an unnecessarily large C, we chose the smallest value leading to an *F-score* that was at most 5% smaller than the maximum or leading to a classification error that was at most 5% larger than the minimum. Compared to the classification error, *F-score* varied more between the training folds.

In general, both measures suggested rather small values for C. However, in scenarios S1 and S2 on unbalanced data, *F-score* chose a larger C than the classification error and vice versa in scenarios S3 and S4 on balanced data.

Because we are interested in increasing the classification performance on minority classes, we selected the hyperparameters based on *F-score*.

4.4.3 Entropy and mutual information

The classification problem on the original data was simpler compared to the other scenarios. Indeed, S1 had the lowest entropy (2.3) because the class distribution was dominated by 'deciduous forest'. Thus, the classifier could obtain a high overall accuracy (0.97, c.f. Supplementary Table 4.17) by classifying the majority classes correctly and ignoring at least some of the minority classes. SMOTE increased the difficulty of the classification in scenarios S3 and S4 as indicated by the larger entropy (3.9 and 4.1, respectively) by balancing the distribution.

The mutual information MI^* between the surface reflectance and the class labels in the training data of scenario S1 was relatively large. However, it varied between the four MODIS bands (Figure 4.6). Especially in spring and late summer, MI^* decreased in the red channel B1 and the near-infrared channel B2, both sensitive to vegetation viability. Additionally, these periods showed the largest variability between the different runs in S1 and S2 and had therefore the broadest 5% to 95% quantile ranges.

The removal of Tomek links in S2 increased MI^* only slightly without affecting its temporal shape. In contrast, compared to the original data, SMOTE raised it noticeably and decreased its temporal variability in scenarios S3 and S4. Indeed, the minima of MI^* in B1 and B2 in spring and summer became less pronounced. In contrast, random undersampling of the majority class 'deciduous forest' in S4 hardly affected the mutual information.

4.4.4 Classification performance

We ran the scenarios five times and evaluated each repetition. To assess the classification of single LULC classes, we used *TPR* and *FPR* (Figures 4.7a for RF and 4.7e for SVM) and for

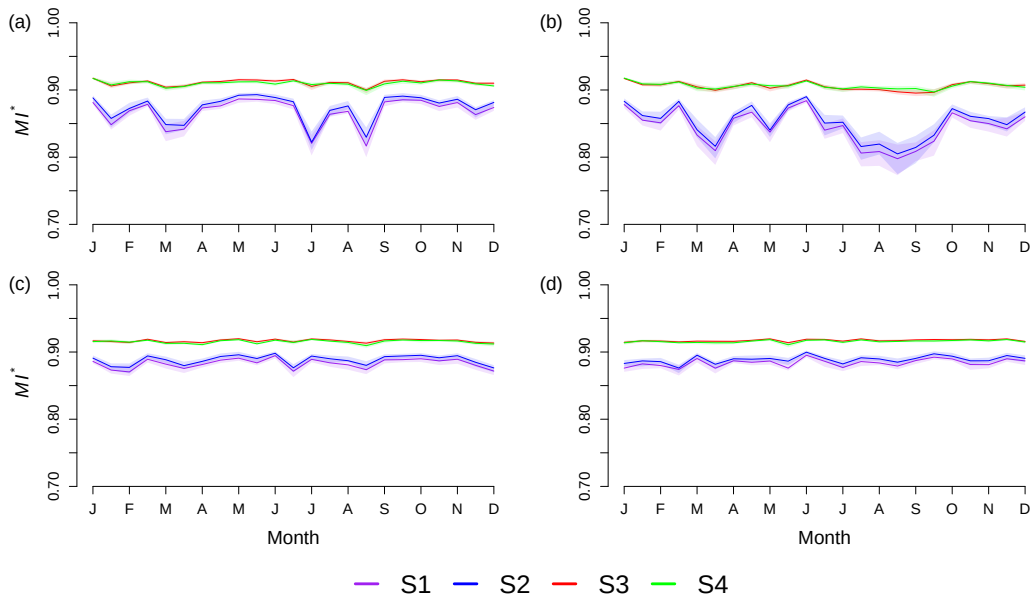

Fig. 4.6 Mutual information MI^* between class labels and predictors (i.e. MODIS spectral bands) for 5 repetitions on 6 training folds in scenarios S1 through S4. (a) red channel B1, (b) near-infrared channel B2, (c) blue channel B3 and (d) mid-infrared channel B7. The plain lines show the median and the shaded areas the 5% to 95% quantile range.

overall performance in a scenario we calculated the measures introduced in Section 4.3.4.

4.4.4.1 Classification of single LULC classes

In S1 the median *TPR*s of the two majority classes 'deciduous forest' and 'paddy rice' were largest and equalled 97% and 89% (RF) and 95% and 85% (SVM), respectively (c.f. Supplementary Tables 4.4, 4.8 and 4.18 for detailed summaries). However, because many minority classes were falsely classified as 'deciduous forest', its median *FPR*s reached 29% (RF) and 26% (SVM). In contrast, the median *FPR*s of 'paddy rice' were low and equalled only 6% (RF) and 5% (SVM).

In general, we observed a positive relationship between the median *TPR*s and the number of pixels in minority classes for both classifiers. However, some comparably large classes like 'mixed forest' and 'bean' or 'shrub' and 'dry field' behaved differently. In addition, RF failed to detect five classes while only two classes had a zero *TPR* when classified with SVM.

In S2 the removal of Tomek links in the majority class 'deciduous forest' decreased its median *FPR* by more than 9%. In contrast, it affected the *FPR*s in the other classes only slightly. Additionally, the median *TPR*s of some of the minority classes increased, particularly when classified with RF (Figure 4.7, b and f, and Supplementary Tables 4.5, 4.9 and 4.18).

The synthetic oversampling in S3 generally increased the median *TPR*s in the minority classes

(Figure 4.7, c and g, and Supplementary Table 4.18), particularly in 'codonopsis', 'urban' and 'shrub'. However, they decreased in the majority classes and in three minority classes ('fallow', 'semi natural' and 'ginseng' when classified RF and 'fallow', 'semi natural' and 'potato' with classified with SVM). The number of classes with zero median *TPR* dropped from five to one when classified with RF and to zero when classified with SVM.

Finally, due to additional random undersampling of 'deciduous forest' in S4 the *TPR*s of some minority classes like 'mixed forest' or 'fallow' increased compared to S3 (Figure 4.7, d and h, and Supplementary Tables 4.7, 4.11 and 4.18). However, the median *FPR*s also increased.

Although the number of 'deciduous forest' pixels was reduced substantially, its median *TPR* decreased only by 9% with RF (from 91% to 80%) and 8% with SVM (from 86% to 78%). Additionally, its *FPR* also dropped from 12% to 3% (RF) and 8% to 3% (SVM).

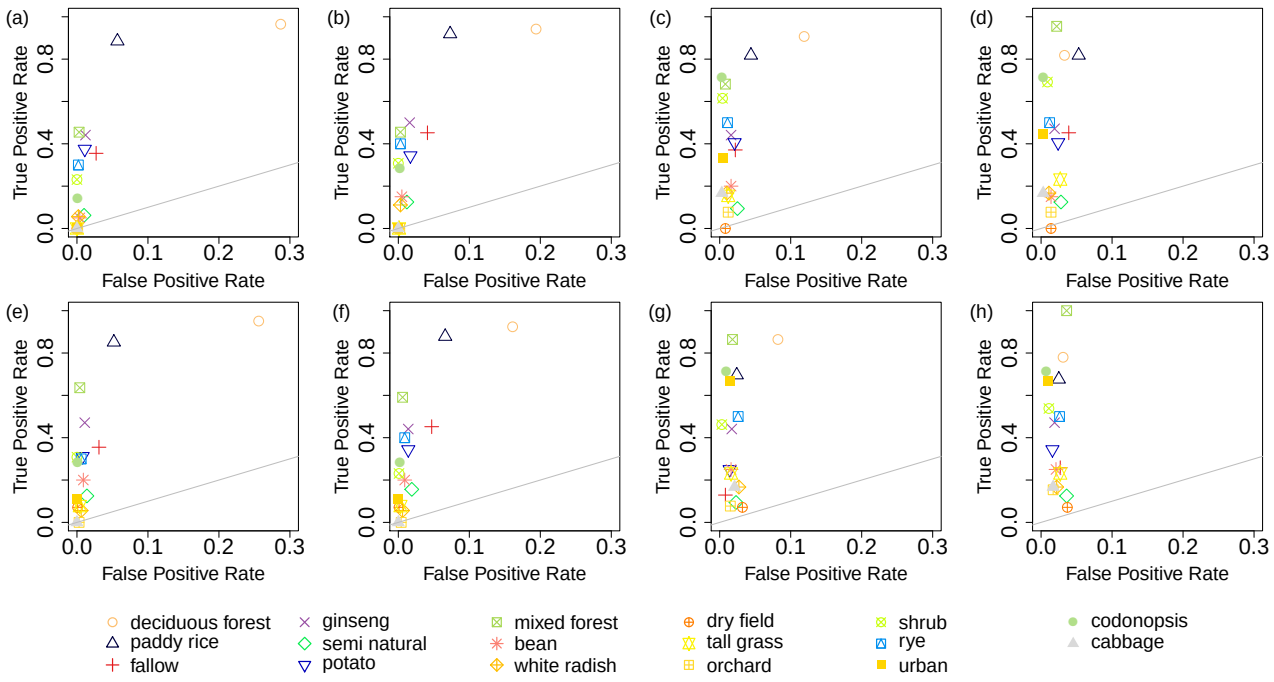

Fig. 4.7 ROC graphs for scenarios S1 through S4 with the RF (upper row) and the SVM (lower row) classifiers. The hyperparameters n_{tree} and *nodesize* for RF and C for SVM were selected based on *F-score*. Median *TPR*s and *FPR*s from 5 repetitions. Note the difference between scales on the x- and y-axis. A point on the diagonal (grey line) indicates a random guess. The order of the classes in the legend reflects the decreasing number of original pixels. The ROC graph based on the parameters selected via the classification error is included in the online Supplementary Material (Figure 4.14) for comparison.

4.4.4.2 Overall performance of scenarios

The median *F-score* based on the five repetitions was comparable between scenarios and ranged from 0.37 to 0.39 (RF) and from 0.35 to 0.38 (SVM), respectively (Supplementary Table 4.12).

The median *NID* varied slightly more, namely between 0.56 and 0.66 (RF) and from 0.62 to 0.65 (SVM) (Supplementary Table 4.13). Due to the zero *TPR*s of the minority types, the *G-mean* was mostly zero except in S4 (0.32) when using SVM (Supplementary Table 4.14).
Although *F-score* remained stable, precision and recall were affected by SMOTE. Actually, precision decreased from 0.6 (RF) and 0.48 (SVM) in scenarios without oversampling (mean value for S1 and S2) to 0.36 (RF) and 0.31 (SVM) in scenarios with SMOTE (mean value for S3 and S4) (Supplementary Table 4.15). In contrast, recall increased from 0.28 (RF) and 0.30 (SVM) to 0.41 (RF) and 0.40 (SVM) (Supplementary Table 4.16). Thus, in scenarios with oversampling the classifiers found more false positives. However, they also identified more actually positive instances.

4.4.4.3 Predicted land use and land cover as a map

Figure 4.8 shows the predicted classes from the repetitions with the largest *F-score* as a map (c.f. Supplementary Table 4.17 for more details). In S1 and S2 minority classes at the boundary between 'deciduous forest' and the agricultural area as well as in the center of the catchment were underrepresented (Figure 4.8, a, b, e and f). In contrast, in scenarios with SMOTE some of them (like 'urban', 'shrub' and 'codonopsis') appeared clearer (Figure 4.8, c, d, g and h). Random undersampling of 'deciduous forest' in S4 particularly affected the 'mixed forest' class in the eastern part of the catchment (Figure 4.8, d and h).
However, synthetic oversampling also increased *FPR*s and decreased precision in S3 and S4. In other words, the falsely classified minority classes increased. For example, the forest edge pixels in the northeastern part of the catchment were misclassified as dry field crops such as 'rye' or 'tall grass' (Figure 4.8, c, d, g and h).

4.5 Discussion

4.5.1 Influence of data resampling on classification performance

SMOTE increased the difficulty of the classification task by balancing the distribution and therefore prevented the classifier from specializing on majority classes. Additionally, it raised the mutual information MI^* between the LULC classes and the predictors. Thus, the predictors contained more information on class labels in data sets augmented by synthetic oversampling than in the original data set.
Both kinds of undersampling – removing Tomek links in S2–S4 and random undersampling in

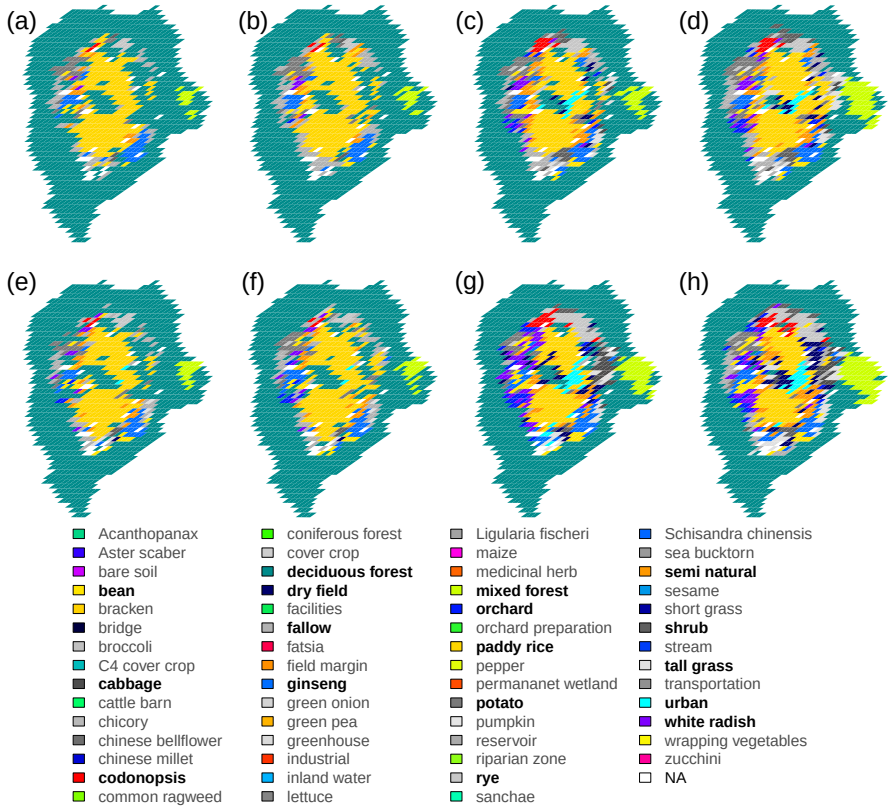

Fig. 4.8 Predicted land use and land cover classes of scenarios (a) S1, (b) S2, (c) S3, and (d) S4 using RF and (e) S1, (f) S2, (g) S3, and (h) S4 using SVM. The Maps from repetitions with the largest *F-score*. Classes with less than 6 original pixels are marked as 'NA'.

S4 – decreased the *FPR* of the majority class 'deciduous forest' substantially and its *TPR* only slightly. While it is quite obvious that removing Tomek links cleans noisy pixels in 'deciduous forest', the improvement due to random undersampling is less clear. One possible explanation is that it balanced the distribution of training data additionally to synthetic oversampling.

Obviously, SMOTE decreased the imbalance ratio. In scenarios S1 and S2 we observed a positive relationship between the number of training pixels (i.e. the size of a class) and the ability of the classifier to recognize a pixel's class correctly. Actually, Spearman's rank correlation coefficient between the median *TPR* and the number of training pixels were approximately 0.7 in S1 and 0.8 in S2, on average over the classifiers (Figure 4.9a). In contrast, the correlations dropped sharply in S3 and S4 regardless of the classifier.

Although SMOTE decreased the imbalance ratio and increased the mutual information between the LULC classes and the predictors, classification of some minority classes remained difficult. To better understand why minority classes behave differently, we evaluated the neighbourhood of the test data. Figure 4.9b shows the relationship between the median *TPR* and the median of the proportion of five nearest neighbours of the test data in the training data that belong to the same class as the test data. The Spearman correlation remained quite high in all scenarios

indicating that pixels with a low number of nearest neighbours of the same class in the training data were often misclassified.

