Behind subjectivity

Factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring

Agnieszka Tarko

Propositions

1. Visual interpretation of land cover must use false colour composite images. (this thesis)

2. Feedback loops are essential for learning visual image interpretation. (this thesis)

3. To preserve biodiversity, endangered tree species must be protected in plantations.

4. Without a Complex Adaptive Systems view, integrating social and biophysical sciences is impossible.

5. With the ubiquity of online mapping applications, map interpretation has become a necessity and must be included in school curricula.

6. Regional knowledge cannot be overstated.

Propositions belonging to the thesis, entitled

Behind subjectivity. Factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring

Agnieszka Tarko
Wageningen, 17 December 2019

Behind subjectivity

Factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring

Agnieszka Tarko

Thesis committee

Promotor

Prof. Dr A. K. Bregt
Professor of Geo-information Science and Remote Sensing
Wageningen University & Research

Co-promotors

Dr S. de Bruin
Associate professor, Laboratory of Geo-information Science and Remote Sensing
Wageningen University & Research

Dr N. E. Tsendbazar
Lecturer, Laboratory of Geo-information Science and Remote Sensing
Wageningen University & Research

Other members

Prof. Dr F. van Langevelde, Wageningen University & Research
Prof. Dr C. C. Fonte, University of Coimbra, Portugal
Prof. Dr A. D. Nelson, University of Twente
Dr P. Wojda, European Commission, Brussels, Belgium

This research was conducted under the auspices of the C.T. de Wit Graduate School of Production Ecology & Resource Conservation (PE&RC)

Behind subjectivity

Factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring

Agnieszka Tarko

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr. A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 17 December 2019
at 11.00 a.m. in the Aula.

Agnieszka Tarko

Behind subjectivity
Factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring

132 pages

PhD Thesis, Wageningen University, Wageningen, NL (2019)
With references, with summary in English

ISBN: 978-94-6395-155-5
DOI: https://doi.org/10.18174/502363

Table of Contents

Chapter 1

Introduction

1.1 Background

Agricultural activities have played a key role in shaping the landscape for millennia and continue to do so (DG AGRI, 2018). Agricultural land is important at global scale. Population growth and increased consumption raise the global demand for food. This results in increased competition for land and causes changes in land use (Godfray et al., 2010). On the other hand, there is the urge to reduce the environment impact of agricultural production. Also the effects of climate change affecting crop performance needs to be taken into account. Many policies aim to ensure food security, the sustainable use of natural resources and the balanced development of rural areas. Agricultural land monitoring is essential for those policies. Large-scale (national, multi-national to global scale) maps of land cover, land use and land change meet those needs (Fritz et al., 2013).

Today, large-scale maps are typically obtained through visual or automated interpretation of remotely sensed images. Information resulting from image interpretation can be presented as a map or as a tabular reference dataset. Both, map production and map quality assessment need reference data. Reference data –as the name implies– serve as a reference of the true situation that is to be represented on the map.

Before digital imagery became common place, trained interpreters visually interpreted analogue aerial images. Nowadays, digital sensors located in air-borne or space-borne platforms acquire remotely sensed images. Digital sensors detect and record the electromagnetic energy reflected by the earth surface without direct contact with this surface. Recorded images are sold or licensed to clients (e.g. governments). Recently they have become more readily available and accessible for the general public as well. Well-known platforms such as Google Earth and Bing Maps provide open-access to remotely sensed images.

Analysis of remotely sensed images involves transforming data displayed in the images into information about land. This can be achieved by visual interpretation or automated methods. Several decades ago, visual interpretation was most prevalent (Bianchetti and MacEachren, 2015). With progress in earth observation, data collection and computational power, today automated methods are the most widely used (Bianchetti and MacEachren, 2015; Coppin et al., 2004; Tewkesbury et al., 2015; White, 2019). Current large-scale land cover maps are typically produced using automated image classification methods. Training and validation of large-scale maps rely on higher quality reference data. Collecting reference data for large-scale maps through field

visits typically is too costly or otherwise infeasible. Instead, reference data collection rely on visual interpretation. Interpretation involves data with higher resolution than those used for map creation (Chen et al., 2015; Hansen et al., 2013; Pengra et al., 2019; Tsendbazar et al., 2018).

Humans excel in image interpretation tasks, where visual problem-solving activity requires knowledge and experience for successful information extraction (White, 2019). On the other hand, recent advances in computer vision, machine learning and artificial intelligence are promising for automated agricultural land monitoring (Patrício and Rieder, 2018). Automated methods may replace human interpreters in a majority of those tasks. Will land map production and agricultural land monitoring be fully automated in the next decades? Similarly to Bianchetti and MacEachren (2015), I still expect human intervention to be indispensable, for example for final quality checks of the products, for product satisfaction assessment, when dealing with new, unexpected situations and when solving newly arrived issues.

Currently, trained human interpreters often perform visual interpretation for reference data collection. There are many factors affecting the quality of the results of visual image interpretation. It depends, among others, on the interpreter's training and experience, on the nature and complexity of interpreted object or phenomena and on the quality of images used for interpretation (Lillesand et al., 2008; Pengra et al., 2019; White, 2019). Expertise in visual image interpretation has always been recognised in remote sensing literature (White, 2019). But visual image interpretation is also known to be subjective and hence inconsistent between interpreters (Olofsson et al., 2014; Pengra et al., 2019; Powell et al., 2004). Understanding of important factors influencing visual image interpretation and ways to improve consistency of interpretation often come from personal experience (White, 2019). Formal research on this subject is limited. This thesis contributes to understanding of visual image interpretation and its subjectivity by assessing factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring. In particular the focus is on preferred image characteristics, agreement between visual interpretations and automatically detected cast shadows, a way to improve consistency of visual image interpretations of land cover reference data and finally, factors influencing interpretation consistency of land change.

1.2 Image interpretation

The analysis of remotely sensed images allows obtaining information about the earth's surface. It emerged as discipline during World War I, when visual interpretation of analogue images taken from the airplane served for reconnaissance (Lillesand et al., 2008; White, 2019). Literature from this period acknowledges the role of humans in the analysis process (White, 2019). With the advent of satellite earth observation, the introduction of digital sensors and increasing computing capabilities, remote sensing research shifted towards automated methods for image interpretation (Bianchetti and MacEachren, 2015).

Nowadays also visual interpretation relies on digital images. Digital images are matrices, where the elements are called pixels. The value of each pixel relates to the intercepted radiance and is recorded as a digital number (DN, pixel value). Spatial size of the pixel determines image resolution. Images can be also described by the spectral characteristics, containing data from single or multiple spectral bands. Pixels of displayed multispectral image include multiple DNs, one for each spectral band. To visualise a colour image of a scene as perceived by the human eye, three separate spectral bands of different wavelength ranges must be displayed. Images are also described by the presence of noise, cloud coverage, cast shadows. Qualities such as colour composite for multispectral images, image brightness and contrast can be set according to the users' appreciation. The image acquisition time informs when the radiance was measured. Digital images can be acquired on different platforms (Lillesand et al., 2008).

For local-scale (e.g. a district or municipality) observation, the most common platforms for acquisition of aerial images are aircrafts and unmanned aerial vehicles (UAVs or "drones"). To observe earth surface at large-scale, satellite images can be used. Imaging satellites are operated by governments and businesses around the world and are sold or licensed to the clients. Recently, with the growing public utility of satellite images, they become more available and accessible not only for scientific use but also for general public, such as openly-accessible European Commission's Sentinel images with 10 m spatial resolution (ESA, 2019).

Democratisation of earth observation data and the more interactive online interfaces for image access encouraged trained interpreters and general public to collect spatial data by means of visual image interpretation. In scientific literature humans performing visual image interpretation are differently referred to. Most commonly they are called "interpreters" (McRoberts et al.,

2018; Olofsson et al., 2014; Pengra et al., 2019; Zhao et al., 2014) but also the terms "operators" (Van Coillie et al., 2014), "annotators" (Jia et al., 2016) or "experts" (See et al., 2013; Tsendbazar et al., 2018) are used. Moreover, there is no clear definition who the expert image interpreter is. For example See et al. (2013) defined experts as "individuals with a background in remote sensing/spatial sciences", Tsendbazar et al. (2018) as ones with "experience in satellite based land cover analysis and image interpretation" and Zhao et al. (2014) as ones with "sufficient image interpretation experience". Spatial data collection performed by a general public is called crowdsourcing (See et al., 2015). See et al. (2013) assessed the quality of crowdsourced data collected through a Geo-Wiki (platform for engaging citizens in environmental monitoring) competition, where a degree of human impact and classified land cover were identified by volunteers and by experts and the results were compared. Results showed that there is little difference between experts and non-experts in identifying human impact but when it comes to land cover identification, experts were better than volunteers. In the medical domain studies have shown explicit difference between expert and novice image analyses (White, 2019).