The proportion of nearest neighbour of the same class is inversely related to the class overlap which occurs if a region in feature space contains comparable numbers of points from different classes. Such classes are particularly difficult to distinguish. SMOTE increased the proportion of neighbours of the same class in some minority classes (Supplementary Figures 4.15 through 4.18). However, classes like 'dry field' and 'orchard' retained a largely varying proportion of nearest neighbours and a median proportion of zero even after synthetic oversampling. Accordingly, these classes were particularly difficult to classify.

The optimized hyperparameter C of the SVM classifier might also indicate a substantial class overlap. Specifically, a small C value induces a large margin and indicates that a further decrease of the margin fails to increase the classification accuracy because the classes are not separable.

SVM performed better than RF in the baseline scenario S1. This is in agreement with Dudoit et al. (2003) who reported that RF performed poorly under data imbalance. However RF marginally outperformed SVM when trained on the synthetically oversampled data. In S3 and S4, the median *F-score*, precision, and recall were larger and *NID* was smaller compared to SVM (Supplementary Table 4.12, 4.13, 4.15, and 4.16). Additionally, the decrease of *TPR*s of majority classes and some of the larger minority classes (like 'fallow', 'semi natural' or 'potato') was less pronounced when classified with RF (Supplementary Table 4.18). Similarly, performance measures indicated a better agreement between the maps and the reference LULC data when RF was used as classifier (Supplementary Table 4.17).

4.5.2 Issues related to learning

Our findings on the difficulty of classification related to class overlap are in agreement with previously published results. Although it has been widely accepted that class imbalance is responsible for a drop in classification performance, several recent studies report that imbalance *per se* does not prevent learning (He et al., 2009; López et al., 2013; Sun et al., 2009). It is rather a combination of imbalance and intrinsic characteristics of the data like small sample size, possible sub-concepts (i.e. different clusters within single classes), class overlap or noisy data.

Among these intrinsic characteristics, class overlap seems to play a particular role. Prati et al. (2004), for example, reported that class overlap was at least as important as imbalance. They

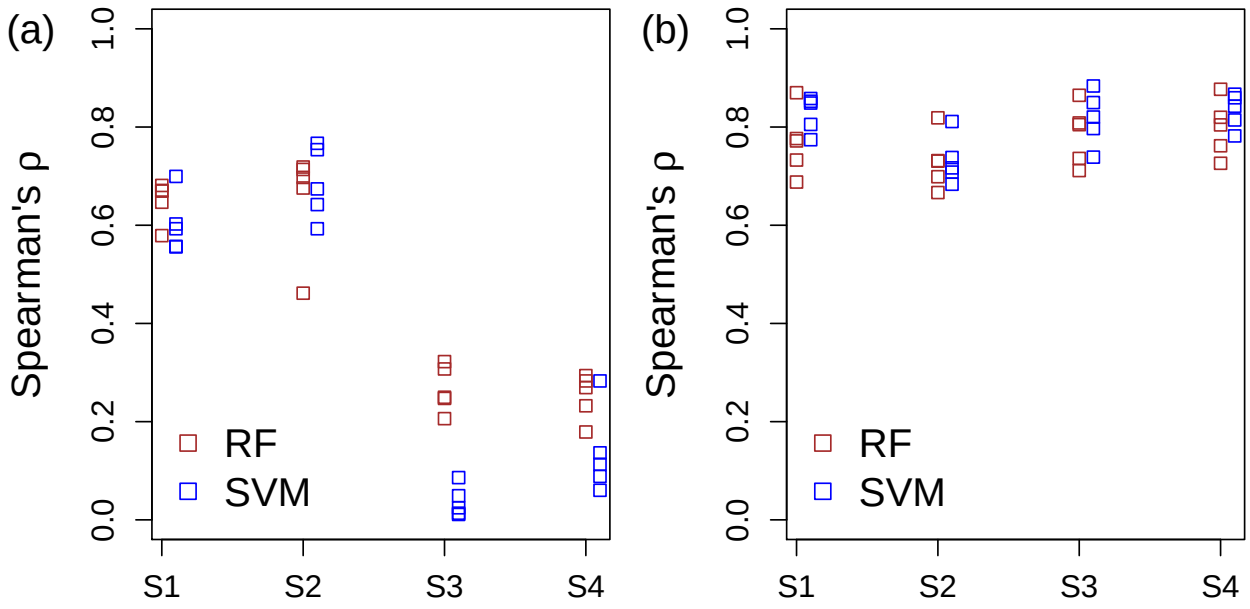

Fig. 4.9 Spearman correlation coefficient between (a) *TPR*s and the class sizes in the training data; (b) *TPR*s and the median of the proportion of five nearest neighbours of the test data in the training data that belong to the same class as the test data. The five points per scenario represent the five repetitions.

have shown on a set of artificial data that for the same imbalance ratio the performance of the classifier decreased with increasing class overlap. Moreover, Denil et al. (2010) described an interaction between overlap and imbalance and mentioned that for a certain degree of overlap, increasing the number of training samples did not improve the classification performance. To our knowledge, there are only few studies in remote sensing explicitly treating class overlap. One of them is the work by Alejo et al. (2013). They also reported that SMOTE could effectively reduce even large class imbalance. However, it failed to increase the classification performance in the presence of class overlap.

Overlap of LULC classes originates from spectral similarities between classes which can be enhanced by mixture of different crops inside the same pixel. Actually, in an agricultural mosaic landscape, mixed pixels containing several crops are common. Therefore, pure pixels that reflect the spectral signature of one particular crop are probably even rarer. Some recent studies reported that SMOTE might increase the class overlap because it does not take the neighbourhood of pixels into account (Bunkhumpornpat et al., 2009; Maciejewski et al., 2011). However, using alternative SMOTE implementations that take the neighbourhood into account will probably not alleviate the substantial overlap already existing in the original data.

Besides the class overlap, the minor crops in our data set are rare in the absolute sense. Indeed, some minority classes contain less than 10 pixels (Table 4.1). This might be insufficient for

a classifier to induce a reliable model about the distribution of the data – an issue known as the "lack of density" or the "lack of information" (López et al., 2013, and references therein). In other words, in imbalanced small data sets the absolute rarity (i.e. small number of data points) and the relative rarity (i.e. large imbalance ratio) amplify each other and make learning difficult.

SMOTE helps to alleviate the issue of class imbalance, increases the number of training points and therefore decreases the absolute rarity. However, splitting small classes for cross validation risks to generate small subclusters (sub-concepts). Indeed, in classes with only 6 points, we left out one point for testing and used 5 points for synthetic oversampling. This single testing point might generate a subcluster and will be particularly difficult to classify in the presence of class overlap. In particular, Japkowicz et al. (2002) suggested that the SVM classifier was insensitive to class imbalance. Instead it is affected by the presence of small subclusters that are frequent in imbalanced data sets with few training points.

4.6 Summary and conclusions

The classification of the original imbalanced data set was particularly challenging due to a small number of training points in the minority classes (and thus a possible presence of small subclusters) and the class overlap. SMOTE helped to alleviate the issue of class imbalance and increased the number of training points.

Balancing the data distribution by synthetically oversampling the minority classes enhanced the relationship between the class labels and the reflectance data (larger mutual information MI^*).

Synthetic oversampling increased the true positive rates of some minority classes substantially compared to the original imbalanced data set. However, due to class overlap some of the minority classes remained difficult to classify. Although it decreased precision and increased recall, the combined measure *F-score* remained stable between scenarios.

SVM outperformed RF when trained on the original unbalanced data set. In contrast, RF performed marginally better than SVM when trained on the synthetically oversampled data and produced maps that agreed slightly better with reference LULC data (smaller *NID*). However, the difference in performance between the classifiers was small.

Data preprocessing with SMOTE to balance the data distribution is independent of the classifier. The implementation of the algorithm is straightforward and its functioning is easy to understand. We have used RF and SVM, however, any other classifier could be used on the preprocessed data.

Therefore, synthetic oversampling can be plugged in into an existing classification framework without further adjustments.

When oversampling fails to increase the classification performance, a detailed analysis of the pixels' neighbourhood can yield important information. In particular, in the presence of class overlap, increasing the number of training points does not guarantee a better classification result.

4.7 Acknowledgements

This study was carried out as part of the International Research Training Group TERRECO (GRK 1565/1) funded by the Deutsche Forschungsgemeinschaft (DFG) at the University of Bayreuth, Germany and the Korean Research Foundation (KRF) at Kangwon National University, Chuncheon, South Korea.

The MOD13Q1 data product is curtsey of the online Data Pool at the NASA Land Processes Distributed Active Archive Center (LP DAAC) at the USGS/Earth Resources Observation and Science (EROS) Center (https://lpdaac.usgs.gov).

References

Akbani, R., S. Kwek & N. Japkowicz (2004). "Applying support vector machines to imbalanced datasets". In: *Machine Learning: ECML 2004*. Springer, pp. 39–50 (cit. on pp. 12, 107).

Alejo, R., J. M. Sotoca, R. M. Valdovinos & G. A. Casan (2009). "The multi-class imbalance problem: cost functions with modular and non-modular neural networks". In: *The Sixth International Symposium on Neural Networks (ISNN 2009)*. Springer, pp. 421–431 (cit. on p. 106).

Alejo, R., R. M. Valdovinos, V. García & J. Pacheco-Sanchez (2013). "A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios". In: *Pattern Recognition Letters* 34.4, pp. 380–388 (cit. on p. 125).

Attarchi, S. & R. Gloaguen (2014). "Classifying complex mountainous forests with L-Band SAR and Landsat data integration: A comparison among different machine learning methods in the hyrcanian forest". In: *Remote Sensing* 6.5, pp. 3624–3647 (cit. on pp. 10, 72, 107).

Batista, G. E., R. C. Prati & M. C. Monard (2004). "A study of the behavior of several methods for balancing machine learning training data". In: *ACM Sigkdd Explorations Newsletter* 6.1, pp. 20–29 (cit. on p. 113).

Bivand, R. & C. Rundel (2014). *rgeos: Interface to Geometry Engine - Open Source (GEOS)* (cit. on pp. 45, 77, 109).

Bontemps, S., P. Defourny, E. Bogaert, O. Arino, V. Kalogirou & J. Perez (2011). *GLOBCOVER 2009 - Products Description and Validation Report*. Tech. rep. European Space Agency (cit. on pp. 4, 5, 7–9, 38, 63, 64, 105).

Breiman, L (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32 (cit. on pp. 10, 12, 72, 73, 76, 107).

Brown, J. C., J. H. Kastens, A. C. Coutinho, D. d. C. Victoria & C. R. Bishop (2013b). "Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data". In: *Remote Sensing of Environment* 130, pp. 39–50 (cit. on p. 106).

Bruzzone, L. & S. B. Serpico (1997). "Classification of imbalanced remote-sensing data by neural networks". In: *Pattern Recognition Letters* 18.11, pp. 1323–1328 (cit. on pp. 12, 106).

Bunkhumpornpat, C., K. Sinapiromsaran & C. Lursinsap (2009). "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem". In: *Advances in Knowledge Discovery and Data Mining*. Springer, pp. 475–482 (cit. on p. 125).

Camps-Valls, G., L. Bruzzone, et al. (2009). *Kernel methods for remote sensing data analysis*. Vol. 26. Wiley Online Library (cit. on p. 107).

Caputo, B, K Sim, F Furesjo & A Smola (2002). "Appearance-based object recognition using SVMs: which kernel should I use?" In: *Proceedings of NIPS workshop on Statitsical methods for computational experiments in visual processing and computer vision*. Whistler, Canada (cit. on p. 117).

Chawla, N. V. (2010). "Data mining for imbalanced datasets: An overview". In: *Data Mining and Knowledge Discovery Handbook*. Springer, pp. 875–886 (cit. on pp. 12, 106).

Chawla, N. V., K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357 (cit. on pp. 12, 106, 111, 118).

Chawla, N. V., N. Japkowicz & A. Kotcz (2004). "Editorial: special issue on learning from imbalanced data sets". In: *ACM Sigkdd Explorations Newsletter* 6.1, pp. 1–6 (cit. on p. 106).

Chhabra, A., H. Geist, R. A. Houghton, H. Haberl, A. K. Braimoh, P. L. Vlek, J. Patz, J. Xu, N. Ramankutty, O. Coomes & others (2006). "Multiple impacts of land-use/cover change". In: *Land-use and land-cover change*. Springer, pp. 71–116 (cit. on pp. 2, 3, 105).

Clark, M. L., T. M. Aide, H. R. Grau & G. Riner (2010). "A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America". In: *Remote Sensing of Environment* 114.11, pp. 2816–2832 (cit. on pp. 10, 64, 72, 106).

Clark, M. L. & D. A. Roberts (2012). "Species-Level Differences in Hyperspectral Metrics among Tropical Rainforest Trees as Determined by a Tree-Based Classifier". In: *Remote Sensing* 4.12, pp. 1820–1855 (cit. on pp. 74, 102, 117).

Dawson, T. P., S. T. Jackson, J. I. House, I. C. Prentice & G. M. Mace (2011). "Beyond predictions: biodiversity conservation in a changing climate". In: *Science* 332.6025, pp. 53–58 (cit. on pp. 2, 105).

Denil, M. & T. Trappenberg (2010). "Overlap versus imbalance". In: *Advances in Artificial Intelligence*. Springer, pp. 220–231 (cit. on p. 125).

Dudoit, S. & J. Fridlyand (2003). "Bagging to Improve the Accuracy of A Clustering Procedure." In: *Bioinformatics* 19.9, pp. 1090–1099 (cit. on p. 124).

Fawcett, T. (2006). "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8, pp. 861–874 (cit. on p. 115).

Fernández, A., S. García & F. Herrera (2011). "Addressing the classification with imbalanced data: open problems and new challenges on class distribution". In: *Hybrid Artificial Intelligent Systems*. Springer, pp. 1–10 (cit. on pp. 12, 106).

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs & others (2005). "Global consequences of land use". In: *Science* 309.5734, pp. 570–574 (cit. on pp. 2, 105).

García, V., J. S. Sánchez & R. Mollineda (2011). "Classification of high dimensional and imbalanced hyperspectral imagery data". In: *Pattern Recognition and Image Analysis*. Ed. by J. Vitrià, J. M. Sanches & M. Hernández. Vol. 6669. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 644–651 (cit. on pp. 12, 106).

GEOS Development Team (2014). *GEOS - Geometry Engine, Open Source*. Open Source Geospatial Foundation (cit. on pp. 45, 77, 109).

Hastie, T., R. Tibshirani & J. Friedman (2009). *The elements of statistical learning*. 2nd Edition. Springer (cit. on p. 117).

He, H. & E. A. Garcia (2009). "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9, pp. 1263–1284 (cit. on pp. 12, 106, 124).

Hoffmann, M., C. Hilton-Taylor, A. Angulo, M. Böhm, T. M. Brooks, S. H. Butchart, K. E. Carpenter, J. Chanson, B. Collen, N. A. Cox & others (2010). "The impact of conservation on the status of the world's vertebrates". In: *Science* 330.6010, pp. 1503–1509 (cit. on pp. 2, 105).

Huete, A., C. Justice & W. Van Leeuwen (1999). *MODIS Vegetation Index (MOD 13): Algorithm Theoretical Basis Document*. Tech. rep. (cit. on pp. 65, 69, 108).

Hüttich, C., U. Gessner, M. Herold, B. J. Strohbach, M. Schmidt, M. Keil & S. Dech (2009). "On the Suitability of MODIS Time Series Metrics to Map Vegetation Types in Dry Savanna Ecosystems: A Case Study in the Kalahari of NE Namibia". In: *Remote Sensing* 1.4, pp. 620–643 (cit. on pp. 10, 64, 72, 86, 106).

Immitzer, M., C. Atzberger & T. Koukal (2012). "Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data". In: *Remote Sensing* 4.12, pp. 2661–2693 (cit. on pp. 10, 12, 72, 107).

Japkowicz, N. & S. Stephen (2002). "The class imbalance problem: A systematic study". In: *Intelligent data analysis* 6.5, pp. 429–449 (cit. on pp. 12, 106, 126).

Joe, H (1989). "Relative entropy measures of multivariate dependence". In: *Journal of the American Statistical Association* 84.405, pp. 157–164 (cit. on p. 114).

Johnson, B. A., R. Tateishi & N. T. Hoan (2013). "A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees". In: *International Journal of Remote Sensing* 34.20, pp. 6969–6982 (cit. on p. 107).

Karatzoglou, A., A. Smola, K. Hornik & A. Zeileis (2004). “kernlab – An S4 Package for Kernel Methods in R”. In: *Journal Of Statistical Software* 11.9, pp. 1–20 (cit. on p. 117).

Kononenko, I. & I. Bratko (1991). “Information-based evaluation criterion for classifier’s performance”. In: *Machine Learning* 6.1, pp. 67–80 (cit. on pp. 110, 111).

Kraskov, A, H Stögbauer & P Grassberger (2004). “Estimating mutual information”. In: *Physical Review E* 69.6, p. 066138 (cit. on p. 114).

Kraskov, A., H. Stögbauer, R. G. Andrzejak & P. Grassberger (2005). “Hierarchical clustering using mutual information”. In: *EPL (Europhysics Letters)* 70.2, p. 278 (cit. on p. 115).

Kubat, M. & S. Matwin (1997). “Addressing the curse of imbalanced training sets: One-sided selection”. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, pp. 179–186 (cit. on p. 115).

Leutner, B. F., B. Reineking, J. Müller, M. Bachmann, C. Beierkuhnlein, S. Dech & M. Wegmann (2012). “Modelling Forest α-Diversity and Floristic Composition — On the Added Value of LiDAR plus Hyperspectral Remote Sensing”. In: *Remote Sensing* 4.12, pp. 2818–2845 (cit. on pp. 74, 117).

Liaw, A. & M. Wiener (2002). “Classification and Regression by randomForest”. In: *R news* 2.3, pp. 18–22 (cit. on pp. 12, 72, 117).

Ling, C. X. & C. Li (1998). “Data mining for direct marketing: Problems and solutions.” In: *Proceedings of the Fourth International Conference of Knowledge Discovery and Data Mining (KDD-98)*. Vol. 98, pp. 73–79 (cit. on pp. 12, 106).

López, V., A. Fernández, S. García, V. Palade & F. Herrera (2013). “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”. In: *Information Sciences* 250, pp. 113–141 (cit. on pp. 124, 126).

Loveland, T. R., B. C. Reed, J. F. Brown, D. O. Ohlen, Z Zhu, L Yang & J. W. Merchant (2000). “Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data”. In: *International Journal Of Remote Sensing* 21.6-7, pp. 1303–1330 (cit. on pp. 4, 8, 37, 38, 43, 64, 105).

Maciejewski, T. & J. Stefanowski (2011). “Local neighbourhood extension of SMOTE for mining imbalanced data”. In: *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, pp. 104–111 (cit. on pp. 118, 125).

Mountrakis, G., J. Im & C. Ogole (2011). “Support vector machines in remote sensing: A review”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.3, pp. 247–259 (cit. on pp. 10, 107).

NASA Land Processes Distributed Active Archive Center (LP DAAC) (2013b). *MOD13Q1 Vegetation Indices 16-Day L3 Global 250m*. 47914 252nd Street, Sioux Falls, South Dakota (cit. on p. 108).

Nitze, I., B. Barrett & F. Cawkwell (2015). "Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series". In: *International Journal Of Applied Earth Observation And Geoinformation* 34, pp. 136–146 (cit. on pp. 10, 72, 106).

Numata, J, O Ebenhöh & E. W. Knapp (2008). "Measuring correlations in metabolomic networks with mutual information". In: *Genome Informatics* 20, pp. 112–122 (cit. on p. 114).

Pittman, K., M. C. Hansen, I. Becker-Reshef, P. V. Potapov & C. O. Justice (2010). "Estimating Global Cropland Extent with Multi-year MODIS Data". In: *Remote Sensing* 2.7, pp. 1844–1863 (cit. on pp. 8, 9, 11, 37, 38, 56, 63, 64, 86, 105).

Pozzolo, A. D., O. Caelen & G. Bontempi (2014). *unbalanced: The package implements different data-driven method for unbalanced datasets* (cit. on p. 113).

Prati, R. C., G. E. Batista & M. C. Monard (2004). "Class imbalances versus class overlapping: an analysis of a learning system behavior". In: *MICAI 2004: Advances in Artificial Intelligence*. Springer, pp. 312–321 (cit. on p. 124).

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on pp. 77, 109, 117).

Rodriguez-Galiano, V. F., B Ghimire, J Rogan, M Chica-Olmo & J. P. Rigol-Sanchez (2012). "An assessment of the effectiveness of a random forest classifier for land-cover classification". In: *Isprs Journal of Photogrammetry and Remote Sensing* 67, pp. 93–104 (cit. on pp. 10, 72, 74, 106, 117).

Sales, G. & C. Romualdi (2012). *parmigene: Parallel Mutual Information estimation for Gene Network reconstruction.* (Cit. on p. 114).

Schistad Solberg, A. & R. Solberg (1996). "A large-scale evaluation of features for automatic detection of oil spills in ERS SAR images". In: *Geoscience and Remote Sensing Symposium, 1996. IGARSS'96.'Remote Sensing for a Sustainable Future.', International*. Vol. 3. IEEE, pp. 1484–1486 (cit. on pp. 12, 106).

Schwieder, M., P. Leitão, S. Suess, C. Senf & P. Hostert (2014). "Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques". In: *Remote Sensing* 6.4, pp. 3427–3445 (cit. on pp. 10–12, 63, 71, 72, 83, 106, 157).

Seo, B., C. Bogner, P. Poppenborg, E. Martin, M. Hoffmeister, M. Jun, T. Koellner, B. Reineking, C. L. Shope & J. Tenhunen (2014). "Deriving a per-field land use and land cover map in an agricultural mosaic catchment". In: *Earth System Science Data* submitted (cit. on pp. 8, 16, 66, 109, 141).

Seo, B., P. Poppenborg, E. Martin, M. Hoffmeister, C. Bogner, H. Elsayed Ali, B. Reineking & J. Tenhunen (2014). *Per-field land use and land cover data set of the Haean catchment, South Korea. doi:10.1594/PANGAEA.823677*. data set (cit. on pp. 15, 19, 21, 67, 96, 109).

Shannon, C. E. (1948). "A mathematical theory of communication". In: *Bell System Technical Journal* 27, pp. 379–423 (cit. on pp. 110, 111, 114).

Sokolova, M. & G. Lapalme (2009). "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4, pp. 427–437 (cit. on p. 115).

Solano, R, K Didan, A Jacobson & A Huete (2010). *MODIS Vegetation Index User's Guide*. v. 2.00. Vegetation Index and Phenology Lab. University of Arizona (cit. on p. 108).

Sterling, S. M., A. Ducharne & J. Polcher (2012). "The impact of global land-cover change on the terrestrial water cycle". In: *Nature Climate Change* 3.4, pp. 385–390 (cit. on pp. 2, 105).

Sun, Y., M. S. Kamel, Y. Wang, et al. (2006). "Boosting for learning multiple classes with imbalanced class distribution." In: *ICDM*. Vol. 6, pp. 592–602 (cit. on p. 115).

Sun, Y., M. S. Kamel, A. K. Wong & Y. Wang (2007). "Cost-sensitive boosting for classification of imbalanced data". In: *Pattern Recognition* 40.12, pp. 3358–3378 (cit. on pp. 12, 106).

Sun, Y., A. K. Wong & M. S. Kamel (2009). "Classification of imbalanced data: A review". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.04, pp. 687–719 (cit. on p. 124).

Thenkabail, P. S., M. Schull & H. Turral (2005). "Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data". In: *Remote Sensing of Environment* 95.3, pp. 317–341 (cit. on pp. 9, 10, 64, 72, 86, 105).

Tomek, I. (1976). "Two modifications of CNN". In: *IEEE Transactions on Systems Man and Cybernetics* 6, pp. 769–772 (cit. on p. 112).

Turner, B. L., E. F. Lambin & A. Reenberg (2007). "The emergence of land change science for global environmental change and sustainability". In: *Proceedings of the National Academy of Sciences* 104.52, pp. 20666–20671 (cit. on pp. 2, 105, 161).

U.S. Geological Survey (2012). *Global Land Cover Characteristics Data Base Version 2.0*. Tech. rep. U.S. Geological Survey (cit. on pp. 4, 8, 38, 53, 64, 105).

Valverde-Albacete, F. J. & C. Peláez-Moreno (2014). “100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox”. In: *PloS one* 9.1, e84217 (cit. on p. 111).

Vinh, N. X., J. Epps & J. Bailey (2010). “Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance”. In: *The Journal of Machine Learning Research* 11, pp. 2837–2854 (cit. on p. 115).

Vuolo, F. & C. Atzberger (2012). “Exploiting the Classification Performance of Support Vector Machines with Multi-Temporal Moderate-Resolution Imaging Spectroradiometer (MODIS) Data in Areas of Agreement and Disagreement of Existing Land Cover Products”. In: *Remote Sensing* 4.12, pp. 3143–3167 (cit. on pp. 10, 107).

Wang, S. & X. Yao (2012). “Multiclass imbalance problems: Analysis and potential solutions”. In: *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 42.4, pp. 1119–1130 (cit. on p. 107).

Waske, B., J. A. Benediktsson & J. R. Sveinsson (2009). “Classifying remote sensing data with support vector machines and imbalanced training data”. In: *Multiple Classifier Systems*. Springer, pp. 375–384 (cit. on pp. 12, 106).

Weiss, G. M. (2004). “Mining with rarity: a unifying framework”. In: 6.1, pp. 7–19 (cit. on p. 115).

Weiss, G. M. & F. J. Provost (2003). “Learning when training data are costly: The effect of class distribution on tree induction”. In: *Journal of Artificial Intelligence Research* 19, pp. 315–354 (cit. on p. 115).

Williams, D. P., V. Myers & M. S. Silvious (2009). “Mine classification with imbalanced data”. In: *Geoscience and Remote Sensing Letters, IEEE* 6.3, pp. 528–532 (cit. on pp. 12, 106).

Supplementary Figures

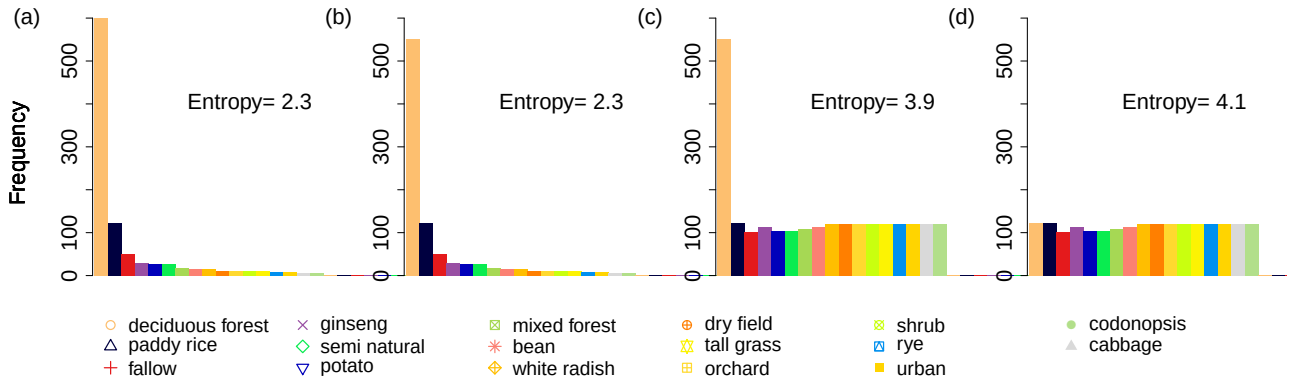

Fig. 4.10 Distribution of the training data sets in different scenarios. (a) S1: original data. (b) S2: original data with Tomek links removed. (c) S3: Tomek links removed and synthetically oversampled minority classes. (d) S4: Tomek links removed, synthetically oversampled minority classes and randomly undersampled majority class 'deciduous forest'.

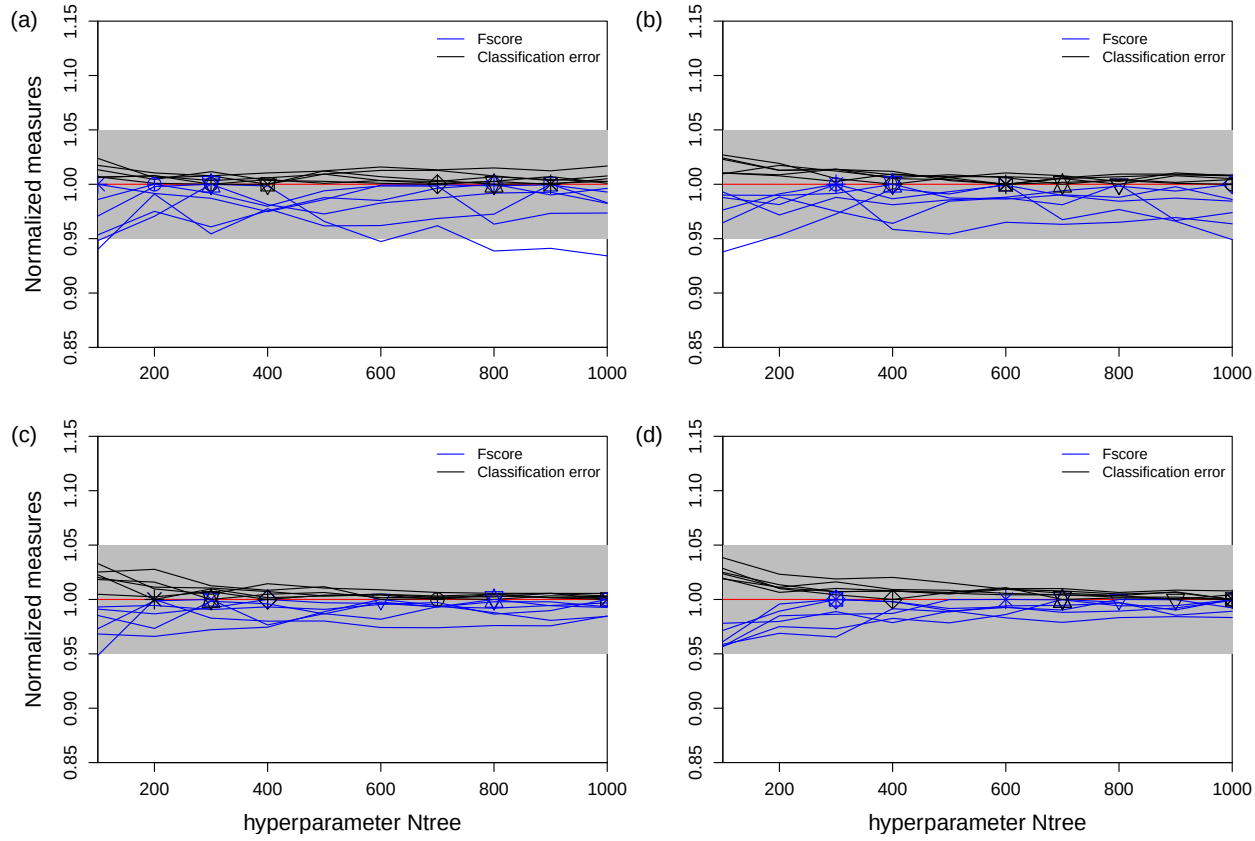

Fig. 4.11 Variation of *F-score* and classification error of RF with changing hyperparameter n_{tree} in 6 training folds in scenarios S1 through S4 (one repetition exemplarily). Both *F-score* and the classification error were normalized by dividing them by their respective maximum or minimum. A horizontal line at one was inserted for convenience. The grey area indicates the 5% threshold and the symbols the chosen n_{tree} for different folds.