To perform image classification, various automated methods have been developed. The objective of automated image classification is to replace visual analysis with quantitative technologies and statistically based decision rules for feature identification (Lillesand et al., 2008). Automated image classification, unlike the one performed by human interpreters, is reproducible, less laborious and often faster. Advances in automated methods in the remote sensing domain have been reviewed regularly. A recent paper of Phiri and Morgenroth (2017) describes developments since 1970s in land cover classification methods based on images acquired by Landsat, the longest running satellite earth observation programme. The first methods for land classification were visual, followed by unsupervised and supervised pixel-based classification methods using different statistical decision rules (maximum likelihood, K-means, Iterative Self-Organizing Data Analysis Technique classifiers ISODATA). A decade later, other methods were developed: sub-pixel-based, knowledge-based, contextual-based, object-based image analysis and hybrid approaches (Phiri and Morgenroth, 2017). Most recently, deep learning algorithms, characterised by neural networks involving more than two layers, have been exploited and researched in the remote sensing domain. Recent advances in deep learning in remote sensing are revived by Zhu et al. (2017). Yet, automated image classification methods are not perfect and require human input in various forms such as interpretation of training data (Phiri and Morgenroth, 2017). Also the reference data used for developing and validating large-scale land

change maps are commonly acquired by visual interpretation. Interpretation involves remotely sensed images with higher resolution than those used for map creation (Olofsson et al., 2014).

Visual interpretations made by multiple interpreters can differ, indicating the subjectivity of visual interpretation (Jia et al., 2016; Pengra et al., 2019; Powell et al., 2004). Although image interpretation is commonly used for reference data collection, only few studies aimed to increase understanding of the causes of inconsistency in visual image interpretations. Gardin et al. (2011) designed a web-based digitisation exercise aiming at assessment of human factors that influence interpreter performance. It was found that the performance was mainly determined by personality factors and that there was a gradual decline in performance accuracy over time (Van Coillie et al., 2014). The reasoning process of image interpretation and the importance to understand and describe it recently re-gained attention (White, 2019). Pengra et al. (2019) assessed duplicate interpretations of randomly selected pixels. Results showed that agreement between interpreters varied. Some agreed more than the others depending on land cover class and regional location of the interpreted pixel.

To increase consistency of visually interpreted reference data, various methods have been proposed and implemented. Zhao et al. (2014) included a review of acquired validation dataset for a global land cover map. Sites were collected by interpreters and later checked by interpreters with "outstanding skills in image interpretation". Tsendbazar et al. (2018) implemented a review with feedback in the process of land cover reference data visual interpretations. Whatever the way forward to increase the consistency of visually interpreted reference dataset, it is important to understand factors that affect visual image interpretation consistency.

1.3 Agricultural land monitoring

In 2003 as much as 11% (1.5 billion ha) of the globe's land surface was estimated to be used for crop production (Bruinsma et al., 2003). Even though it was estimated that roughly a fifth of the global land still have crop production potential, in some regions the perception was that no more or very little land can be brought into cultivation (Bruinsma et al., 2003). There are various programs implemented around the world to monitor agricultural land at large-scale level. Examples are the US Farm Service Agency (FSA, 2017) and the Chinese GIS-based land registry system (Rabley and Yuen, 2009). Large-scale maps are crucial to support those programs. Thus various remotely sensed global and large-scale land maps were produced. Examples are:

- the Chinese Earth land cover map GlobeLand30 (GLOBELAND30, 2019; Jun et al., 2014);
- the U.S. Geological Survey Land Change Monitoring, Assessment, and Projection land cover change global map (USGS LCMAP) (USGS, 2019; Young, 2017);
- the Pan-European land cover inventory CORINE (CORINE, 2019);
- an ongoing programme of yearly global land cover maps at 100m spatial resolution within the Copernicus Global Land Service: Dynamic Land Cover project (CGLS, 2019).

Large-scale maps vary in the used input images, definitions of classified land and in the validation approach. Training and validation of such maps relies on higher quality reference data. Collecting reference data through field visits typically is too costly or otherwise infeasible. Instead, visual interpretation of sampled areas on images with higher resolution can be performed. Such an approach resembles the validation practice adopted for GlobeLand30 (Chen et al., 2015) and training data collection for Copernicus Global Land Service (CGLS, 2019). Since visual interpretation differs between interpreters and is subjective, the visually interpreted reference data cannot be considered as definite "ground truth".

Consistency of interpretations depends, among others, on the interpreted land type (Pengra et al., 2019). Definitions of classified land on maps or reference data can be described as land cover, land use or both combined. Di Gregorio and Jansen (1998) defined land cover as the observed biophysical cover of the earth's surface. They proposed and developed a universally applicable parametric Land Cover Classification System (LCCS). LCCS became widely applied in remote sensing community. With time LCCS emerged as a standard, replying to the need for a map to be comparable and meet multi-user requirements. Land cover is widely used to categorise image classification, but it does not reflect the function of land. Following Di Gregorio and Jansen (2000), land use is "characterised by the arrangements, activities and inputs people undertake in a certain land cover type to produce, change or maintain it". Even though land use classification system has been previously proposed (Gong et al., 2009), there is no single international standard in use so far. The distinction between land cover and land use is not always respected in classifications. For example in CORINE land cover inventory the land cover criteria are inconsistently applied (Jansen and Gregorio, 2002).

Monitoring changes related to agricultural land is both important and challenging. There are large-scale projects involving land cover change, such as above mentioned USGS LCMAP (USGS, 2019; Young, 2017). To detect the

presence of land cover change, change matrices are derived from land cover maps representing different time frames. Another approach to monitor changes was proposed by Comber and Wulder (2019) and Lesiv et al. (2018). In this approach information about change can be acquired through direct interpretation of land transformation.

Another example of a large-scale change detection and near-real time monitoring program is linked to the European Union (EU) Common Agricultural Policy (CAP). The CAP policy aims to ensure food security, the sustainable use of natural resources and balanced development of Europe's rural areas (DG AGRI, 2018). Agricultural land is monitored using the Land Parcel Identification System (LPIS). It is a monitoring approach implemented by the EU Member States (European Commission, 2013). So-called reference parcels are delineated on the basis of very high resolution (VHR) images in the scope of the Control with Remote Sensing program (LPIS TG ETS, 2017a). Along with other sources such as farmers' declarations, VHR satellite images are used not only for LPIS updating but also for quality assurance of the system through the annual Executable Test Suite (ETS) (European Commission, 2014a; LPIS TG ETS, 2017b). Accurate land identification and area quantification are of key importance for implementation of the European CAP (Devos and Milenov, 2013), because payments to European farmers are area-based. LPIS data allow localisation of the agricultural parcels claimed by farmers and quantification of their eligible areas (LPISQA, 2014a). Failures of LPIS may lead to under- or over-declaration, which implies substantial financial risk to the EU (LPISQA, 2014a), as CAP payments amount to around 30% of the EU budget (data from 2011) (European Commission, 2014b). In 2013, CAP expenditures totalled roughly 44 billion euros (European Commission, 2014c), underscoring the importance of LPIS.

Quality assessment of LPIS is regulated by article 6 of Commission Delegated Regulation (EU) No 640/2014 and is performed by the Member States on a yearly basis. In order to assess and improve LPIS quality, the Joint Research Centre (JRC) established the LPIS Quality Assurance (LPISQA) framework in 2010. Within this framework, satellite and aerial images (acquired every year) are used as the main basis for inspection procedures, which include land identification, data description and reporting (LPISQA, 2014a). Within the LPISQA framework, agricultural land identification is performed by visual interpretation of images. The outcome of the inspection process is highly dependent on correct image interpretation, which, in turn, depends on interpreter skills, experience and knowledge of the area of interest, as well as on image quality (Jensen, 2000). The LPISQA protocol requires that images are

of sufficient quality to allow determination of the nature of objects and, especially, identification of eligible agricultural land types.

In line with the legal framework, new requirements were adopted to promote agricultural practices beneficial for the climate and environment called the "greening" of CAP (Tóth and Kučas, 2016). Greening, among others, requires capturing ecological focus areas within LPIS to support biodiversity, contribute to climate resilience and protect soils and habitats as well as ground and surface water (Tóth and Kučas, 2016), following the urge to accommodate environment impact reduction. On the other hand, greening results in more complex interpretation of agricultural land on images. Changes in the legal framework, developments in image acquisition and availability, together with rising computational capacities set new challenges in remote sensing, spatial data management and also in visual image interpretation of agricultural land.

1.4 Research gaps in visual image interpretation

Visual interpretation of remotely sensed data is important for validation and training of large-scale maps. However, it is also subjective and hence there is inconsistency between interpreters (Olofsson et al., 2014; Pengra et al., 2019; Powell et al., 2004). Thus reference data obtained by means of visual interpretation and often used for assessment of land cover, land use, land change products may not be consistent. Typically, characteristics of visual interpretation process in remote sensing literature comes from personal experience (White, 2019). Comprehensive research and knowledge on factors affecting visual image interpretation are limited.

The quality of images used for agricultural land interpretation and classification is influenced by many factors, such as spatial resolution affecting the distinguishability of features and geometric quality affecting geolocation accuracy (Poli, 2014). Geometric image quality is well-understood and can be described using pixel size or statistics such as mean-squared error (Smits et al., 1999). On the other hand, little is known about preferred image quality such as colour composite, image brightness or contrast settings for image interpretation. Trained interpreters can adjust those setting according to their personal preferences. But when relying on readily available VHR images available through applications such as Google Earth or Bing Maps, image settings are impossible to adjust. Little is known about interpreter' preferences regarding photometric image properties.