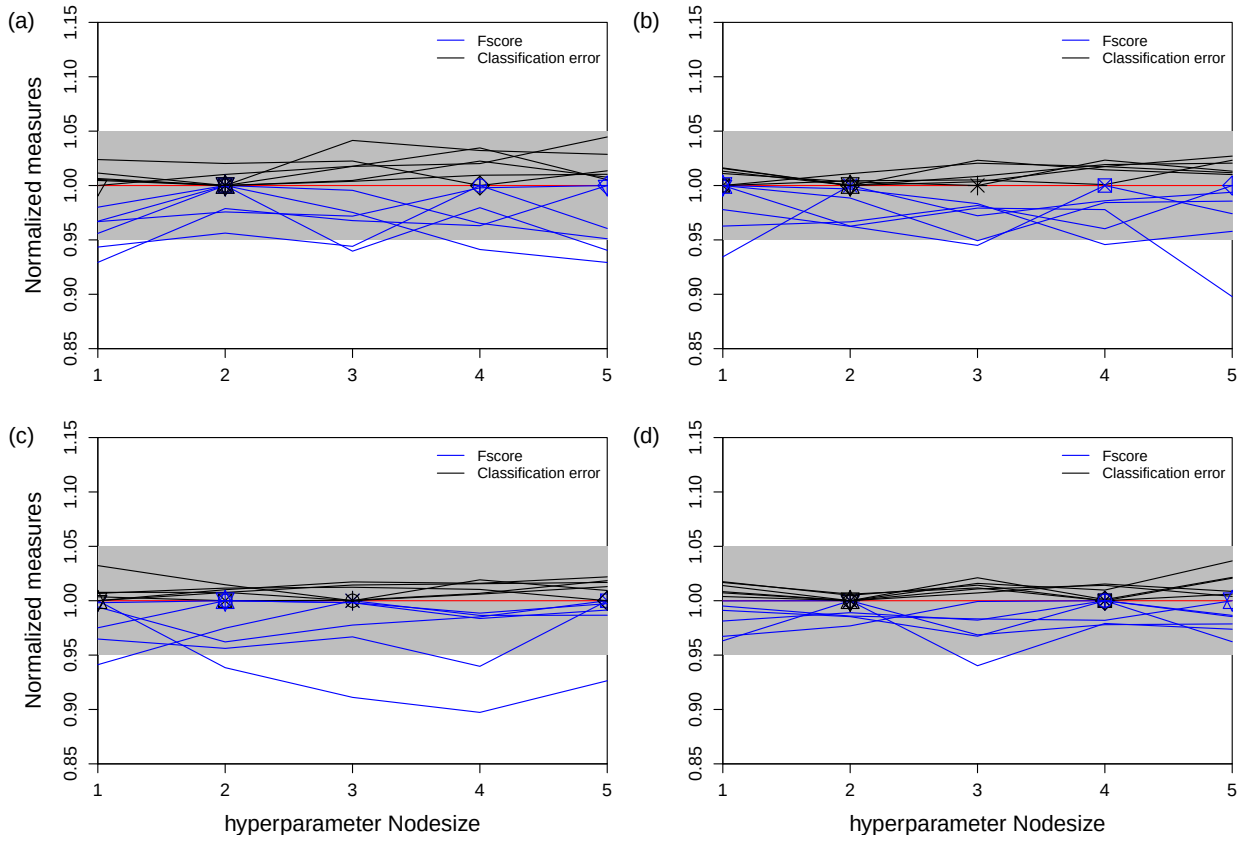

Fig. 4.12 Variation of *F-score* and classification error of RF with changing hyperparameter *nodesize* in 6 training folds in scenarios S1 through S4 (one repetition exemplarily). Both *F-score* and the classification error were normalized by dividing them by their respective maximum or minimum. A horizontal line at one was inserted for convenience. The grey area indicates the 5% threshold and the symbols the chosen *nodesize* for different folds.

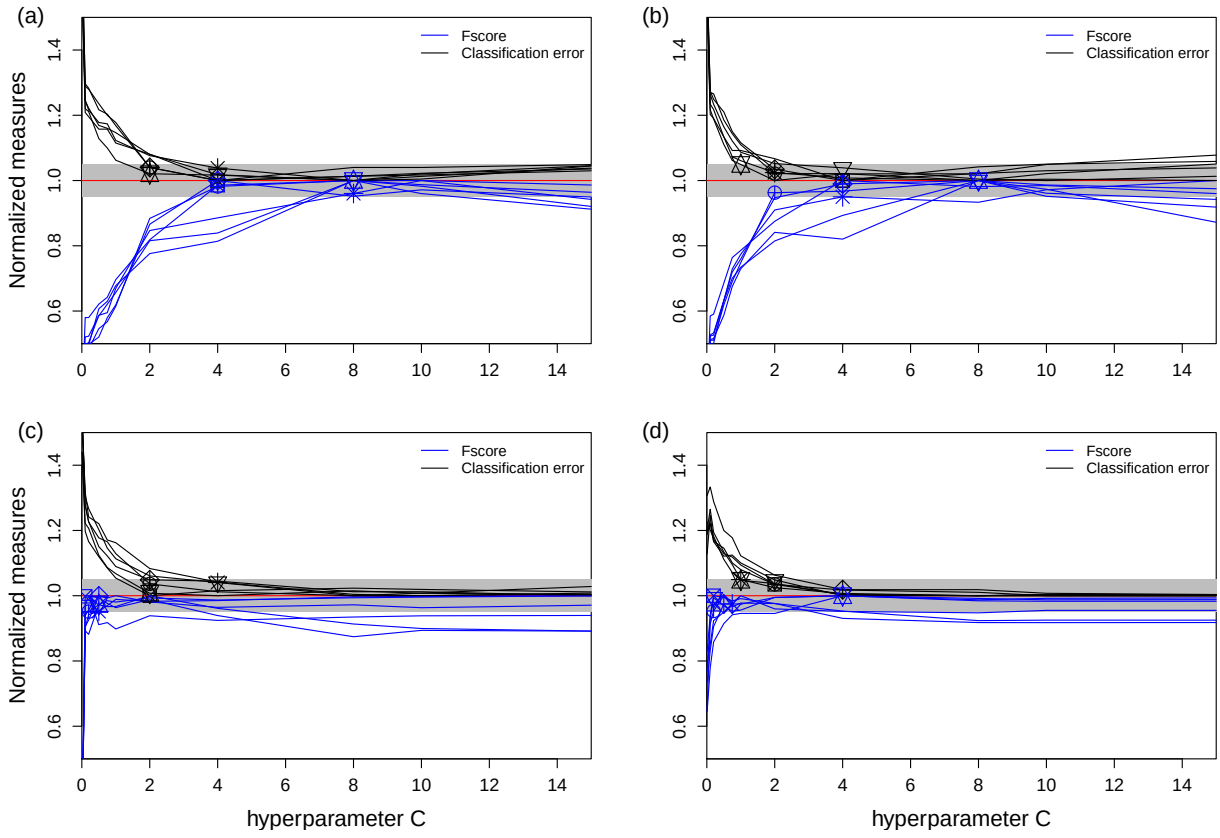

Fig. 4.13 Variation of *F-score* and classification error of SVM with changing hyperparameter C in 6 training folds in scenarios S1 through S4 (one repetition exemplarily). Both *F-score* and classification error were normalized by dividing them by their respective maximum or minimum. A horizontal line at one was inserted for convenience. The grey area indicates the 5% threshold and the symbols the chosen C for different folds.

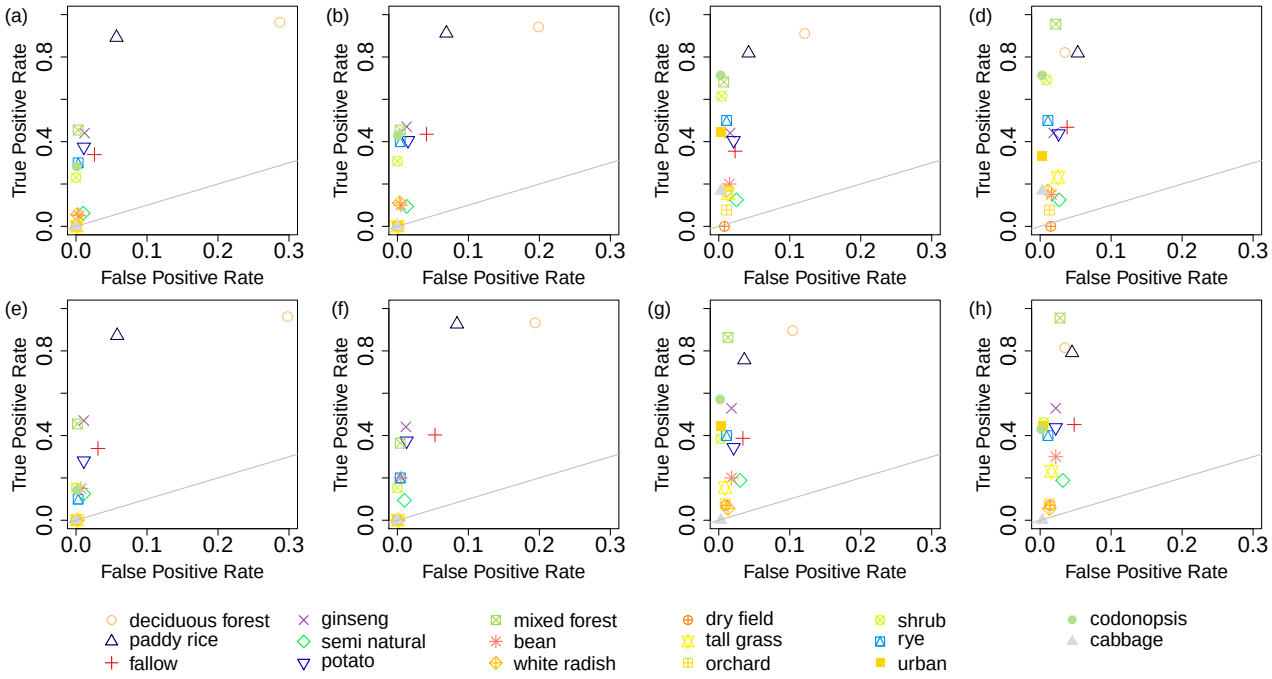

Fig. 4.14 ROC graphs for scenarios S1 through S4 using RF (upper row) and SVM (lower row). The hyperparameters *ntree* and *nodesize* for RF and C for SVM were selected based on the classification error. Median TPRs and FPRs from 5 repetitions. Note the difference between scales on the x- and y-axis. A point on the diagonal (grey line) indicates a random guess. The order of the classes in the legend reflects the decreasing number of original pixels.

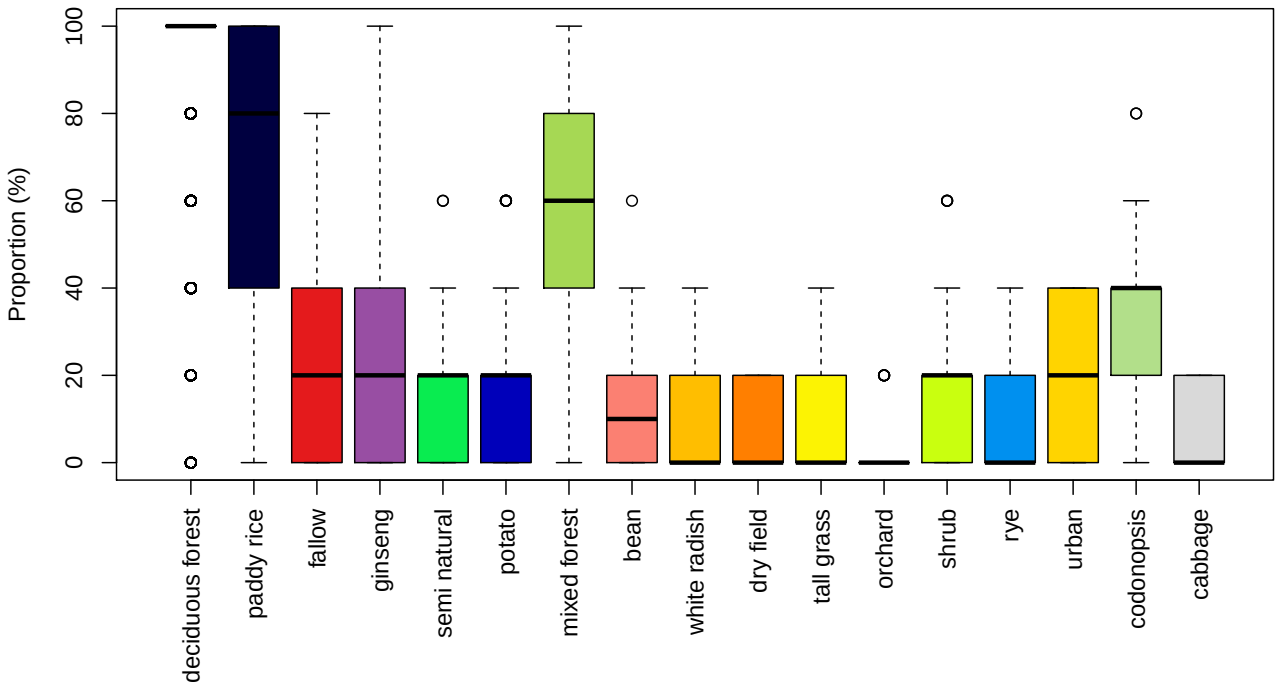

Fig. 4.15 Proportion of five nearest neighbours of the test data in the training data that belong to the same class as the test data in scenario S1.

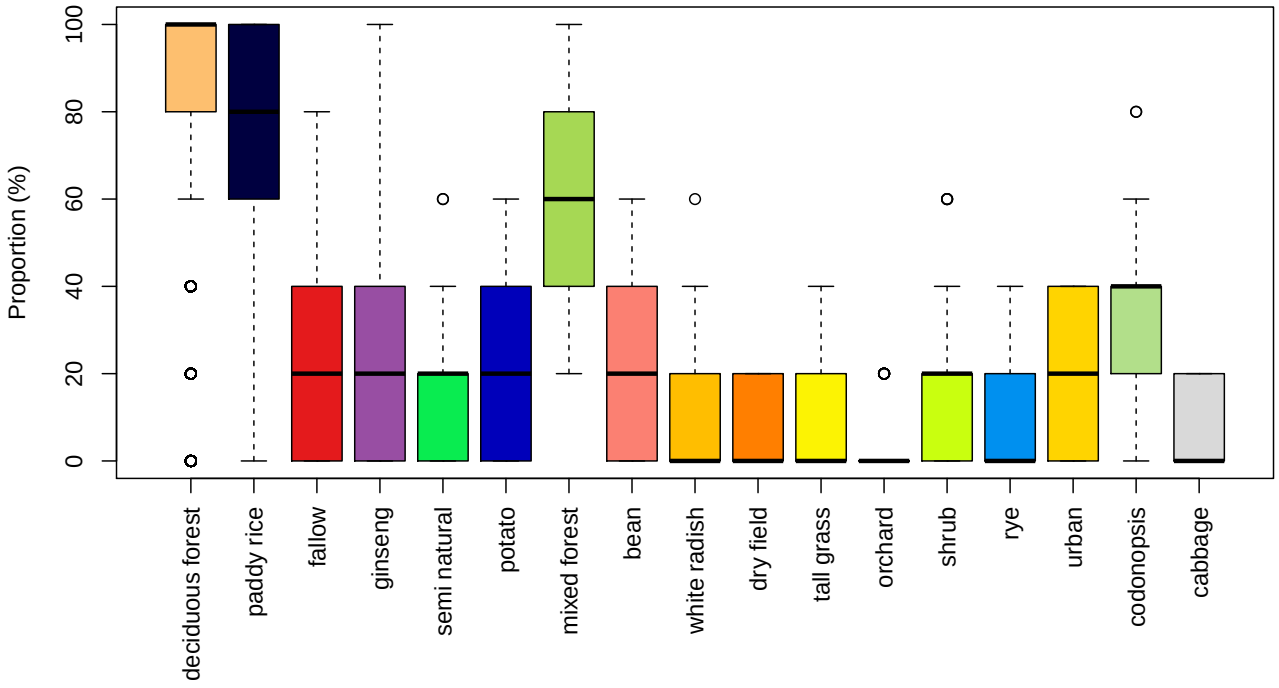

Fig. 4.16 Proportion of five nearest neighbours of the test data in the training data that belong to the same class as the test data in scenario S2.