Another image characteristic which may hinder successful interpretation is presence of shadow. Shadows can complicate delineation of agricultural land, hindering monitoring programs. A cast shadow covering a boundary of an agricultural parcel can affect visual interpretation in applications such as LPIS. While there are several automated methods for shadow detection on remotely sensed images, their accuracy is often assessed against one visual interpretation result treated as "truth"(Adeline et al., 2013; Tsai, 2006). Bearing in mind that visual interpretation is subjective, it is unclear to what extent interpreters are consistent in their assessment of cast shadows. Little is known how visual interpretations of cast shadows differ from an automated method for cast shadow detection in the LPIS context.

One way to deliver data of higher quality than the assessed map is by visually interpreting images with higher resolution than the ones used for map generation. (Olofsson et al., 2014). Reference data collected by means of visual interpretation has been used by several large-scale land cover map initiatives (Pengra et al., 2019; Tsendbazar et al., 2018). However, visual image interpretation only approximates the ground truth and may be inconsistent. Knowledge on ways to improve consistency of visual interpretation often comes from personal experience (White, 2019). Little is known about the influence of review and feedback on consistency of visually interpreted land cover reference data.

Successful agricultural land monitoring requires change identification. Similarly to interpretation of land cover, visual interpretation of change is subjective and inconsistent between interpreters. Nevertheless, reference data collected by the means of visual interpretation on VHR images is considered of higher quality than the large-scale change map product. Visually interpreted reference data is used for training and validation of land change maps. Little is known about factors influencing the quality and consistency of changes detected by visual interpretation.

1.5 Research objectives

Within the context of the research gaps described in the previous section, the overall objective of this thesis is to identify factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring.

Based on this objective, the following four research questions are answered:

1) What image characteristics are preferred by visual interpreters?

2) What is the agreement between visual interpretations and automatically detected cast shadows?

3) How can the consistency of land cover reference data acquired by visual image interpretations be improved?

4) Which are important factors influencing interpretation consistency of land change reference data?

1.6 Thesis outline

The remainder of this thesis is structured along five chapters. Chapters 2 through 5 address each of the research questions listed above in sequential order. Next, chapter 6 synthetises the main findings of this thesis by answering the four research questions and giving recommendations for further research.

Chapter 2

Users' assessment of image photometric quality for visual interpretation of agricultural land

Agnieszka Tarko, Sytze de Bruin, Dominique Fasbender, Wim Devos, Arnold K. Bregt

This chapter is based on:

Tarko, A., de Bruin, S., Fasbender, D., Devos, W., Bregt, A., 2015. Users' assessment of orthoimage photometric quality for visual interpretation of agricultural fields. Remote Sens. 7, 4919–4936. https://doi.org/10.3390/rs70404919

Ancillary data and high resolution figures can be found in the online publication.

Abstract

Land cover identification and area quantification are key aspects of implementing the European Common Agriculture Policy. Legitimacy of support provided to farmers is monitored using the Land Parcel Identification System (LPIS), with land cover identification performed by visual image interpretation. While the geometric image quality required for correct interpretation is well understood, little is known about the photometric quality needed for LPIS applications. This paper analyses the image quality characteristics chosen by authors as being most suitable for visual identification of agricultural land. We designed a survey to assess users' preferred brightness and contrast ranges for images used for LPIS purposes. Survey questions also tested the influence of a background colour on the preferred image brightness and contrast, the preferred image format and colour composite, assessments of images with shadowed areas, appreciation of image enhancements and, finally, consistency of individuals' preferred brightness and contrast settings across multiple sample images. We find that image appreciation is stable at the individual level, but preferences vary across respondents. We therefore recommend that LPIS operators be enabled to personalise photometric settings, such as brightness and contrast values, and to choose the displayed band combination from at least four spectral bands.

Keywords

image quality assessment; photometric quality; Land Parcel Identification System (LPIS); quality assurance; visual image interpretation

2.1 Introduction

With the ongoing prevalence of digital sensors, both air-borne and space-borne imaging techniques allow efficient identification of land cover. Accurate land cover identification and area quantification are of key importance for implementation of the European Common Agriculture Policy (CAP) (Devos and Milenov, 2013), because payments to European farmers are area-based. Eligibility for CAP support is managed via the LPIS, as individually implemented by the Member States (MS) of the European Union (EU). LPIS data allow localisation of the agricultural parcels claimed by farmers and quantification of their eligible areas (LPISQA, 2014a). Failures of LPIS may lead to under- or over-declaration, which implies substantial financial risk to the EU (LPISQA, 2014a), as CAP payments amount to around 30% of the EU budget (data from 2011) (European Commission, 2014b). In 2013, CAP expenditures totalled roughly 44 billion euros (European Commission, 2014c), underscoring the importance of LPIS precision. Quality assessment of the identification system for agricultural parcels is regulated by the Article 6 of Commission Delegated Regulation (EU) No 640/2014 and is performed by the MS on a yearly basis. In order to assess and improve LPIS quality, the Joint Research Centre (JRC) established the LPIS Quality Assurance (LPISQA) framework in 2010. Within this framework, satellite and aerial images (acquired every year) are used as the main basis for inspection procedures, which include land cover identification, data description and reporting (LPISQA, 2014a).

Within the LPISQA framework, land cover identification is performed by visual interpretation of images. Visual image interpretation includes determination of the nature of objects on an image and a judgment on their significance (Paine and Kiser, 2012). Object significance is particularly important for parcel delineation and for distinguishing eligible land. The outcome of the inspection process is highly dependent on correct image interpretation, which, in turn, depends on operator skills, experience and knowledge of the area of interest, as well as on image quality (Jensen, 2000). Currently, most air-borne and satellite images are multispectral. They are provided in digital format, with a spatial resolution of 0.5 m or less. The LPISQA protocol requires that images be of sufficient quality to allow determination of the nature of objects and, especially, identification of eligible land cover types.

The quality of imagery data and their fitness for use are influenced by many factors. Among these are spatial resolution, which affects the distinguishability of features and the scale at which the image can be displayed, and geometric

quality, which affects geolocation accuracy (Poli, 2014). Geometric image quality is well understood and can be described using standardised measures such as mean-squared error (Smits et al., 1999). The LPISQA guidelines recommend that the spatial resolution of geometric images be 1 m or less and that the visual scale at which image interpretation is performed be larger than 1:5000 (LPISQA, 2012). Temporally, images are acquired on a yearly basis in the crop-growing season. Although the LPISQA framework does not define a required spectral resolution, in practice for visual image interpretation of agricultural land four bands are used: blue, green, red and near infrared (LPISQA, 2014b, 2014c).

The LPISQA guidelines recommend a radiometric resolution of images of at least 8 bits, but 10–11 bits per channel is strongly advised. Imagery quality is also addressed by the minimum look angle, which is recommended to be at least 56 degrees (Astrand et al., 2014; LPISQA, 2014b, 2014c). In contrast to the geometric image quality required for LPISQA applications (LPISQA, 2014c), little is known about the photometric image quality that is needed (LPISQA, 2014d). There is no standard measure for photometric image quality, yet data objectivity and their comparison potential depend, among other things, on the photometric quality of the images (Honkavaara et al., 2009). Assessment of the first two years of LPISQA implementation revealed a suboptimal quality of some of the image data used, e.g., degraded photometric quality or use of a low-quality image where a better alternative was available (Devos et al., 2012).

Studies of the photometric quality of images do exist, but the topic is multifaceted and hence complex (Wang et al., 2002). While there is a broad literature on radiometric and photometric aspects of images, its main focus is on radiometry during image production (Honkavaara et al., 2009; Markelin and Honkavaara, 2008; Olsen et al., 2010). Typically, quality assessment metrics are designed for natural, close-range images (e.g., tested using the Tampere Image Database 2008 (TID2008)) and intended for evaluation of full-reference image (Ponomarenko and Lukin, 2009) or video (Pinson and Wolf, 2003). There are also proposals for no-reference image quality assessment models (Mittal et al., 2013; Sheikh et al., 2005; Zhang and Chandler, 2013). Furthermore, studies in the medical domain have examined, e.g., digital radiography and base images quality parameters applicable to a wide variety of imaging tasks (Krupinski et al., 2007). A study by Pyka (2009) considers photometric image quality for orthophoto map production. Another study compares the radiometric quality of satellite images (GeoEye-1 versus WorldView-2) using diverse quality indicators, including visual assessment (Aguilar and Saldaña, 2013).

As the LPISQA framework offers no standards for image photometric quality and image processing is in MS management, the current study investigates users' assessments of the suitability of images with various characteristics for the task of visual identification of land cover and agricultural land within the LPIS context. The aim here is to determine whether there are general preferences for image brightness and contrast settings, colour composite, noise and shadows and file formats. This research focuses on the image settings set by the operator before the actual delineation step. Understanding users' preferences regarding these properties may help in deriving an image standard for LPISQA applications.

2.2 Data and Methods

To investigate users' preferences, we developed an online survey in which multiple images were presented on a computer screen and respondents were asked to indicate the best and sometimes also the worst image sample for the purpose of agricultural land delineation. The survey is available as ancillary data to the article. The survey was relatively short (containing 29 questions), so the average time for participants to complete it did not exceed 15 min. Some of the images used were repeated in later questions presented in a slightly different way. In order to ensure that respondents did not go back and reconsider earlier choices, the process of answering was one-way: answers were saved as they were submitted, and could not be changed later.

2.2.1 Image Selection and Processing

The aerial and satellite images included in the survey were sampled from those used for the LPISQA in the years 2011 and 2012. The pixel size of aerial images was 0.25 m or 0.50 m; of satellite images the pixel size was 0.50 m. Pansharpened georeferenced satellite images (WorldView-2, QuickBird-2 and GeoEye-1) were taken from the JRC Community Image Data Portal (European Commission, 2014d). Both false colour composites and normal colour composites were selected and the images were presented in either the lossless tagged image file format (TIFF, without compression) or the lossy Enhanced Compression Wavelet (ECW) format, as delivered by the MS. The ECW format compress large images and retain their visual quality. TIFF and ECW are the two most commonly used image formats for LPISQA purposes (LPISQA, 2014e, 2014f). Image samples were chosen from the broadest possible range of European landscapes. They were from the following MS: Bulgaria, Germany, France, Ireland, the Netherlands, Poland and Slovenia (Figure 2.1). The survey asked participants to compare images displayed with

various levels of brightness and contrast and with other enhancements typically used in LPISQA.

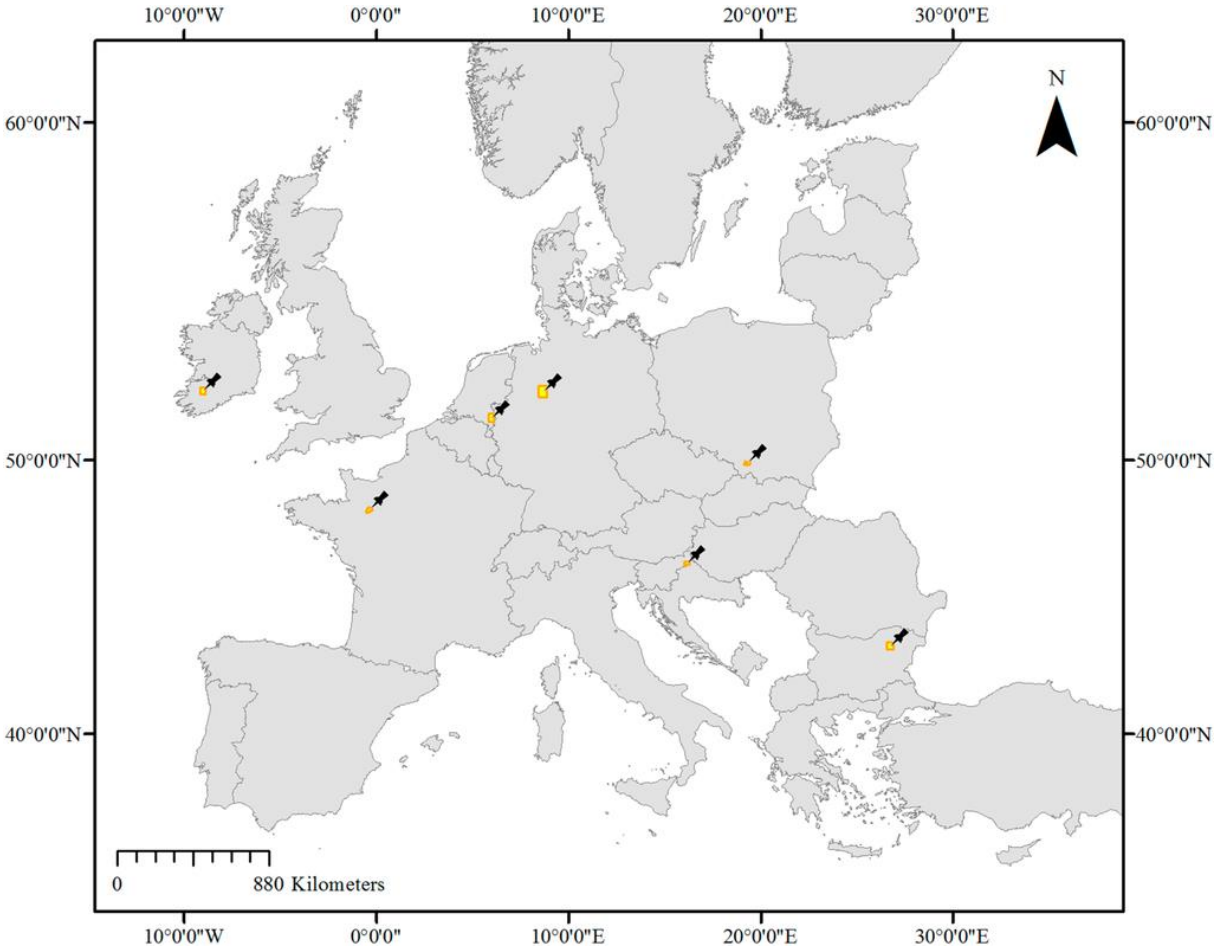


Figure 2.1. The images selected for the online survey are from the zones outlined in yellow (indicated with a black pin).

The selected images were modified for the survey. The original image, as delivered by the MS, served as the default one, i.e., with image brightness and contrast centred at zero. Brightness and contrast were then varied using the open-source GIMP2 software (GNU, 2014). Image enhancement was done using ENVI (ENVI, 2014) interactive display functions. All default histogram stretch options were used: linear stretch using the data minimum and maximum to perform a linear contrast stretch (without clipping), linear 0–255 (with the digital number (DN) values of the pixels displayed as a range from 0 to 255), square root stretch (taking the square root of the values in the input histogram and then applying a linear stretch), and finally the linear 2% stretch (a linear method with a 2% clip on both ends of the distribution of each band).

The survey consisted of four main sections. The first section (seven questions) focused on characteristics of the respondent. Specifically, it queried respondents' level of education, geographic location of practice, years of experience in visual image interpretation, and activities and years of experience with LPIS. Since we were unable to control for colour blindness in our sampling, the participants' colour vision was tested using part of the Ishihara test (Ishihara, 1972). We used three plates, designed to give a quick assessment of colour vision deficiency. The second section of the survey (two questions) focused on preferred image brightness and contrast. Seven versions of a single image were provided in each question, and participants were asked to choose the one they considered best for the purpose of agricultural land delineation. The third section (11 questions) tested preferences for combinations of brightness and contrast. Each question presented two or three renditions of a single image. Participants were asked either to indicate the best image for the purpose of agricultural land delineation or to choose both the best and the worst image for this purpose. Some of these questions repeated the same image samples used earlier but with a different colour background (either black or white, Figure 2.2). The fourth section of the survey (8 questions) tested preferences for false colour composite or natural colour composite images, as well as for TIFF or ECW formats and standard image enhancements.

Figure 2.2. Two questions concerning brightness and contrast using the same images but presenting them on different background colours—white (question 3, Left) and black (question 11, Right). The images displayed have the following properties (brightness and contrast): A (-30, 0), B (0, 0) and C (15, 0).

2.2.2 Sample Selection

Respondents in our sample represented three groups:
1. Technical and administrative staff involved in the LPISQA,

2. Professionals with visual image interpretation experience, and
3. Students.

The majority of respondents were LPISQA technical and administrative staff from a variety of MS. This group represents the main "target group" of the study, as they are the major users of image evaluations. They completed the survey during a LPISQA workshop in Baveno, Italy, in 2013. At that workshop three stations were set up for completing the online survey. Each was equipped with a similar laptop and using identical screen settings. The second group of respondents consists of employees of JRC (Ispra, Italy) and Wageningen University (Wageningen, The Netherlands). These were all experienced in image interpretation but not directly involved with LPIS activities. The final group of participants was made up of students from the Laboratory of Geo-information Science and Remote Sensing at Wageningen University and from the Remote Sensing, Photogrammetry and Geoinformation Department of the Krakow Academy of Science and Technology (Krakow, Poland). The survey was administered to all three groups from September to November 2013. During that period, 197 valid complete records were collected (Table 2.1).

Visual inspection of processed images is the traditional means of evaluation, as human observers can recognise distortion and degradation of imagery without referring to the original image (Choi et al., 2009). Our survey clearly specified the aim of image interpretation as to delineate agricultural land. The aim here was to pinpoint specific issues connected with image interpretation for this purpose.

Table 2.1. Three groups of respondents completed our online survey.

Group of Respondents	Subgroup of Respondents	Nr of Respondents
1. LPISQA technical and administrative staff		82
2. Professionals with experience in visual image interpretation	Wageningen University Joint Research Centre	25 16
3. Students	Wageningen University Krakow Academy of Science and Technology	32 42
Total		197

2.2.3 Analysis Methods

Table 2.2 lists the methods used to analyse the survey results. Most analyses were performed in R (R Core Team, 2017).

Table 2.2. Methods of analysis of the survey results (the survey can be found as ancillary data to the article).