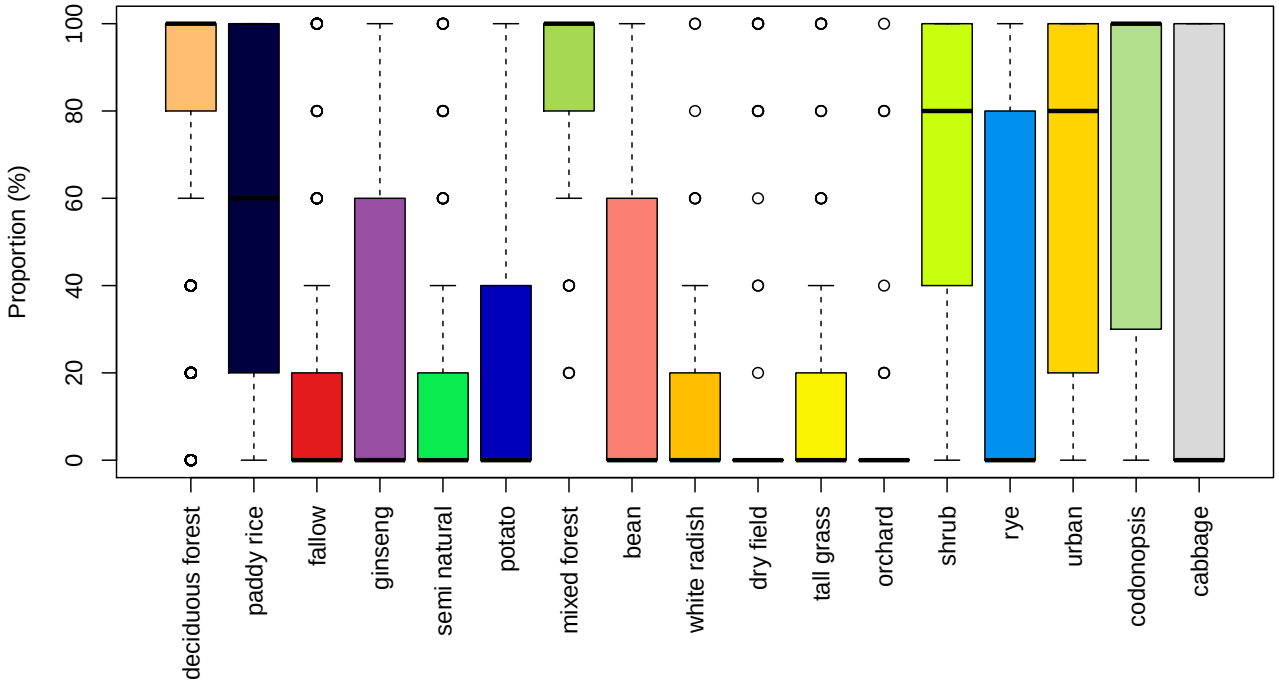

Fig. 4.17 Proportion of five nearest neighbours of the test data in the training data that belong to the same class as the test data in scenario S3.

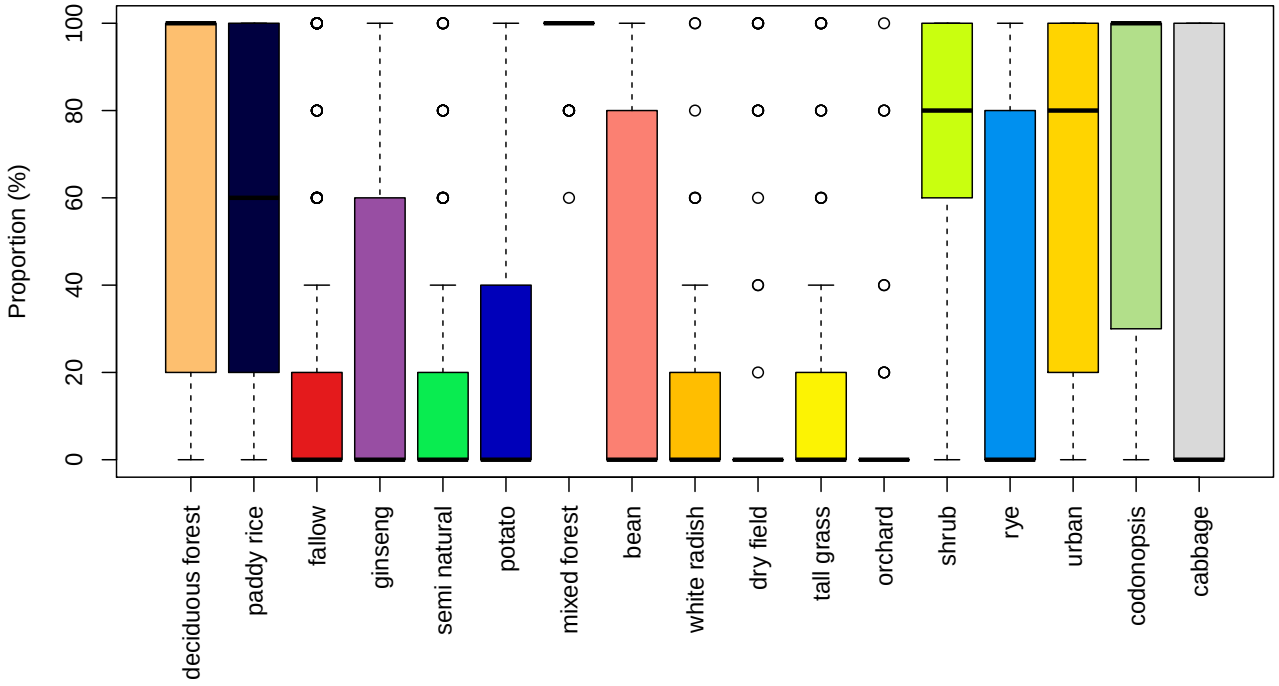

Fig. 4.18 Proportion of five nearest neighbours of the test data in the training data that belong to the same class as the test data in scenario S4

.

Supplementary Tables

Table 4.2. Modification of the LULC classification scheme by Seo et al. (2014). Similar minority classes were merged reducing the number of classes from 67 to 59.

Original class	Aggregated class
dry field, mixed dry field	dry field
barren, bare soil	bare soil
chinese cabbage, european cabbage, cabbage	cabbage
apple, peach, grape, orchard	orchard
coniferous forest, pine forest	coniferous forest

Table 4.3. The average oversampling rate N in the training data of the SMOTEd scenarios (S3 and S4) in 5 repetitions.

Classes	N (%)
deciduous forest	-
paddy rice	-
fallow	100
ginseng	300
semi natural	300
potato	300
mixed forest	500
bean	600
white radish	700
dry field	933
tall grass	1017
orchard	1017
shrub	1017
rye	1333
urban	1500
codonopsis	1967
cabbage	2300

Table 4.4. ROC summary of 5 repetitions in scenario S1 using RF.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	96.40	96.40	96.50	96.66	96.90	97.10	28.30	28.70	28.70	28.70	28.70	29.10
paddy rice	87.20	87.20	88.50	88.26	88.50	89.90	5.30	5.40	5.70	5.70	5.90	6.20
fallow	29.00	35.50	35.50	36.46	37.10	45.20	2.40	2.70	2.70	2.72	2.90	2.90
ginseng	41.20	44.10	44.10	46.46	50.00	52.90	1.10	1.10	1.20	1.22	1.20	1.50
semi natural	3.10	6.20	6.20	7.48	9.40	12.50	0.90	0.90	1.00	1.00	1.10	1.10
potato	28.10	28.10	37.50	35.62	40.60	43.80	1.10	1.10	1.10	1.26	1.20	1.80
mixed forest	40.90	45.50	45.50	44.58	45.50	45.50	0.00	0.20	0.30	0.28	0.30	0.60
bean	0.00	0.00	5.00	3.00	5.00	5.00	0.30	0.40	0.40	0.42	0.50	0.50
white radish	0.00	5.60	5.60	5.58	5.60	11.10	0.10	0.20	0.20	0.22	0.30	0.30
dry field	0.00	0.00	0.00	1.42	0.00	7.10	0.00	0.00	0.00	0.00	0.00	0.00
tall grass	0.00	0.00	0.00	3.08	7.70	7.70	0.00	0.00	0.00	0.06	0.10	0.20
orchard	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.06	0.10	0.10
shrub	23.10	23.10	23.10	23.10	23.10	23.10	0.00	0.00	0.00	0.00	0.00	0.00
rye	20.00	30.00	30.00	32.00	40.00	40.00	0.20	0.20	0.20	0.24	0.30	0.30
urban	0.00	0.00	0.00	2.22	0.00	11.10	0.00	0.00	0.00	0.00	0.00	0.00
codonopsis	14.30	14.30	14.30	25.74	42.90	42.90	0.10	0.10	0.10	0.14	0.20	0.20
cabbage	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.5. ROC summary of 5 repetitions in scenario S2 using RF.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	93.50	93.60	94.20	94.06	94.40	94.60	18.80	19.40	19.40	19.36	19.60	19.60
paddy rice	90.50	90.50	91.90	91.48	91.90	92.60	6.30	6.60	7.30	7.08	7.50	7.70
fallow	40.30	41.90	45.20	44.84	46.80	50.00	3.60	3.80	4.10	4.08	4.30	4.60
ginseng	35.30	41.20	50.00	45.30	50.00	50.00	1.10	1.20	1.60	1.44	1.60	1.70
semi natural	6.20	6.20	12.50	9.98	12.50	12.50	1.00	1.10	1.20	1.16	1.20	1.30
potato	34.40	34.40	34.40	36.26	37.50	40.60	1.50	1.50	1.70	1.70	1.80	2.00
mixed forest	45.50	45.50	45.50	47.30	50.00	50.00	0.30	0.30	0.30	0.38	0.50	0.50
bean	0.00	5.00	15.00	10.00	15.00	15.00	0.40	0.40	0.50	0.52	0.60	0.70
white radish	11.10	11.10	11.10	11.10	11.10	11.10	0.10	0.20	0.30	0.24	0.30	0.30
dry field	0.00	0.00	0.00	1.42	0.00	7.10	0.00	0.00	0.00	0.00	0.00	0.00
tall grass	0.00	0.00	0.00	1.54	0.00	7.70	0.00	0.00	0.10	0.06	0.10	0.10
orchard	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.20	0.16	0.20	0.20
shrub	23.10	23.10	30.80	29.26	30.80	38.50	0.00	0.00	0.00	0.00	0.00	0.00
rye	40.00	40.00	40.00	40.00	40.00	40.00	0.20	0.30	0.30	0.32	0.40	0.40
urban	0.00	0.00	0.00	2.22	0.00	11.10	0.00	0.00	0.00	0.00	0.00	0.00
codonopsis	28.60	28.60	28.60	37.16	28.60	71.40	0.10	0.20	0.20	0.22	0.30	0.30
cabbage	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.6. ROC summary of 5 repetitions in scenario S3 using RF.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	90.40	90.70	90.70	90.84	91.00	91.40	10.80	10.80	11.90	11.60	12.10	12.40
paddy rice	77.70	80.40	81.80	81.36	82.40	84.50	3.40	4.30	4.40	4.20	4.40	4.50
fallow	30.60	33.90	37.10	35.48	37.10	38.70	1.80	2.20	2.20	2.36	2.70	2.90
ginseng	41.20	41.20	44.10	47.06	52.90	55.90	1.30	1.60	1.60	1.60	1.70	1.80
semi natural	3.10	6.20	9.40	9.36	12.50	15.60	2.40	2.50	2.50	2.68	2.70	3.30
potato	34.40	37.50	40.60	41.26	43.80	50.00	2.00	2.00	2.10	2.16	2.20	2.50
mixed forest	63.60	63.60	68.20	70.90	72.70	86.40	0.60	0.70	0.80	0.76	0.80	0.90
bean	10.00	15.00	20.00	19.00	25.00	25.00	1.10	1.40	1.60	1.52	1.70	1.80
white radish	11.10	11.10	16.70	15.56	16.70	22.20	0.60	0.90	1.00	0.92	1.00	1.10
dry field	0.00	0.00	0.00	1.42	0.00	7.10	0.70	0.80	0.80	0.82	0.90	0.90
tall grass	7.70	15.40	15.40	13.86	15.40	15.40	1.10	1.10	1.20	1.28	1.50	1.50
orchard	7.70	7.70	7.70	10.78	15.40	15.40	0.80	1.00	1.20	1.10	1.20	1.30
shrub	46.20	53.80	61.50	58.44	61.50	69.20	0.30	0.40	0.40	0.44	0.50	0.60
rye	50.00	50.00	50.00	50.00	50.00	50.00	0.90	1.00	1.10	1.08	1.10	1.30
urban	33.30	33.30	33.30	39.98	44.40	55.60	0.10	0.30	0.40	0.32	0.40	0.40
codonopsis	57.10	71.40	71.40	68.54	71.40	71.40	0.20	0.30	0.30	0.28	0.30	0.30
cabbage	16.70	16.70	16.70	16.70	16.70	16.70	0.20	0.30	0.30	0.28	0.30	0.30

Table 4.7. ROC summary of 5 repetitions in scenario S4 using RF.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	81.40	81.80	81.80	81.90	81.90	82.60	2.40	3.10	3.30	3.32	3.80	4.00
paddy rice	80.40	81.80	81.80	82.44	83.10	85.10	4.80	5.20	5.30	5.28	5.50	5.60
fallow	40.30	41.90	45.20	45.16	48.40	50.00	3.20	3.50	3.90	3.78	4.00	4.30
ginseng	41.20	44.10	47.10	48.24	52.90	55.90	1.70	1.80	1.90	1.92	2.10	2.10
semi natural	6.20	9.40	12.50	11.24	12.50	15.60	2.60	2.60	2.80	2.84	3.10	3.10
potato	37.50	37.50	40.60	41.88	43.80	50.00	2.10	2.30	2.40	2.38	2.50	2.60
mixed forest	90.90	95.50	95.50	94.58	95.50	95.50	2.00	2.20	2.20	2.20	2.30	2.30
bean	10.00	15.00	15.00	17.00	20.00	25.00	1.00	1.40	1.40	1.44	1.60	1.80
white radish	11.10	16.70	16.70	15.58	16.70	16.70	1.00	1.00	1.10	1.18	1.20	1.60
dry field	0.00	0.00	0.00	1.42	0.00	7.10	1.30	1.30	1.40	1.52	1.70	1.90
tall grass	23.10	23.10	23.10	23.10	23.10	23.10	2.20	2.20	2.70	2.50	2.70	2.70
orchard	7.70	7.70	7.70	9.24	7.70	15.40	1.10	1.10	1.40	1.28	1.40	1.40
shrub	53.80	53.80	69.20	64.58	69.20	76.90	0.70	0.90	0.90	0.92	1.00	1.10
rye	40.00	50.00	50.00	48.00	50.00	50.00	0.90	1.10	1.20	1.12	1.20	1.20
urban	33.30	33.30	44.40	44.44	55.60	55.60	0.30	0.30	0.30	0.36	0.40	0.50
codonopsis	57.10	71.40	71.40	68.54	71.40	71.40	0.20	0.30	0.30	0.30	0.30	0.40
cabbage	16.70	16.70	16.70	23.34	33.30	33.30	0.30	0.30	0.30	0.36	0.40	0.50

Table 4.8. ROC summary of 5 repetitions in scenario S1 using SVM.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	94.30	95.10	95.10	95.18	95.70	95.70	24.30	25.60	25.60	25.52	25.60	26.50
paddy rice	84.50	85.10	85.10	85.26	85.80	85.80	4.80	4.90	5.20	5.12	5.20	5.50
fallow	29.00	32.30	35.50	34.52	37.10	38.70	3.00	3.00	3.10	3.10	3.20	3.20
ginseng	41.20	44.10	47.10	45.90	47.10	50.00	0.60	1.00	1.10	1.00	1.10	1.20
semi natural	12.50	12.50	12.50	13.74	15.60	15.60	1.20	1.40	1.40	1.48	1.60	1.80
potato	28.10	31.20	31.20	30.58	31.20	31.20	0.70	0.80	0.80	0.92	1.10	1.20
mixed forest	54.50	59.10	63.60	62.72	68.20	68.20	0.20	0.30	0.40	0.34	0.40	0.40
bean	15.00	15.00	20.00	18.00	20.00	20.00	0.70	0.80	0.90	0.88	1.00	1.00
white radish	0.00	0.00	5.60	3.36	5.60	5.60	0.40	0.60	0.60	0.66	0.80	0.90
dry field	0.00	0.00	7.10	4.26	7.10	7.10	0.00	0.10	0.10	0.12	0.20	0.20
tall grass	0.00	0.00	7.70	4.62	7.70	7.70	0.20	0.30	0.40	0.38	0.50	0.50
orchard	0.00	0.00	0.00	1.54	0.00	7.70	0.20	0.20	0.30	0.26	0.30	0.30
shrub	23.10	23.10	30.80	29.26	30.80	38.50	0.00	0.00	0.00	0.04	0.10	0.10
rye	10.00	20.00	30.00	28.00	40.00	40.00	0.50	0.60	0.60	0.66	0.70	0.90
urban	11.10	11.10	11.10	17.76	22.20	33.30	0.00	0.00	0.00	0.00	0.00	0.00
codonopsis	28.60	28.60	28.60	34.32	42.90	42.90	0.10	0.10	0.10	0.12	0.10	0.20
cabbage	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.10