No.	Aspect	Items	Analysis Method
1.	Brightness	Question about the preferred brightness, to be chosen from seven different levels (survey section II, item 1)	– Count and plot (line chart)
2.	Contrast	Question about the preferred contrast, to be chosen from seven different levels (survey section II, item 2)	– Count and plot (line chart)
3.	Brightness and contrast combined	Questions asking the participant to choose the best image or the best and worst images out of two or three samples, each with a different combination of brightness and contrast. Three duplicate questions were included to determine the effect of a white or black background (section III of the survey, items 1–11)	– For choice of best brightness and contrast: Count and plot (bar chart) – Estimated Shannon entropy (Shannon, 1948; Singh, 2013; Tuomisto, 2010) and its estimated standard deviation (Harris, 1975) (for image triplets, boxplot) – Plot (boxplot) of distance from the preference indicated earlier and the answer to the current question for each respondent
4.	Format and colour composite	Questions asking the participant to choose the best or the best and worst of two or three images of different format and colour composite, all with default (as delivered by the MS) brightness and contrast (survey section IV, items 1, 4 and 6)	– Normalised index (the ratio of the frequency of an image being chosen as best and as worst, normalised to the [0,1] interval) – Contingency table (the worst image properties as a function of best ones)
5.	Standard enhancement	Questions asking the participant to choose the best or the best and worst of two or three images with different enhancements: – Three questions, each presenting three images with four different types of enhancements used throughout (survey section IV, items 2, 5 and 8) – Two questions, each presenting a pair of images, one with default settings (as delivered by the MS) and a second with 2% stretch applied. In one question, part of the border of land under inspection is obscured by a shadow (survey section IV, items 3 and 7)	– Percent of count – Estimated Shannon entropy (Shannon, 1948; Singh, 2013; Tuomisto, 2010)

The brightness and contrast levels chosen as best in the first questions were later considered as the reference values in the checks for individual consistency of choices, later on referred to as "canters".

For a given question, the estimated Shannon entropy (as a measure of dispersion) was calculated as follows:

$$\hat{H}(X) = -\sum_{i=1}^{g} \hat{P}(x_i) \ln \hat{P}(x_i) \tag{2.1}$$

where
$\hat{H}(X)$ = the estimated Shannon entropy,
x_i = the event of choosing image i in the question,
$\hat{P}(x_i)$ = the estimated probability mass from the histogram,
g = the number of images used in the question (either two or three).

To analyse the precision of the estimated Shannon entropy values, the standard deviation was also calculated (Harris, 1975). This enabled us to assess the spread of the estimated entropy over the number of possible responses to a question:

$$\hat{\sigma}_{\hat{H}} = \sqrt{\frac{1}{N}\left[\sum_{i=1}^{g} \hat{P}(x_i) \ln^2 \hat{P}(x_i) - \hat{H}^2\right] + \frac{g-1}{2N^2}} \tag{2.2}$$

where
$\hat{\sigma}$ = the estimated standard deviation of the estimated Shannon entropy,
N = the sample size (in our study 197).

Image format and composite appreciation (Table 2.2; row 4) was calculated for each given image sample as the ratio between the frequency of it being chosen as best and as worst; normalised to the [0,1] interval. Normalisation allowed to compare survey questions that use different numbers of sample images. A contingency table was developed of the image format appreciations using the least appreciated image properties as a function of the most appreciated ones. The index was calculated using the following equations:

$$c_i = \frac{a_i}{b_i} \text{ and } c_i^{'} = \frac{c_i}{\sum_{i=1}^{g} c_i} \tag{2.3}$$

where
c_i = the partial ratio value,
c'_i = the normalised ratio value,
a_i = the number of times the image was chosen as best in response to the question,
b_i = the number of times the image was chosen as worst in response to the question.

In the discussion below, the estimated Shannon entropy and its standard deviation are termed simply entropy and standard deviation of entropy, respectively.

Preferred brightness and contrast values were expected to be relatively consistent throughout the survey. Both are expressed as the distance from the centre, or the preferred value for each respondent. First, the difference was calculated between the preferred brightness and contrast values (chosen in response to the first two questions of section II of the survey) and the contrast or brightness values chosen in the further questions. Second, within a given question, the difference was determined between the sample image "closest" to the initially chosen brightness and contrast values and these values in the respondent's actual answer.

2.3 Results and Discussion

In total, 197 respondents completed the survey. Of these, 11 failed the Ishihara test, indicating colour blindness. Less than 6% of respondents were likely to have been colour blind, as this percentage corresponds with averages for populations of Northern European origin. Among them, the frequency of red-green colour vision defects is around 8% for men and 0.5% for women (Deeb, 2005). All responses for respondents who could have been colour-blind were included in the further analysis.

2.3.1 Preferred Brightness and Contrast Ranges

Figure 2.3 presents preferences for brightness and contrast modification levels applied to false colour composite images. The central zero values correspond to the default, unprocessed images as delivered by the MS. Note that based on a preliminary comparison, the range of contrast levels of modification was set to twice the range of brightness modification levels. More than half of the participants preferred the default brightness setting (zero adjustment). On average, there was a tendency to favour slightly reduced brightness settings (–

4.2). There was also a slight preference for somewhat increased contrast levels (3.9).

The spread of the preferred brightness and contrast indicate clear variation between participants in their preferences. This suggests that LPIS operators should be enabled to personalise brightness and contrast settings for visual image interpretation. This finding should be taken into account in the LPISQA technical guidelines for delivering georeferenced map images.

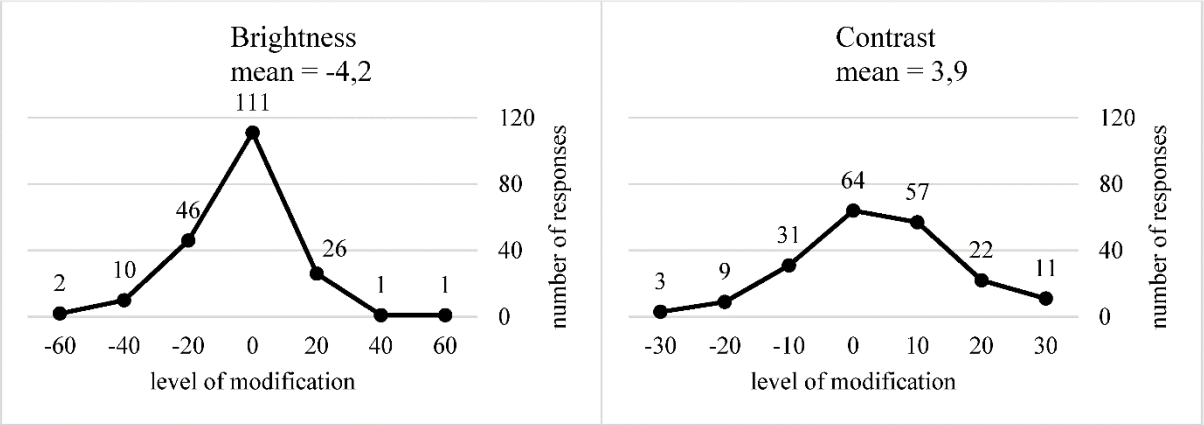


Figure 2.3. Preferred modification levels for brightness (Left) and contrast (Right).

2.3.2 Influence of Background Colour on the Image Brightness and Contrast Appreciation

Figure 2.4 presents responses to three pairs of questions (A, B and C) in which each pair depicts the same sample images on, respectively, a white background and a black background. Note that Pair A compares brightness values only, as the contrast value was set at default (0) for all of the image samples presented. Pair B compares contrast values only, as here the brightness value was left at default. Against the white background, the image chosen as best was the one with the higher contrast (set at 20). Against the black background, the most appreciated image was the one with contrast set at 0. Pair C compares images with default values with those with modified brightness and contrast.

Figure 2.5 presents the entropy and its standard deviation (see Equation (2.2)) times the 97.5% quantile of the student distribution with 196 degrees of freedom across all respondents. Each question asks users to indicate the best and the worst image, first against the white background and then against the black background. The entropy was greater for the choice of best image than for the worst one, meaning that there was more consensus on the choice of the

worst image. In contrast, the standard deviation of entropy was greater for the choice of the worst image, meaning that the entropy value is less precise. The background colour (white or black) did not significantly influence the choice of best and worst images.

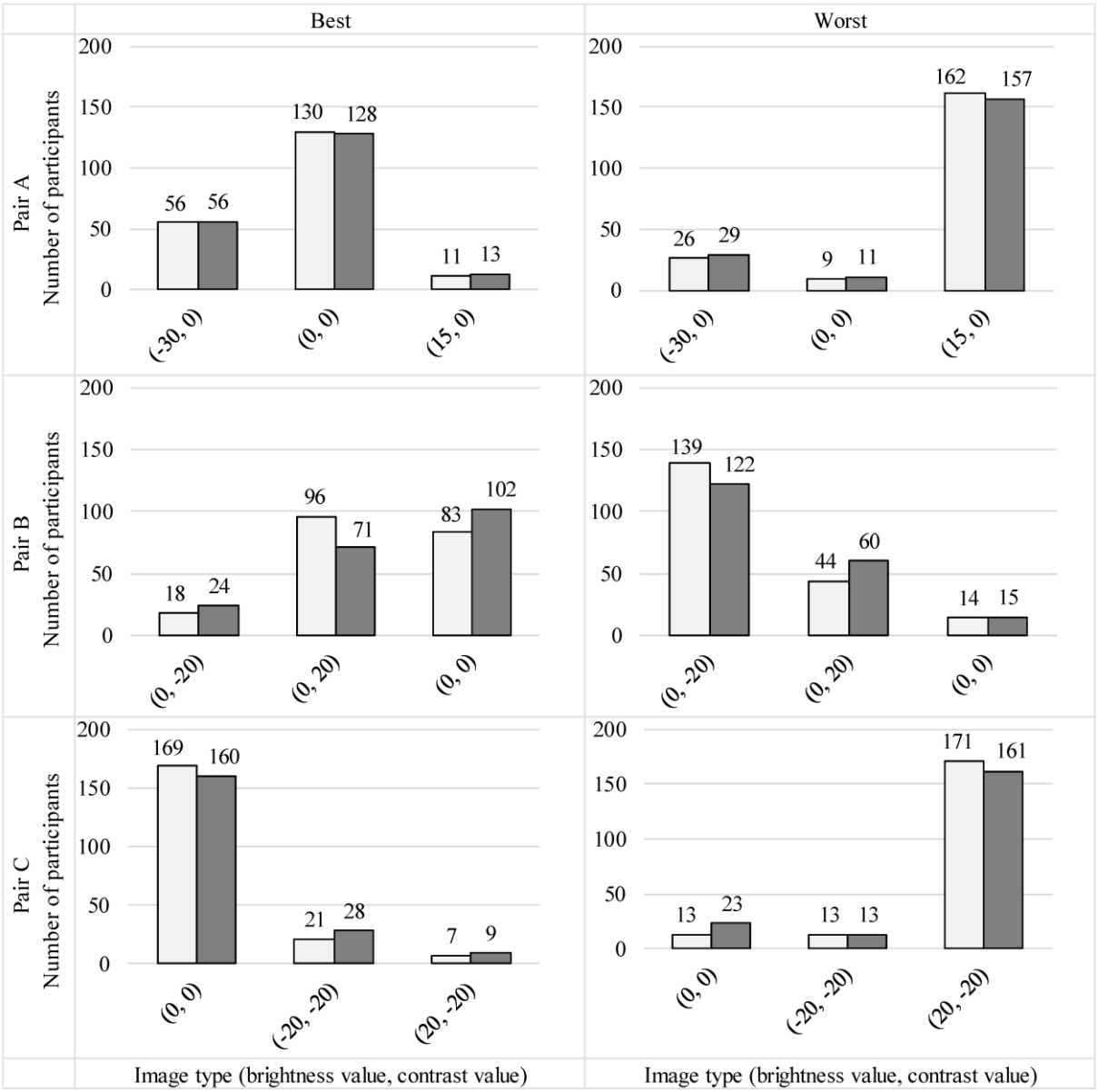


Figure 2.4. Respondents' choice of the best and the worst images against the white and the black background (the brightness and contrast combinations used in the sample images are specified). Figures A, B and C each represent a pair of questions using the same sample images against a different background colour (best on the left and worst on the right; white column = white background, black column = black background).

The entropy value was highest for Pair B, which means that a less consistent preference was expressed for a specific image. In this Pair, the least appreciated image against both background colours was the one with contrast reduced to −20. However, the high entropy values indicate that respondents' choices were

rather dispersed. The lower entropy values found here indicate more consistency in the choices of best and worst images against a white background. Entropy was highest for the questions presenting samples against a black background, and there was a more outspoken preference for a higher contrast against a white background.

These results suggest that preferences of image brightness and contrast may vary depending on the background colour used. Again, this suggests that some means of personalising brightness and contrast should be made available to operators involved in LPISQA visual image interpretation.

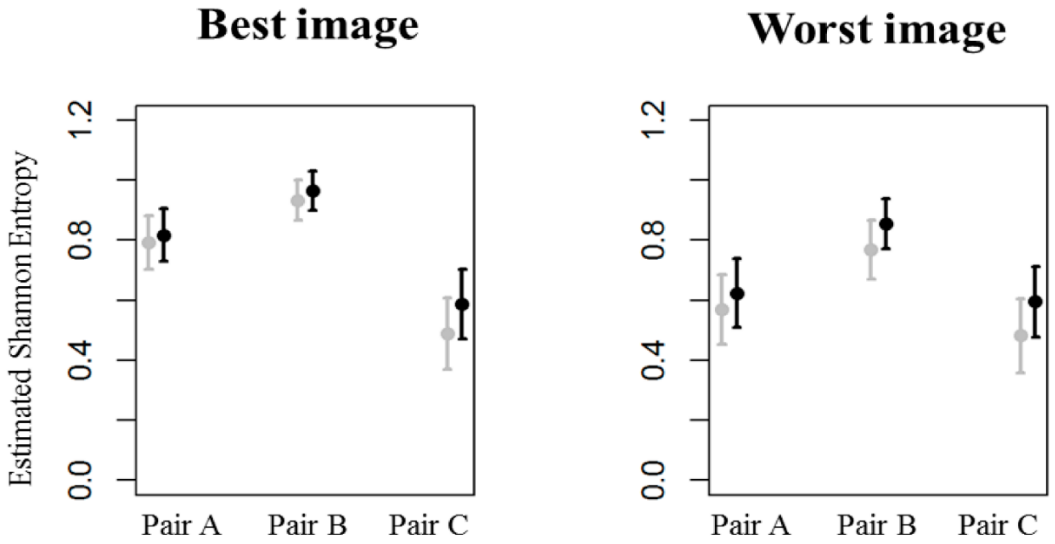


Figure 2.5. Estimated Shannon entropy (points) and its standard deviation times the 97.5% quantile of the Student distribution with 196 degrees of freedom (whiskers) for choices of the most and the least appreciated image triplets, paired with the white background (grey line) and black background (black line). The pairs of questions referred to as A, B and C are the same as those in Figure 2.4.

2.3.3 Image Format and Colour Composite Appreciation

For questions on image format and colour composite, Table 2.3 presents the normalised index of the most appreciated images, calculated using Equation (2.3). The most appreciated image type was a false colour composite in TIFF format. The least appreciated type was a natural colour composite in ECW.

When the image chosen as best was a natural colour composite (in either ECW or TIFF format), the majority of respondents (72% and 85%) selected the false colour composite as the worst image (Figure 2.6). This means that in the context of the survey, appreciation of colour composite was more decisive in determining overall appreciation than the image format type. However, when the image chosen as best was a false colour composite TIFF, the ECW was

selected as the worst sample image. These results imply a strong recommendation for LPISQA administrators to acquire or order images in at least four bands –visible (red, green and blue) and near infrared– to allow production of false colour composites. Beyond having four (or more) bands, operators should be given a means to change the bands displayed.

Table 2.3. Normalised index of most appreciated images from survey items testing image format and colour composite preferences (FCC = false colour composite, NCC = natural colour composite, TIFF = tagged image file format, ECW = Enhanced Compression Wavelet).

Colour Composite/ Format	Normalised Index of Most Appreciated Images			Average	Median
	Question Comparing Three Image Samples	Question Comparing Two Image Samples in NCC Format	Question Comparing Two Image Samples in TIFF Format		
FCC/TIFF	0.64	–	0.86	0.75	0.75
NCC/TIFF	0.24	0.98	0.14	0.45	0.24
NCC/ECW	0.12	0.02	–	0.07	0.07

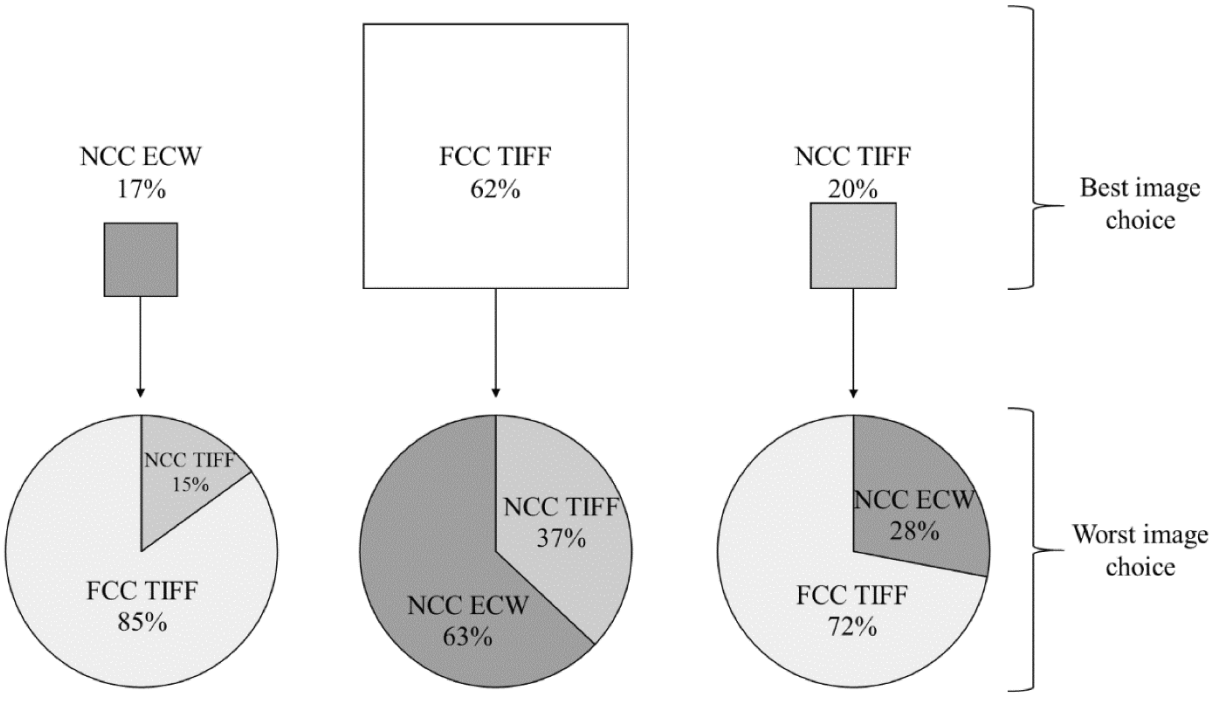


Figure 2.6. Influence of best image format choice on the subsequent worst one (FCC = false colour composite, NCC = natural colour composite, TIFF = tagged image file format, ECW = Enhanced Compression Wavelet).