Table 4.9. ROC summary of 5 repetitions in scenario S2 using SVM.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	91.90	91.90	92.40	92.50	92.80	93.50	14.80	15.70	16.10	16.08	16.60	17.20
paddy rice	87.20	87.20	87.80	88.12	88.50	89.90	6.20	6.30	6.60	6.60	6.90	7.00
fallow	37.10	38.70	45.20	42.28	45.20	45.20	3.60	4.00	4.70	4.44	4.90	5.00
ginseng	41.20	44.10	44.10	46.46	50.00	52.90	1.00	1.10	1.40	1.26	1.40	1.40
semi natural	9.40	12.50	15.60	13.74	15.60	15.60	1.50	1.70	1.90	1.84	1.90	2.20
potato	28.10	31.20	34.40	33.74	37.50	37.50	1.10	1.20	1.40	1.40	1.60	1.70
mixed forest	54.50	59.10	59.10	60.90	63.60	68.20	0.50	0.50	0.60	0.64	0.80	0.80
bean	15.00	15.00	20.00	20.00	25.00	25.00	0.70	0.80	0.90	0.88	0.90	1.10
white radish	0.00	5.60	5.60	4.48	5.60	5.60	0.50	0.60	0.60	0.70	0.80	1.00
dry field	0.00	0.00	7.10	5.70	7.10	14.30	0.00	0.00	0.10	0.08	0.10	0.20
tall grass	0.00	0.00	7.70	6.16	7.70	15.40	0.10	0.30	0.30	0.32	0.40	0.50
orchard	0.00	0.00	0.00	3.08	7.70	7.70	0.30	0.30	0.40	0.42	0.50	0.60
shrub	23.10	23.10	23.10	24.64	23.10	30.80	0.00	0.10	0.10	0.12	0.10	0.30
rye	10.00	20.00	40.00	30.00	40.00	40.00	0.40	0.60	0.90	0.88	1.20	1.30
urban	0.00	0.00	11.10	13.32	22.20	33.30	0.00	0.00	0.00	0.04	0.00	0.20
codonopsis	0.00	14.30	28.60	20.02	28.60	28.60	0.10	0.10	0.20	0.16	0.20	0.20
cabbage	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.10

Table 4.10. ROC summary of 5 repetitions in scenario S3 using SVM.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	85.80	86.20	86.40	86.30	86.50	86.60	7.70	8.20	8.20	8.66	8.60	10.60
paddy rice	64.90	67.60	69.60	68.52	69.60	70.90	1.60	2.00	2.40	2.28	2.70	2.70
fallow	4.80	8.10	12.90	11.28	14.50	16.10	0.20	0.40	0.80	0.70	0.80	1.30
ginseng	38.20	38.20	44.10	44.70	47.10	55.90	1.50	1.60	1.70	1.82	2.10	2.20
semi natural	6.20	6.20	9.40	8.12	9.40	9.40	1.80	2.20	2.30	2.20	2.30	2.40
potato	18.80	18.80	25.00	23.76	25.00	31.20	1.10	1.10	1.40	1.38	1.60	1.70
mixed forest	81.80	81.80	86.40	85.46	86.40	90.90	1.70	1.70	1.80	1.78	1.80	1.90
bean	15.00	20.00	25.00	23.00	25.00	30.00	1.10	1.50	1.60	1.58	1.70	2.00
white radish	11.10	16.70	16.70	16.68	16.70	22.20	2.40	2.50	2.70	2.66	2.80	2.90
dry field	0.00	0.00	7.10	5.70	7.10	14.30	2.70	3.10	3.20	3.22	3.40	3.70
tall grass	15.40	23.10	23.10	24.64	30.80	30.80	1.60	1.60	1.60	1.72	1.70	2.10
orchard	0.00	7.70	7.70	9.24	7.70	23.10	1.20	1.50	1.50	1.46	1.50	1.60
shrub	30.80	38.50	46.20	41.58	46.20	46.20	0.30	0.30	0.30	0.40	0.40	0.70
rye	40.00	50.00	50.00	52.00	50.00	70.00	2.00	2.40	2.60	2.70	3.20	3.30
urban	55.60	66.70	66.70	64.48	66.70	66.70	1.00	1.40	1.40	1.46	1.60	1.90
codonopsis	71.40	71.40	71.40	74.26	71.40	85.70	0.60	0.90	0.90	0.86	0.90	1.00
cabbage	16.70	16.70	16.70	23.34	33.30	33.30	1.90	2.00	2.10	2.18	2.40	2.50

Table 4.11. ROC summary of 5 repetitions in scenario S4.

	True Positive Rate (%)						False Positive Rate (%)					
Classes	Min	25% Q	Median	Mean	75% Q	Max	Min	25% Q	Median	Mean	75% Q	Max
deciduous forest	76.60	77.30	78.00	77.78	78.40	78.60	2.40	2.60	3.10	3.22	3.80	4.20
paddy rice	66.20	66.20	67.60	67.84	68.90	70.30	1.60	1.80	2.50	2.32	2.80	2.90
fallow	19.40	25.80	25.80	24.52	25.80	25.80	2.00	2.40	2.70	2.66	3.00	3.20
ginseng	41.20	41.20	47.10	45.90	47.10	52.90	1.60	1.80	1.90	2.02	2.30	2.50
semi natural	6.20	6.20	12.50	10.60	12.50	15.60	3.30	3.30	3.60	3.60	3.90	3.90
potato	28.10	31.20	34.40	34.38	34.40	43.80	1.20	1.40	1.60	1.56	1.80	1.80
mixed forest	95.50	100.00	100.00	99.10	100.00	100.00	3.30	3.60	3.60	3.60	3.70	3.80
bean	20.00	25.00	25.00	24.00	25.00	25.00	1.30	1.80	2.10	1.94	2.20	2.30
white radish	5.60	5.60	16.70	12.26	16.70	16.70	2.00	2.00	2.20	2.16	2.30	2.30
dry field	0.00	7.10	7.10	7.12	7.10	14.30	2.50	3.10	3.70	3.46	3.80	4.20
tall grass	23.10	23.10	23.10	26.18	30.80	30.80	2.20	2.70	2.70	2.70	2.80	3.10
orchard	7.70	15.40	15.40	15.40	15.40	23.10	1.20	1.50	1.70	1.64	1.80	2.00
shrub	38.50	46.20	53.80	49.22	53.80	53.80	0.80	0.90	1.10	1.04	1.10	1.30
rye	40.00	40.00	50.00	48.00	50.00	60.00	2.40	2.60	2.60	2.64	2.70	2.90
urban	44.40	55.60	66.70	60.02	66.70	66.70	0.80	0.90	1.00	1.02	1.20	1.20
codonopsis	57.10	71.40	71.40	68.54	71.40	71.40	0.30	0.50	0.70	0.62	0.80	0.80
cabbage	0.00	16.70	16.70	20.00	33.30	33.30	1.20	1.50	1.70	1.58	1.70	1.80

Table 4.12. *F-score* in 5 repetitions of scenarios S1 through S4.

Classifier	Scenarios	Min	25% Q	Median	Mean	75% Q	Max
RF	S1	0.35	0.37	0.37	0.37	0.38	0.38
	S2	0.38	0.38	0.39	0.39	0.40	0.40
	S3	0.38	0.38	0.38	0.39	0.40	0.41
	S4	0.37	0.37	0.38	0.38	0.39	0.41
SVM	S1	0.34	0.36	0.38	0.38	0.40	0.40
	S2	0.33	0.33	0.37	0.36	0.38	0.40
	S3	0.34	0.35	0.36	0.35	0.36	0.36
	S4	0.33	0.35	0.35	0.35	0.35	0.35

Table 4.13. *NID* in 5 repetitions of scenarios S1 through S4.

Classifier	Scenarios	Min	25% Q	Median	Mean	75% Q	Max
RF	S1	0.65	0.65	0.66	0.66	0.66	0.66
	S2	0.61	0.62	0.62	0.62	0.62	0.63
	S3	0.56	0.56	0.56	0.56	0.56	0.56
	S4	0.60	0.60	0.60	0.60	0.60	0.60
SVM	S1	0.62	0.63	0.64	0.63	0.64	0.64
	S2	0.60	0.61	0.62	0.61	0.62	0.62
	S3	0.62	0.62	0.62	0.62	0.63	0.63
	S4	0.64	0.64	0.65	0.65	0.66	0.66

Table 4.14. *G-mean* in 5 repetitions of scenarios S1 through S4.

Classifier	Scenarios	Min	25% Q	Median	Mean	75% Q	Max
RF	S1	0.00	0.00	0.00	0.00	0.00	0.00
	S2	0.00	0.00	0.00	0.00	0.00	0.00
	S3	0.00	0.00	0.00	0.06	0.00	0.29
	S4	0.00	0.00	0.00	0.07	0.00	0.33
SVM	S1	0.00	0.00	0.00	0.00	0.00	0.00
	S2	0.00	0.00	0.00	0.00	0.00	0.00
	S3	0.00	0.00	0.00	0.12	0.28	0.30
	S4	0.00	0.00	0.32	0.19	0.32	0.33

Table 4.15. Precision in 5 repetitions of scenarios S1 through S4.

Classifier	Scenarios	Min	25% Q	Median	Mean	75% Q	Max
RF	S1	0.57	0.57	0.62	0.62	0.66	0.67
	S2	0.55	0.55	0.58	0.58	0.59	0.63
	S3	0.37	0.37	0.38	0.38	0.40	0.41
	S4	0.33	0.34	0.34	0.35	0.36	0.37
SVM	S1	0.47	0.47	0.50	0.50	0.52	0.57
	S2	0.41	0.42	0.46	0.47	0.48	0.55
	S3	0.30	0.32	0.32	0.32	0.32	0.34
	S4	0.30	0.30	0.30	0.30	0.30	0.30

Table 4.16. Recall in 5 repetitions of scenarios S1 through S4.

Classifier	Scenarios	Min	25% Q	Median	Mean	75% Q	Max
RF	S1	0.26	0.26	0.27	0.27	0.27	0.27
	S2	0.28	0.29	0.29	0.30	0.30	0.31
	S3	0.38	0.38	0.39	0.39	0.40	0.42
	S4	0.41	0.41	0.42	0.42	0.42	0.46
SVM	S1	0.27	0.30	0.30	0.30	0.31	0.32
	S2	0.27	0.28	0.30	0.30	0.31	0.32
	S3	0.38	0.38	0.39	0.39	0.40	0.40
	S4	0.38	0.41	0.41	0.41	0.42	0.42

Table 4.17. Evaluation of the maps with the largest *F-score* in scenarios S1 through S4.

	RF				SVM			
	S1	S2	S3	S4	S1	S2	S3	S4
accuracy	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.96
G-mean	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.32
precision	0.67	0.63	0.40	0.37	0.57	0.55	0.32	0.30
recall	0.27	0.30	0.42	0.46	0.31	0.31	0.41	0.42
F-score	0.38	0.40	0.41	0.41	0.40	0.40	0.36	0.35
NID	0.66	0.62	0.56	0.60	0.63	0.61	0.62	0.65

Table 4.18. Changes of the median *TPR*s and *FPR*s in S1 through S4.

LULC	Median *TPR* (RF)			Median *TPR* (SVM)			Median *FPR* (RF)			Median *FPR* (SVM)		
	S1 to S2	S2 to S3	S3 to S4	S1 to S2	S2 to S3	S3 to S4	S1 to S2	S2 to S3	S3 to S4	S1 to S2	S2 to S3	S3 to S4
deciduous forest	-2.30	-3.50	-8.90	-2.70	-6.00	-8.40	-9.30	-7.50	-8.60	-9.50	-7.90	-5.10
paddy rice	3.40	-10.10	0.00	2.70	-18.20	-2.00	1.60	-2.90	0.90	1.40	-4.20	0.10
fallow	9.70	-8.10	8.10	9.70	-32.30	12.90	1.40	-1.90	1.70	1.60	-3.90	1.90
ginseng	5.90	-5.90	3.00	-3.00	0.00	3.00	0.40	0.00	0.30	0.30	0.30	0.20
semi natural	6.30	-3.10	3.10	3.10	-6.20	3.10	0.20	1.30	0.30	0.50	0.40	1.30
potato	-3.10	6.20	0.00	3.20	-9.40	9.40	0.60	0.40	0.30	0.60	0.00	0.20
mixed forest	0.00	22.70	27.30	-4.50	27.30	13.60	0.00	0.50	1.40	0.20	1.20	1.80
bean	10.00	5.00	-5.00	0.00	5.00	0.00	0.10	1.10	-0.20	0.00	0.70	0.50
white radish	5.50	5.60	0.00	0.00	11.10	0.00	0.10	0.70	0.10	0.00	2.10	-0.50
shrub	7.70	30.70	7.70	-7.70	23.10	7.60	0.00	0.40	0.50	0.10	0.20	0.80
rye	10.00	10.00	0.00	10.00	10.00	0.00	0.10	0.80	0.10	0.30	1.70	0.00
codonopsis	14.30	42.80	0.00	0.00	42.80	0.00	0.10	0.10	0.00	0.10	0.70	-0.20
dry field	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.60	0.00	3.10	0.50
tall grass	0.00	15.40	7.70	0.00	15.40	0.00	0.10	1.10	1.50	-0.10	1.30	1.10
orchard	0.00	7.70	0.00	0.00	7.70	7.70	0.10	1.00	0.20	0.10	1.10	0.20
urban	0.00	33.30	11.10	0.00	55.60	0.00	0.00	0.40	-0.10	0.00	1.40	-0.40
cabbage	0.00	16.70	0.00	0.00	16.70	0.00	0.00	0.30	0.00	0.00	2.10	-0.40

Chapter 5

Synopsis

In this dissertation, improved quantification of Land Use and Land Cover (LULC) in complex agricultural landscapes is explored. Specifically, extraction of spatially and thematically detailed LULC information from existing, spatially coarse, multi-spectral satellite products is pursued through three studies: high-quality ground LULC data collection (Chapter 2), derivation of fractional cover, i.e. continuous LULC representation (Chapter 3), and multi-crop LULC classification (Chapter 4).

In chapter 2, the high-quality LULC census data was introduced. The LULC data provide a comprehensive characterisation of the study site, and have been utilised in a wide range of projects (e.g. Nguyen et al., 2012; Poppenborg et al., 2013; Reineking et al., 2013; Shope et al., 2014). The comparison of the observed data with the MODIS land cover product (GLC) revealed the limitations of this GLC product in the cultivated landscape. In chapter 3, a fractional LULC model was developed and it presented an attractive way to spatially improve GLC databases based on existing satellite products. In chapter 4, a multi-crop LULC classification model was developed to thematically enrich LULC representation. By means of a data rebalancing technique, the problem of the LULC data imbalance was addressed. Artificial balancing of the training data substantially increased the classification performance of some minority, i.e. rare, LULC types.

Overall, the presented dissertation contributes to the field of LULC quantification and provides an assessment of methods that can help in improving LULC quantification in complex heterogeneous cultivated landscapes. In the following pages, key results and finding of the dissertation are summarised. Afterwards, a general discussion on the value of the dissertation, the future

outlook, and concluding remarks will be given.

5.1 Summary

Deriving a per-field land use and land cover map in an agricultural mosaic catchment (Chapter 2)

The Haean Land Use and Land Cover (LULC) observation data is a per-field vector GIS data of the Haean catchment, South Korea. The area is characterised by intensive agriculture. The information provided here includes the land use and land cover census data (67 LULC types) in a polygon shape file for three years (2009 – 2011). The raw data was complied regarding Land Cover Classification System (LCCS) developed by Food and Agricultural Organizations of the United Nations (FAO) and United Nations Environment Programme (UNEP) (Cord et al., 2010; Di Gregorio, 2005). It is an elaborated land cover description scheme and enables us to describe LULC consistently. Data quality information is included in the dataset following the scheme. The resulting dataset includes detailed crop type information in a consistent and complete manner. Additionally, the original data was reclassified into a number of classification systems for compatibility with the land cover type definitions of various GLC products.

A total of 67 LULC types were identified. During the three-year study period, Haean catchment underwent discernible changes of LULC both at the field level (displacement) and at the landscape level (composition change). A number of rice paddies and dry fields were converted to alternative perennial crops such as “ginseng” and “orchards” in response to subsidies and other policy measures. The alternative crops essentially replaced annual crops. It may be concluded that the governmental policy was successful in introducing more environmentally friendly perennial crops thus possibly preventing soil erosion. Field-level changes are also partially due to crop rotation, which is common for the annual crops in the region. Note that these displacements will be reflected neither in static (i.e. unchanging over time) GLC products nor in local statistics.