2.3.4 Images with Shadowed Areas

One of the biggest challenges in visual image interpretation and mapping is retrieving information from areas obscured by shadows. One of the survey questions included a parcel boundary that was partially shaded. After manipulation of one of the images (application of a 2% stretch), the shadowed border was no longer distinguishable (please consult section IV, Question 7 of the online survey available as ancillary data).

Table 2.4 presents respondents' choice of the best image for two questions in which one of the images was not stretched and a 2% stretch was applied to the second. In both, the image chosen as best was the one with the 2% stretch, even if this implied that the parcel border was blurred by shadow. A possible explanation for this rather counterintuitive result is that respondents did not really consider the intended use of the image, which was indication of the land represented by the Reference Parcel.

Table 2.4. Percentage of respondents that chose the respective image samples (no stretch versus 2% stretch) as best for the purpose of delineating agricultural land.

Image Sample	Images without High Objects nor Shadows	Images with Part of a Border Obscured by Shadow
No stretch	24	28
2% stretch	76	72

The LPISQA methodology already incorporates an angle of view restriction to minimise areas with occlusions. Although the effect of shadow is recognised, there are no measures as yet addressing shadow length and information loss.

2.3.5 Standard Image Enhancement Appreciation

Image enhancement appreciation was investigated by testing four standard enhancements, ranging from no stretch to square root stretch. Table 2.5 presents the results.

Table 2.6 presents the entropy for the best and worst image choices for the same questions as in Table 2.5. The most favoured enhancement was the 2% stretch, followed by the linear stretch and then the unenhanced image. The least favoured one was the square root stretch. The display enhancement of 2% stretch is commonly used in image processing and GIS software. It was probably familiar to the respondents and perhaps therefore most appreciated.

Table 2.5. Frequency of image samples being chosen as best (%). Each question (I, II and III) offered three images from which to choose.

Image Sample	Frequency of Being Chosen as Best Image (%)		
	I	II	III
No stretch	25	33	9
Linear stretch	71	–	27
2% stretch	–	54	63
Sq root stretch	4	13	–

Table 2.6. Estimated Shannon entropy of responses to the three questions referred to in Table 2.5.

	I		II		III	
	Best	Worst	Best	Worst	Best	Worst
Estimated Shannon entropy	0.72	0.75	0.97	0.82	0.86	0.60

2.3.6 Consistency of Brightness and Contrast Preferences

Respondents' brightness and contrast preferences were expected to be relatively consistent throughout the survey. This was measured by examining how far from the "centre" respondents' answers were; that is how similar they were to the brightness or contrast values initially indicated as preferred. First, the difference was calculated between the preferred brightness and contrast values (chosen in response to the first two questions of section II of the survey) and the image contrast or brightness chosen in the further questions. Second, within a given question, the difference was determined between the sample image closest to the initially chosen brightness and contrast and these values in the respondent's actual answer. This is depicted on the horizontal axis of Figure 2.7. These values thus show the difference between the closest possible answer in a given question and the preferred brightness or contrast. The size of the point marker in Figure 2.7 reflects the number of times a choice was made. The grey dashed line indicates the theoretically most consistent choice.

Figure 2.7A's preferred brightness value is well represented in answers to the further questions (in section III of the survey). Moreover, this same preference at individual level was chosen again in later questions in the majority of cases. For brightness, respondents' second preference (when the favourite was not available) was slightly less bright. The most frequent answers are concentrated around the dashed grey line, confirming a high consistency in choices. Figure 2.7C shows that the least preferred brightness is quite dispersed, and the images selected as worst differ markedly from the preferred brightness. Furthermore,

the least preferred brightness values are higher than the brightness most preferred.

Figure 2.7B,D show similar plots for preferred contrast values. Throughout the survey, the preferred contrast is largely consistent with that chosen as preferred in the second question of section II (indicated in Figure 2.7B as the largest value at the 0,0 intersection). We see on closer inspection of Figure 2.7B that sometimes, even when the preferred contrast value was represented in the question, a higher contrast value than the preferred one was likely to be chosen.

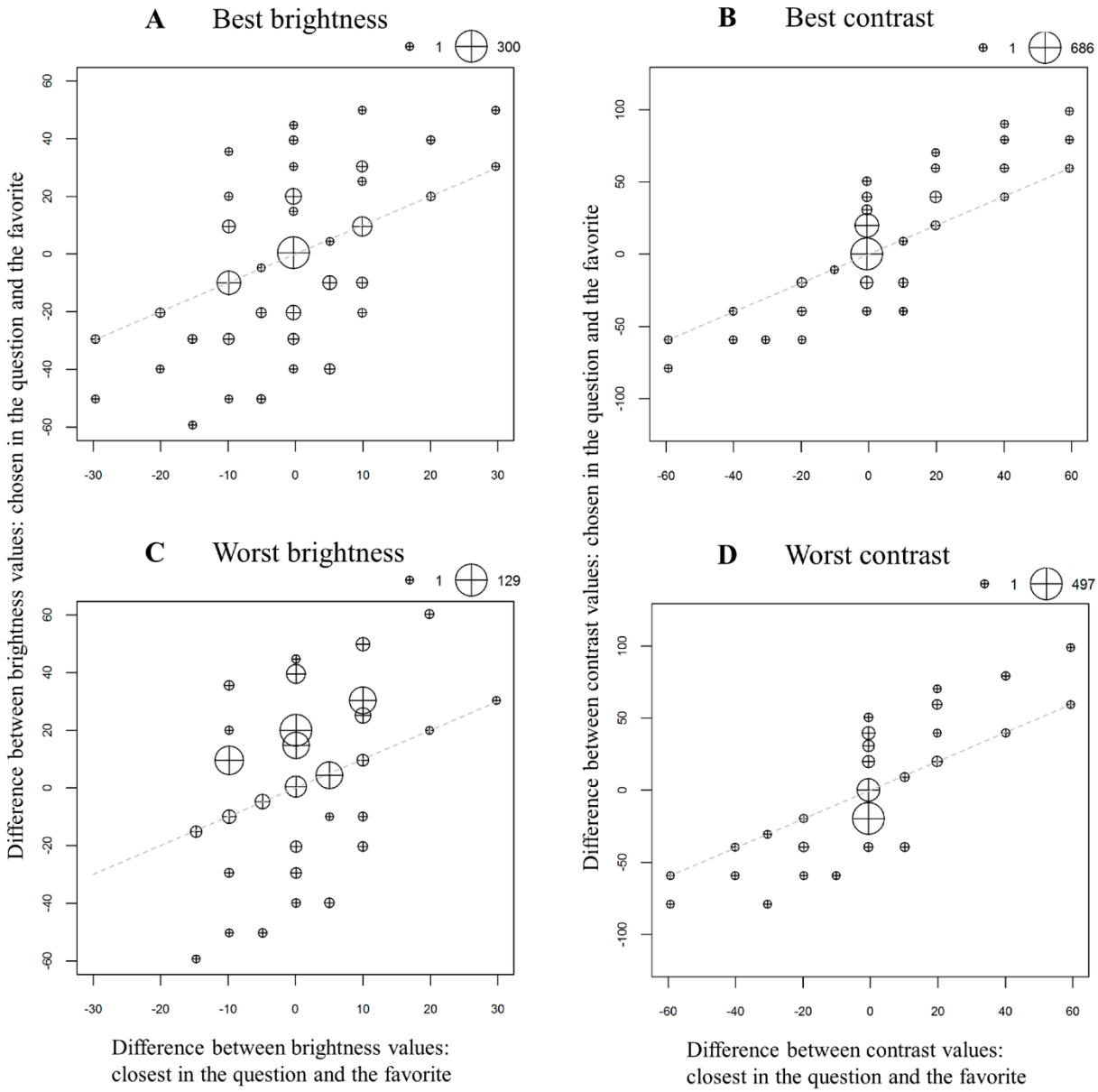


Figure 2.7. Number of choices (indicated by the marker size) for most and least preferred brightness and contrast. The grey dashed line indicates the theoretically most consistent choice possibilities.

To summarise, the preferred brightness values were not always confirmed in the further questions. Nonetheless, there was substantial consistency in the brightness chosen as preferred. When the preferred brightness value was not represented, a lower brightness value was typically selected. Higher brightness values were less appreciated. There was much consensus among respondents on the preferred contrast values. Moreover, in the majority of cases, the preferred contrast was chosen again and again in the further questions. Otherwise, images with higher contrast values were preferred over those with lower contrast.

These results demonstrate that image preferences are relatively consistent for an individual respondent. However, they do vary within the group of respondents. This again suggests the merit of providing LPISQA operators a means to personalise settings and make photometric adjustments in images.

2.4 Conclusions and Recommendations

This study confirms that brightness and contrast are both important attributes of LPISQA images. Users generally prefer higher contrast values combined with lower brightness values. Where the background colour for the image comparison is white, an even higher contrast value is preferred. Our findings, furthermore, reveal variety among users in their preferred brightness and contrast, although individual respondents exhibited a high degree of consistency in the choices they made. This suggests the usefulness of providing individual operators with a means to personalise image settings. Respondents to our survey were quite consistent in choosing a specific brightness and contrast value. Among respondents, there was also overall agreement on the least preferred images, though there was less consensus on the most appreciated ones.