The census data was compared with the MODIS Land Cover Type product (MCD12Q1). The Cohen’s κ between the rasterised ground truth and the MODIS land cover was which is fair but not substantial (avg. κ=0.41 for 2009 and 2010). The MODIS product failed to capture the drastic changes presumably due to its low spatial and thematic resolution. For example, minor crops such as “Potato” were not captured as they were not defined in the classification scheme. Linear elements such as “Water Bodies” and “Urban and Built-Up Lands” were completely

missing in the MODIS product due to the pixel size (500 m). This phenomenon becomes more problematic for land cover types smaller than the MODIS pixel in its typical dimension. The comparison revealed that the limitation of GLC products such as lack of irrigated fields and vaguely defined cropland types could cause substantial misrepresentation of LULC in future applications (e.g. Ecosystem Services research) particularly for complex agricultural landscapes.

Mapping fractional land use and land cover in a monsoon region: the effects of data processing options (Chapter 3)

To estimate multi-type fractional cover, a Random Forest (RF) regression model was developed using globally available multi-spectral satellite products. The main objective was to evaluate different data-processing options regarding a fractional cover LULC regression problem. The efficacy of various spectral predictors, smoothing filters, and data interval options were evaluated based on its impact on the regression performance. For rigorous evaluation of the data-processing options, a spatial cross-validation scheme was adopted. Additionally, relative importance of the spectral bands and the data acquisition dates were estimated using a RF variable importance metric.

The regression model reproduced spatial distributions of the LULC fractions. However, predicting absolute fractions remained difficult especially for the minor types. The model performance differed between types due to the distributions of the observed fraction data. It is primarily due to the imbalanced nature of the LULC data (i.e. unequal distributions of LULC types) as the RF regression model is biased towards the major types. With elaborations such as the use of the Hurdle formulation or the use of data-balancing techniques, the model performance may be improved to a certain degree.

The analysis of the relative importance of input features revealed that the monsoon period was not the most influential period on the regression performance. Moreover, the most influential periods varied by LULC type. Therefore, the use of full time series is recommended in future applications. Smoothing by the Savitzky-Golay filter was disadvantageous. It suggests that the original MODIS maximum value composite algorithm already sufficiently suppresses noise. The surface reflectance bands B1 and B2 were important in modelling fractional cover. In contrast, the bands B3 and B7 were rather uninformative, especially for the minor types.

Improving the classification of rare land use and land cover types using synthetic data (Chapter 4)

The classification of the imbalanced LULC data set was challenging due to both absolute and relative rarities as well as the class overlap. The synthetic oversampling method alleviated the issue of class imbalance by increasing the number of minor type data points. Balancing the data by the synthetic minority oversampling technique (SMOTE) increased the true positive rate of some monitory LULC types substantially, however some other minority types remained difficult to classify.

The mechanism by which data resampling affected model performance was analysed by looking at the relationship between LULC type labels and surface reflectance (mutual information) and the difficulty of classification (entropy). The low classification performance on some minority types was attributed to a substantial class overlap (i.e. different classes contain comparable data points in the same region of a feature space) already present in the original data set.

Support vector machine (SVM) outperformed RF when trained on the original unbalanced data set, whereas RF performed marginally better than SVM when trained on the synthetically oversampled data. RF produced maps that agreed slightly better with reference LULC data. However, the difference in performance between the classifiers was small.

5.2 Prospective applications

To generalise the lessons for LULC modelling, further studies covering different types of landscapes and different spatial and temporal scales using the presented frameworks would be necessary. However, in the course of this work, a number of interesting applications have emerged that might profit from the methodological advances made in the dissertation.

The LULC census data presented in chapter 2 can be useful in developing/validating a high-resolution LULC product for complex agricultural landscapes as it is a vector-form data with fine spatial information. As it is described by a consistent and complete scheme 'FAO-LCCS', it can be translated in any simpler scheme (e.g. IGBP-Discover), thus used to evaluate a variety of LULC products such as MODIS land cover or GlobCover. For instance, we are planning to use the census data to train a LULC model for the larger surrounding area, namely Soyang watershed.

The Haean catchment has been studied intensively as it undergoes a conflict between agriculture and environmental protection (Nguyen et al., 2012; Poppenborg et al., 2013). The LULC data

can be also used as input for regional environmental modelling (e.g. Reineking et al., 2013; Shope et al., 2014; Zhao et al., 2012), ecosystem services and decision making analysis (e.g. Poppenborg et al., 2013). The LULC data can be combined with the regional economic statistics to be used in economic and other social science research (Nguyen et al., 2014).

By using the fractional LULC framework described in chapter 3, further studies covering different types of landscapes (e.g. Eurasian Steppe) can produce interesting outcomes. Several studies have shown examples of deriving fractional land cover information (Asner et al., 2000; Defries et al., 2000; DeFries et al., 1995; Fernandes et al., 2004; Foody et al., 1996; Guerschman et al., 2009; Lu et al., 2003; Schwarz et al., 2005; Schwieder et al., 2014) but with a small number of LULC types (e.g. few green vegetation types). Fractional cover estimation of multi-crop LULC would be a very useful application of the presented framework. The developed fractional cover model and multi-crop LULC model can be applied to the past-time data to trace back historic LULC records. In such an area like dry land regions of the Northeast Asia, which has suffered from desertification due to both climate change and livestock pastoralism (Narantsetseg et al., 2014), back-tracing of continuous LULC trend would be a very interesting work. As there has been MODIS spectral data for the last 15 years, there is a possibility to trace back fractional LULC for the period (e.g. Wu et al., 2014) . Certainly, securing high-quality ground truth data would be an important prerequisite task.

It is an essential problem for almost all modelling studies to 'objectively' determine modelling parameters and select data-processing options. The overall modelling frameworks established in chapter 3 and 4 used sound and rigorous evaluation procedures (e.g. spatial cross-validation). These can be a useful guide for other modelling tasks in choosing optimal modelling parameters as well as data-processing options.

When monitoring agricultural ecosystems in cultivated landscapes, minority LULC types can be even more important than majority LULC types; they might indicate a beginning of LULC change or impact of climate change and land policy decisions. To cope with the inherent data imbalance in LULC data, synthetic resampling techniques presented in chapter 4 can be useful. It can be conveniently used to obtain balanced data for training learning algorithms, hence improve learning performance. For example, the MODIS land cover is trained by the System for Terrestrial Ecosystem Parameterization (STEP) database in which training data is imbalanced in land cover type. Data rebalancing of the STEP database data may produce better quality LULC data for the globe.

5.2.1 Research outlook

During the course of this dissertation, several opportunities based on the results of the dissertation have been identified. In general, the value of the dissertation will be increased by continuing efforts to relate this research with ecosystem services and decision-making analysis in additional case study areas. Extension of the presented approach toward a larger study area would be an important step to consider in future research.
In particular, the following points constitute attractive further research directions.

5.2.1.1 Standardised acquisition of high-quality LULC data

Developing a strategy to massively acquire high-quality LULC data would be an important step to consider. Training data for GLC products are generally not very extensive. For example, the collection 5 MODIS land cover products are trained using STEP version 6 database which includes 2095 training sites distributed across the terrestrial part of the Earth (Friedl et al., 2010; Sulla-Menashe et al., 2011). For agricultural cover types, limitation of training data sites is noteworthy as it includes small number of sites for the whole cultivated zones. Precisely, the version 6 STEP database includes 499 pixels for cultivated zones (i.e. $>$ 60% agriculture) and 119 pixels for agricultural mosaics (i.e. 30-60% agriculture) for the whole globe. For the sites, specific crop type information is missing but five broadly defined crop type classes are recorded – namely cereal crop, broadleaf crop, mixed crop, rice, and orchards/vineyards. This in total bring a problem of training data for specific crop types. Moreover,Foody (2007) noted that errors in ground data sets used in remote sensing of land cover may be large. These types of limitations affect the thematic quality (i.e. simplified agricultural LULC types) in most of the GLC databases.
The geographic scale of the LULC quantification is beyond individual studies (Chen et al., 2014) as any direct survey on LULC is limited in its spatial coverage. Thus, for acquiring high-quality LULC for a wider area, it is necessary to develop a data collection platform for LULC data. For bird-watching data, there are operational networks of data collection (e.g. eBird http://ebird.org/content/ebird/) which engage a large number of people reporting locally observed bird data. These data collection networks with open and free access to data have successfully contributed to many large-scale biodiversity research programs (Chen et al., 2014; Turner et al., 2015). When it comes to spatial data, the Geo-Wiki (http://geo-wiki.org) is an active data-sharing platform available to the public (Fritz et al., 2011).
Automatic LULC data retrieval from street view images of collaborative mapping platforms

(e.g. OpenStreetMap https://www.openstreetmap.org/ can be another source of data at a wider spatial scale. Machine learning and artificial intelligence research groups (e.g. Clarifai http://www.clarifai.com) have been analysing publicised images for various purposes such as automatic object recognition or topic analysis.

To routinely acquire high-quality LULC data, it would be very useful to collect published data from online data sharing platforms and produce LULC data sets with proper descriptions. The resulting datasets can be valuable for many LULC related applications.

5.2.1.2 Application of the adopted methods to larger areas

Transferability of the presented data and the LULC quantification models is an important issue to be addressed. As the target area was a small catchment (64 km^2), the results of this dissertation may not be sufficient to fully justify the applicability of the methods (i.e. SMOTE) over other (larger) areas. For instance, given a LULC classification problem from spectral data, the within class heterogeneity of a LULC class (i.e. spectral and temporal variations of the instances of the LULC class) can be very large. In such cases, standard classifiers might fail to identify the LULC classes located outside of the target sites. Because standard classifiers generally search for a specific pattern of features for a class, they may fail to detect classes from noisy and/or transformed (but informative) feature data sets. Due to spatio-temporal variations such as in climate variables, this transform in feature data prevails (e.g. varying phenology). This also applies to the case of a multi-year classification where the classifiers trained in a specific year fail to recognise the LULC classes for a different test year.

To address this issue, first, additional LULC survey campaigns on different areas and on various other LULC systems should be implemented. Second, it may be useful to develop an approach which can learn from such a case. Recently, scale- and transform-invariant learning techniques are being introduced to the remote sensing communities (Jones et al., 2014). For example, feature extraction using scale invariant feature transform (SIFT) could resolve some issues related to such heterogeneities by capturing local features in training data (Jones et al., 2014). Hu et al. (2015) proposed to use convolutional neural networks (CNN) to learn from hyperspectral data to classify LULC classes and their experimental results showed that proposed method can achieve reasonable performance. These new learning algorithms should be considered in future works to improve learning from multi-temporal/multi-regional ground truth data sets.

5.2.1.3 Data and model assimilation

Existing spectral data available globally are coarse in resolution till now. Spatially rich Landsat data (e.g. 30 m Thematic Mapper) are also often used for land monitoring (Vittek et al., 2014; Watts et al., 2010). Due to its 16-day repeating interval, the Landsat products are severely restricted in monitoring LULC in monsoon or tropical regions, where frequent cloud contamination occurs. In contrast, NASA's MODIS products are less prone to noise due to its shorter repeating interval (e.g. daily reflectance data) and its composition procedure (NASA Land Processes Distributed Active Archive Center (LP DAAC), 2013a). However, MODIS products are coarse in spatial resolution ($>$ 250 m). Yet multi-spectral data is likely the best spectral data one can get for a wide range of areas since hyper-spectral data is rare at the global level. Hence, data assimilation between spectral data from different sensors would be needed to prepare better input data for LULC quantification models. A fusion of a spatially rich (e.g. 30 m LANDSAT 8 data) and a temporally rich (e.g. MODIS daily 250 m data) would be a practical option for obtaining high-quality spectral input data for LULC quantification models. As these existing medium resolution multi-spectral products are available for the last decades (e.g. LANDSAT 5 is available from 1984 and MODIS data from 2001), data assimilation could increase the potential of the LULC quantification models in looking back at historical LULC records (e.g. Gao et al., 2006; Gomez-Chova et al., 2015).

In addition to data assimilation, model assimilation needs to be further developed. The resampling technique used in the multi-crop classification can be used to alleviate the problem of the data imbalance of training data in the fractional LULC study, in which also the mixture classes caused low model performance.

5.2.1.4 New learning algorithms

Finally, an interesting topic in future research would be to test the framework with new machine learning algorithms such as generative models (Murphy, 2012). For instance, deep learning has led to substantial improvements in model performance especially in object recognition, natural language processing, and multi-media data processing (Deng et al., 2014). In LULC science, Hu et al. (2015) and Lv et al., 2014 reported their experimental results showing that the deep learning algorithms are useful. These new algorithms may be able to resolve the mixed pixel problem which undermines the model performance to a certain degree. To my knowledge, it is still rare to use the recent learning algorithm in land use science. Moreover, to generalise the lessons of this study, further studies using new algorithms with diverse settings may be

needed.

To deal with the potential class overlapping, it is necessary to explore new data, learning algorithms, and pre- and post-processing methods. For instance, applying SMOTE to a high resolution dataset would be an interesting topic as it lessens the problem of the mixed pixels. Data pre-processing with SMOTE to balance the data distribution is independent of the classifier and synthetic oversampling can be plugged in into an existing classification framework without further adjustments. We only tested SMOTE with two classifiers (i.e. RF and SVM). Tests with other state-of-the-art classifiers such as deep neural networks (Deng et al., 2014) would be necessary to secure the value of the results.

5.3 Conclusions

Land use and land cover information is crucial for ecosystem services researches, decision making and studies on global change in general (Hansen et al., 2013; Schulp et al., 2011; Turner et al., 2007). Especially, it influences significantly the outcomes of environmental and ecological models (Mahecha et al., 2010; Matthews, 1983) as well as decision making studies. The quality of LULC information is important for these applications, thus acquisition of appropriate LULC data is an essential issue.

GLC data provides valuable information about various land systems such as urban, forested, shrubland, and agriculture and it remains a key data source for scientific communities and decision making groups. As shown in the dissertation, the existing GLC products have limitations in complex heterogeneous agricultural landscapes due to their coarse thematic, spatial, and temporal resolution. Therefore, the use of the GLC products may lead to an inadequate representation of the actual landscape. To deal with complex heterogeneous agricultural landscapes, improvements of the GLC data in spatial, temporal, and thematic resolution are strongly desired.

For cultivated landscapes, acquisition of detailed LULC data is an essential need, which is not sufficiently satisfied by the use of the existing GLC products so far. Therefore, this dissertation focused on the extraction of spatially and thematically detailed LULC information from existing, medium resolution, multi-spectral satellite products.

In the course of the study, three research objectives were pursued: high-quality LULC data collection, continuous representation of LULC, and classification of multi-crop LULC. The results showed that the existing GLC product was restricted in representing the studied landscape (Chapter 2) and there were feasible strategies to achieve spatially and thematically improved

LULC representation (Chapter 3, 4). Two statistical learning techniques (i.e. RF and SVM) and various data-processing techniques (e.g. SMOTE) were used to improve the performance of the LULC quantification models.

The per-field LULC census data presented in chapter 2 revealed the complex and heterogeneous distribution of LULC in the study site. The comparison between the census data and the MODIS land cover revealed that the limitations of the GLC products (e.g. inability to deal with linear elements and irrigated fields) could cause a substantial misrepresentation of the real LULC. Consequently, the agreement between the rasterised ground truth and the MODIS land cover was rather poor. For complex agricultural landscapes, global LULC ground truth datasets are still lacking (Sulla-Menashe et al., 2011).

The study presented in chapter 3 is one of the few studies addressing the fractional LULC estimation in monsoonal agricultural landscapes. Still continuous LULC representation especially with multiple land cover types is underdeveloped at the global level. The result showed that, when properly chosen and processed, coarse satellite products can be a useful information source about continuous representation of LULC. Estimating fractional LULC from available coarse resolution satellite data can be a useful strategy for obtaining that information.

In addition, the fractional cover study demonstrated how to choose optimal data-processing options. For complex cultivated landscapes such as an agricultural mosaic catchment, appropriate data processing options should be adopted to boost LULC modelling performance. The study established an evaluation framework for the options with a rigorous cross-validation scheme. Identification of a best combination of data-processing options would be a practical help for similar modelling studies.

Furthermore, the study presented in chapter 4 tackled the multi-crop LULC classification problem for the cultivated landscape. In cultivated ecosystems, LULC data are often imbalanced which undermines performance of standard learning techniques (i.e. classification and regression algorithms). The goal of the study was to improve the classification of rare classes in an agricultural mosaic catchment by using SMOTE. Artificially balancing the LULC data distribution enhanced the classification performance of some minority types and substantially improved LULC representation of the study site (i.e. LULC prediction maps). This study demonstrated how essential is the distribution of the LULC type labels in cultivated landscapes as well as how to overcome such an imbalance. As the data balancing technique is independent from other modelling steps, it can be plugged in into any learning framework without further adjustments.

The studies in chapter 3 and 4 shed a light on further application of coarse satellite products to

yield useful information. Retrieving detailed LULC information from existing satellite products can be a convenient way to augment new information with a relatively small cost. Aided by the presented modelling frameworks, publicly available remote sensing data can be used to extract such information in cultivated landscapes. This approach may also alleviate computational burden imposed by the use of high-resolution spectral data.

Enhancement of learning techniques, as well as improved data-processing methods can foster developments of new high-quality LULC databases. This dissertation presents the data and the methodological approaches to overcome the aforementioned limitations of the GLC products and satisfy the need of sound LULC data in complex heterogeneous agricultural landscapes. Aided by recent machine learning algorithms (e.g. SVM) and data-processing techniques (e.g. SMOTE), the modelling frameworks produced spatially and thematically rich LULC information for the study site. The outcome of the studies can be used as an input data for regional environmental modelling, ecosystem services research, and decision making analysis. The lessons and findings in this study can be used to create a local dataset which can complement the GLC products in complex heterogeneous landscapes. As GLC products are based on the same type of data (e.g. global satellite data) and similar algorithms (e.g. decision trees), the modelling approaches and finding of this study can be easily applied to the development of a new GLC product.

5.4 Record of publications

Peer-reviewed articles

Martin, E.; Reineking, B.; Seo, B.; Steffan-Dewenter, I. (2015). Pest control of aphids depends on landscape complexity and natural enemy interactions. PeerJ 3:e1095, doi:10.7717/peerj.1095

Harter, D.; Iri, S.; Seo, B.; Steinbauer, M.; Triantis, K.A.; Fernández-Palacios, J.M.; Beierkuhnlein, C. (2015). Impacts of global climate change on the floras of oceanic islands – projections, implications and current knowledge. Perspectives in Plant Ecology, Evolution and Systematics, 17, 160–183, doi:10.1016/j.ppees.2015.01.003

Seo, B.; Bogner, C.; Poppenborg, P.; Martin, E.; Hoffmeister, M.; Jun, M.; Koellner, T.; Reineking, B.; Shope, C.L.; Tenhunen, J. (2014). Deriving a per-field land use and land cover map in an agricultural mosaic catchment. Earth System Science Data, 6, 339-–352, doi:10.5194/essd-6-339-2014

Shope, C.L.; Maharjan, G.; Tenhunen, J.; Seo, B.; Kim, K.; Riley, J.; Arnhold, S.; Koellner, T.; Ok, Y-S; Peiffer, S.; Kim, B.; Park, J-H; Huwe, B. (2013). An interdisciplinary swat ecohydrological model to define catchment-scale hydrologic partitioning, Hydrology and Earth System Sciences, 10, 7235–7290, doi:10.5194/hessd-10-7235-2013

Martin, E.; Reineking, B.; Seo, B.; Steffan-Dewenter, I. (2013). Natural enemy interactions constrain pest control in complex agricultural landscapes. Proceedings of the National Academy of Sciences, 110, 5534-–5539, doi:10.1073/pnas.1215725110

Thanh, N.T.; Hoang, V-N; Seo, B. (2012). Cost and environmental efficiency of rice farms in South Korea, Agricultural Economics, 43, 367–376, doi:10.1111/j.1574-0862.2012.00589.x

K'Otuto, G.O.; Otieno D.O.; Seo, B.; Ogindo, H.O.; Onyango, J.C. (2012). Carbon dioxide exchange and biomass productivity of the herbaceous layer of a managed tropical humid savanna ecosystem in western Kenya, Journal of Plant Ecology, 5, 286–297, doi:10.1093/jpe/rts038

Dataset

Seo, B.; Poppenborg, P.; Martin, E.; Hoffmeister, M.; Bogner, C.; Jun, M.; Koellner, T.; Reineking, B.; Shope, C.L.; Tenhunen, J. (2014). Per-field land use and land cover data set of the Haean catchment, South Korea. PANGAEA, doi:10.1594/PANGAEA.823677

References

Asner, G. & D. B. Lobell (2000). "A biogeophysical approach for automated SWIR unmixing of soils and vegetation". In: *Remote Sensing of Environment* 74.1, pp. 99–112 (cit. on pp. 11, 67, 72, 157).

Chen, G., T. X. Han, Z. He, R. Kays & T. Forrester (2014). "Deep convolutional neural network based species recognition for wild animal monitoring." In: *ICIP*, pp. 858–862 (cit. on p. 158).

Cord, A, C Conrad, M Schmidt & S Dech (2010). "Standardized FAO-LCCS land cover mapping in heterogeneous tree savannas of West Africa". In: *Journal of Arid Environments* 74.9, pp. 1083–1091 (cit. on pp. 7, 154).

Defries, R. S., M. C. Hansen & J. Townshend (2000). "Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8 km AVHRR data". In: *International Journal Of Remote Sensing* 21.6-7, pp. 1389–1414 (cit. on pp. 11, 63, 64, 157).

DeFries, R. S., C. B. Field, I. Fung, C. O. Justice, S. Los, P. A. Matson, E. Matthews, H. A. Mooney, C. S. Potter, K. Prentice, et al. (1995). "Mapping the land surface for global atmosphere-biosphere models: Toward continuous distributions of vegetation's functional properties". In: *Journal of Geophysical Research: Atmospheres (1984–2012)* 100.D10, pp. 20867–20882 (cit. on pp. 11, 72, 157).

Deng, L. & D. Yu (2014). "Deep learning: methods and applications". In: *Foundations and Trends in Signal Processing* 7.3–4, pp. 197–387 (cit. on pp. 160, 161).

Di Gregorio, A (2005). *Land Cover Classification System: Classification Concepts and User Manual: LCCS*. Rome (Italy). Food and Agriculture Organization of the United Nations (FAO) (cit. on pp. 2, 3, 7, 43, 52, 56, 154).

Fernandes, R., R. Fraser, R. Latifovic, J. Cihlar, J. Beaubien & Y. Du (2004). "Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region". In: *Remote Sensing of Environment* 89.2, pp. 234–251 (cit. on pp. 11, 37, 63, 68, 72, 75, 85, 157).

Foody, G. M. (2007). "Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data". In: *dx.doi.org* 17.7, pp. 1317–1340 (cit. on p. 158).

Foody, G. M. & M. K. Arora (1996). "Incorporating mixed pixels in the training, allocation and testing stages of supervised classifications". In: *Pattern Recognition Letters* 17.13, pp. 1389–1398 (cit. on pp. 11, 13, 72, 157).

Friedl, M. A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley & X. Huang (2010). "MODIS Collection 5 global land cover: Algorithm refinements and characterization

of new datasets". In: *Remote Sensing of Environment* 114.1, pp. 168–182 (cit. on pp. 9, 37, 43, 54, 158).

Fritz, S., I. McCallum, C. Schill, C. Perger, L. See, D. Schepaschenko, M. van der Velde, F. Kraxner & M. Obersteiner (2011). "Geo-Wiki: An online platform for improving global land cover". In: *Environmental Modelling and Software*, pp. 1–14 (cit. on pp. 10, 158).

Gao, F., J Masek, M Schwaller & F Hall (2006). "On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance". In: *Geoscience and Remote Sensing, IEEE Transactions on* 44.8, pp. 2207–2218 (cit. on p. 160).

Gomez-Chova, L., D. Tuia, G. Moser & G. Camps-Valls (2015). "Multimodal Classification of Remote Sensing Images: A Review and Future Directions". In: *Proceedings of the IEEE* 103.9, pp. 1560–1584 (cit. on p. 160).

Guerschman, J. P., M. J. Hill, L. J. Renzullo, D. J. Barrett, A. S. Marks & E. J. Botha (2009). "Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors". In: *Remote Sensing of Environment* 113.5, pp. 928–945 (cit. on pp. 11, 12, 19, 63, 65, 71, 72, 157).

Hansen, M. C., P. V. Potapov, R Moore, M Hancher, S. A. Turubanova, A Tyukavina, D Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A Kommareddy, A Egorov, L Chini, C. O. Justice & J. R. G. Townshend (2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change". In: *Science* 342.6160, pp. 850–853 (cit. on pp. 1–3, 7, 37, 56, 161).

Hu, W., Y. Huang, L. Wei, F. Zhang & H. Li (2015). "Deep Convolutional Neural Networks for Hyperspectral Image Classification". In: *Journal of Sensors* 2015, p. 12 (cit. on pp. 159, 160).

Jones, H. & X. Sirault (2014). "Scaling of Thermal Images at Different Spatial Resolution: The Mixed Pixel Problem". In: *Agronomy* 4.3, pp. 380–396 (cit. on p. 159).

Lu, H., M. R. Raupach, T. McVicar & D. J. Barrett (2003). "Decomposition of vegetation cover into woody and herbaceous components using AVHRR NDVI time series". In: *Remote Sensing of Environment* 86.1, pp. 1–18 (cit. on pp. 11, 12, 64, 71, 72, 157).

Lv, Q, Y Dou, X Niu, J Xu & B Li (2014). "Classification of land cover based on deep belief networks using polarimetric RADARSAT-2 data". In: *Geoscience and Remote ...* (Cit. on p. 160).

Mahecha, M. D., L. M. Fürst, N. Gobron & H. Lange (2010). "Identifying multiple spatiotemporal patterns: A refined view on terrestrial photosynthetic activity". In: *Pattern Recognition Letters* 31.14, pp. 2309–2317 (cit. on pp. 1, 3, 37, 161).

Matthews, E. (1983). "Global vegetation and land use: New high-resolution data bases for climate studies". In: *Journal of Climate and Applied Meteorology* 22.3, pp. 474–487 (cit. on pp. 1–3, 37, 161).

Murphy, K. P. (2012). *Machine Learning*. A Probabilistic Perspective. MIT Press (cit. on p. 160).

Narantsetseg, A., S. Kang, B.-E. Lkhamsuren & D. W. Ko (2014). "Determinants of Caragana microphylla density distribution in the Mongolian steppe". In: *Ecological Research* 29.5, pp. 855–865 (cit. on p. 157).

NASA Land Processes Distributed Active Archive Center (LP DAAC) (2013a). *MOD13A1 Vegetation Indices 16-Day L3 Global 500m*. Tech. rep. 47914 252nd Street, Sioux Falls, South Dakota (cit. on pp. 9, 68, 69, 160).

Nguyen, T. T., H. V. Ngu & B. Seo (2012). "Cost and environmental efficiency of rice farms in South Korea". In: *Agricultural Economics* 43.4, pp. 369–378 (cit. on pp. 39, 153, 156).

Nguyen, T. T., M. Ruidisch, T. Koellner & J. Tenhunen (2014). "Synergies and tradeoffs between nitrate leaching and net farm income: The case of nitrogen best management practices in South Korea". In: *Agriculture Ecosystems & Environment* 186, pp. 160–169 (cit. on pp. 3, 157).

Poppenborg, P. & T. Koellner (2013). "Do attitudes toward ecosystem services determine agricultural land use practices? An analysis of farmer's decision-making in a South Korean watershed". In: *Land Use Policy* 31.0, pp. 422–429 (cit. on pp. 2, 3, 37, 39, 153, 156, 157).

Reineking, B. & B. Seo (2013). "Natural enemy interactions constrain pest control in complex agricultural landscapes." In: *Proceedings of the National Academy of Sciences* 110.14, pp. 5534–5539 (cit. on pp. 37, 39, 153, 157).

Schulp, C. & R Alkemade (2011). "Consequences of uncertainty in global-scale land cover maps for mapping ecosystem functions: an analysis of pollination efficiency". In: *Remote Sensing* 3.9, pp. 2057–2075 (cit. on pp. 1–3, 37, 38, 63, 161).

Schwarz, M. & N. E. Zimmermann (2005). "A new GLM-based method for mapping tree cover continuous fields using regional MODIS reflectance data". In: *Remote Sensing of Environment* 95.4, pp. 428–443 (cit. on pp. 11, 63, 64, 72, 157).

Schwieder, M., P. Leitão, S. Suess, C. Senf & P. Hostert (2014). "Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques". In: *Remote Sensing* 6.4, pp. 3427–3445 (cit. on pp. 10–12, 63, 71, 72, 83, 106, 157).

Shope, C. L., G. R. Maharjan, J Tenhunen, B Seo, K Kim, J Riley, S Arnhold, T Koellner, Y. S. Ok, S Peiffer, B Kim, J. H. Park & B Huwe (2014). "Using the SWAT model to improve process descriptions and define hydrologic partitioning in South Korea". In: *Hydrology And Earth System Sciences* 18.2, pp. 539–557 (cit. on pp. 19, 39, 47, 56, 153, 157).

Sulla-Menashe, D., M. A. Friedl, O. N. Krankina, A. Baccini, C. E. Woodcock, A. Sibley, G. Sun, V. Kharuk & V. Elsakov (2011). "Hierarchical mapping of Northern Eurasian land cover using MODIS data". In: *Remote Sensing of Environment* 115.2, pp. 392–403 (cit. on pp. 9, 158, 162).

Turner, B. L., E. F. Lambin & A. Reenberg (2007). "The emergence of land change science for global environmental change and sustainability". In: *Proceedings of the National Academy of Sciences* 104.52, pp. 20666–20671 (cit. on pp. 2, 105, 161).

Turner, W, C Rondinini, N Pettorelli, B Mora, A. K. Leidner, Z Szantoi, G Buchanan, S Dech, J Dwyer, M Herold, L. P. Koh, P Leimgruber, H Taubenboeck, M Wegmann, M Wikelski & C Woodcock (2015). "Free and open-access satellite data are key to biodiversity conservation". In: *Biological Conservation* 182, pp. 173–176 (cit. on p. 158).

Vittek, M., A. Brink, F. Donnay, D. Simonetti & B. Desclée (2014). "Land Cover Change Monitoring Using Landsat MSS/TM Satellite Image Data over West Africa between 1975 and 1990". In: *Remote Sensing* 6.1, pp. 658–676 (cit. on pp. 9, 68, 160).

Watts, J. D., S. L. Powell, R. L. Lawrence & T. Hilker (2010). "Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery". In: *Remote Sensing of Environment* 115.1, pp. 66–75 (cit. on pp. 9, 68, 160).

Wu, D., H. Wu, X. Zhao, T. Zhou, B. Tang, W. Zhao & K. Jia (2014). "Evaluation of Spatiotemporal Variations of Global Fractional Vegetation Cover Based on GIMMS NDVI Data from 1982 to 2011". In: *Remote Sensing* 6.5, pp. 4217–4239 (cit. on p. 157).

Zhao, P & J Lüers (2012). "Improved determination of daytime net ecosystem exchange of carbon dioxide at croplands". In: *Biogeosciences Discussions* 9, pp. 2883–2919 (cit. on p. 157).

(Eidesstattliche) Versicherungen und Erklärungen

(§ 8 S. 2 Nr. 6 PromO)

Hiermit erkläre ich mich damit einverstanden, dass die elektronische Fassung meiner Dissertation unter Wahrung meiner Urheberrechte und des Datenschutzes einer gesonderten Überprüfung hinsichtlich der eigenständigen Anfertigung der Dissertation unterzogen werden kann.

(§ 8 S. 2 Nr. 8 PromO)

Hiermit erkläre ich eidesstattlich, dass ich die Dissertation selbständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe.

(§ 8 S. 2 Nr. 9 PromO)

Ich habe die Dissertation nicht bereits zur Erlangung eines akademischen Grades anderweitig eingereicht und habe auch nicht bereits diese oder eine gleichartige Doktorprüfung endgültig nicht bestanden.

(§ 8 S. 2 Nr. 10 PromO)

Hiermit erkläre ich, dass ich keine Hilfe von gewerbliche Promotionsberatern bzw. -vermittlern in Anspruch genommen habe und auch künftig nicht nehmen werde.

. . . . . .

Ort, Datum, Unterschrift