The false colour composite was revealed to be preferred over the natural one. This finding should encourage MS to order four-band images (with visible bands and near infrared) rather than limiting images to only the visible bands. The preferred image format was TIFF, which was favoured over the lossy ECW.

Considering the standard stretches and image enhancements, the 2% stretch was preferred. This popular display enhancement is commonly applied in image processing and GIS software and therefore was probably familiar to the respondents. Where an image contained a shadowed area, loss of information (in the shadow) appeared less important to the respondents than the overall

image appearance, as the image sample chosen as best was one with a higher contrast and lower brightness, though this rendered the land use delineation indiscernible. This finding suggests that there is still a role for expert input in designing image quality standards for shaded areas. Images with shadows require further investigation, as it is crucial to find an optimal balance between visual appreciation of an image and loss of information.

The results of our survey indicate that, for the purpose of agricultural land delineation, the best image format is a TIFF, false colour composite, with the enhancement of a 2% stretch and providing operators a means to adjust brightness and contrast to their own individual needs. Further research should focus on evaluation of existing metrics for assessing the photometric image quality (Pyka, 2009) for LPISQA objectives, or if needed, the design of new metrics.

Acknowledgments

The authors thank Augusta Bande and Romuald Franielczyk from the European Commission Joint Research Centre (JRC), Institute for Environment and Sustainability (IES), Monitoring Agricultural Resources Unit for creating and maintaining the online survey.

Chapter 3

Comparison of manual and automated shadow detection on satellite images for agricultural land delineation

Agnieszka Tarko, Sytze de Bruin, Arnold K. Bregt

This chapter is based on:

Tarko, A., de Bruin, S., Bregt, A.K., 2018. Comparison of manual and automated shadow detection on satellite imagery for agricultural land delineation. Int. J. Appl. Earth Obs. Geoinf. 73, 493–502. https://doi.org/10.1016/j.jag.2018.07.020

High resolution figures can be found in the online publication.

Abstract

Land cover identification and area quantification are key aspects in determining support payments to farmers under the European Common Agricultural Policy. Agricultural land is monitored using the Land Parcel Identification System and visual image interpretation. However, shadows covering reference parcel boundaries can hinder effective delineation. Visual interpretation of shadows is labour intensive and subjective, while automated methods give reproducible results. In this paper we compare shadow detection on satellite imagery obtained by visual interpretation to a proposed automated, data-driven method. The latter automated method is a thresholding approach employing both panchromatic and multispectral imagery, where the former has a finer spatial resolution than the latter. Thresholds are determined from automatically generated training data using a risk-based approach. Comparison of the total shadow area per scene showed that more pixels were labelled as shadow by the automatic procedure than by visual interpretation. However, the union of shadow area independently identified by twelve interpreters on a subscene was larger than the automatically determined shadow area. The limited intersection of the shadow areas identified by the interpreters demonstrated that interpreters strongly disagreed in their interpretations. The shadow area labelled by the automated method was in between the intersection and the union of the areas interpreted by interpreters. Furthermore, the automated shadow detection method is reproducible and reduces the interpretation effort and skill required.

Keywords

photointerpretation; risk-based classification; data-driven approach; Land Parcel Identification System

3.1 Introduction

Shadows are present on the majority of remotely sensed images, and their presence can affect information abstraction. In image interpretation, shadows can be indicative of land morphology or a feature's height and shape (Lillesand et al., 2008). Yet, shadows can complicate delineation of agricultural lands, hindering monitoring programs. A shadow covering a boundary of a reference parcel can affect delineation, impeding the updating of field geometries (European Commission, 2014a; LPIS TG ETS, 2017b). Since some agricultural subsidies are area-based, delineation of parcels may impact farm subsidies (Astrand et al., 2004), worth some €50,000 million in 2017 (European Commission, 2014e). Identification of shadows on images is therefore of key interest.

An example of a voluntary monitoring approach is the Land Parcel Identification System (LPIS) implemented by the European Union (EU) Member States (European Commission, 2013). So-called reference parcels are delineated on the basis of very high resolution (VHR) images in the scope of the Control with Remote Sensing program (LPIS TG ETS, 2017a). Along with other sources such as farmers' declarations, VHR satellite images are used not only for LPIS updating but also for quality assurance of the system through the annual Executable Test Suite (ETS) (European Commission, 2014a; LPIS TG ETS, 2017b). Such testing is performed by EU Member States using a limited number of VHR satellite images; this involves re-delineation of a sample of reference parcels. Shadows may influence the reference parcel boundary re-delineation, impacting ETS inspection. If too many parcel boundaries are masked by shadows, both sample randomness and sample size may be jeopardised, influencing the ETS results. Detection of cast shadows in a satellite scene can help to determine if available imagery is usable for the reference parcel monitoring process. Beyond the LPIS in the EU (IACS, 2017), knowledge of cast shadows is useful for other agricultural monitoring programs as well, such as those implemented by the US Farm Service Agency (FSA, 2017) and the Chinese GIS-based land registry system (Rabley and Yuen, 2009).

Traditionally, manual interpretation of agricultural systems based on aerial or satellite images has sought to delineate field boundaries, including those partly hidden by shadows on an image. To pinpoint reference parcel boundaries, image enhancement is performed and auxiliary data is used, such as additional images, maps and field visits. Although manual mapping performed by interpreters is subjective and has a low reproducibility (Van Westen, 2000), it is nonetheless still widely applied.

The results of manual mapping of shadows has been compared with the results of automated methods. In many such exercises, a manually digitised shadow mask serves as a reference to assess or test proposed shadow detection methods (Adeline et al., 2013; Tsai, 2006). As reviewed by Adeline et al. (2013) and Shahtahmassebi et al. (2013), there are various automated methods for shadow detection on a digital image. Some deterministic methods require additional input data, such as a digital surface model for model-based geometrical shadow detection methods (Adeline et al., 2013; Li et al., 2005). Physically based methods require additional information on the atmosphere and acquisition details (Adeline et al., 2013). Other methods, like machine learning, require users to input data such as training areas (Adeline et al., 2013). Invariant colour model methods rely on RGB channels (for the colours red, green and blue) (Adeline et al., 2013; Tsai, 2006).

Histogram thresholding is among the most popular methods for automated shadow detection due to its simplicity, independence from auxiliary data and good overall performance. However, setting a proper threshold is an issue with these methods. Yamazaki et al. (2009) proposed threshold determination by visual interpretation. Dare (2005) identified the threshold as the mean value between two peaks of a panchromatic (PAN) band histogram, while Otsu (1979) proposed an automatic threshold estimator for grey-level images. The best thresholding results were obtained using first valley detection on Nagao's modified intensity (Adeline et al., 2013; Nagao et al., 1979). However, the valley detection algorithm fails if intensity histograms lack a bimodal distribution (Nagao et al., 1979).

Most automated methods have focused on urban zones and city canters with high buildings and urban valleys (Adeline et al., 2013; Dare, 2005; Li et al., 2005; Sarabandi et al., 2004; Tsai, 2006; Yamazaki et al., 2009). An exception is the use of aerial imagery of an alpine terrain in Central Taiwan in which shadows were identified by first valley detection thresholding using Nagao's modified intensity (Wu et al., 2014). Another exception concerns aerial imagery covering a rural area in Italy on which RGB band spectral ratioing and Otsu's threshold finding method were used (Movia et al., 2015).

While there are several automated methods for shadow detection on remotely sensed imagery, they either require user interaction for threshold detection, or depend on detailed ancillary data, such as a digital surface model. Such data may not be available for many rural areas. Therefore, this study proposes a relatively simple thresholding method for shadow detection with limited ancillary data requirements. The method is compared to manual visual interpretation of shadows in the context of agricultural land delineation.

The aim of the comparison is to assess whether the results of the proposed automated procedure for shadow detection are similar to those of visual interpretation and whether the former are suitable for quick labelling of image scenes that, due to too much shadow, have limited usability for agriculture monitoring. The objective of this paper is twofold: (1) to propose a reproducible method for shadow detection on satellite images using a readily available auxiliary training area and (2) to compare the method with interpreter manual interpretation of shadows. While shadow detection can be used in a processing chain prior to image enhancement operations, the latter are beyond the scope of this paper.

3.2 Methods

This chapter first describes the manual method, then explains the automated procedure and our case-study. Finally the comparison method is briefly presented.

3.2.1 Manual method

Twelve interpreters independently digitised shadows on an agricultural study area. No specific instructions for the visual interpretation were provided and the interpreters were free to choose their preferred software for image display (including image enhancement) and digitising. Following Tarko et al. (2015), the interpreters were asked to set their own preferred image enhancement, which could be adapted while interpreting (e.g., colour stretching). Shadows could be digitised as polygons in a vector layer or by labelling pixels in a raster. The used zooming level was left to the interpreter's discretion; for the used input data the typical mapping scale would be 1:100–1:1000 The interpreters were encouraged to use widely accessible open-layer images (such as Google Earth, Google Maps and Bing Maps) to assist in identification of potential shadows (these are referred to as auxiliary data). The provided data were PAN images with 0.5 m pixel size and a polygon indicating the area to be interpreted (see green frame in Figure 3.2). Multispectral (MS) imagery were not included, because the 2 m pixel size of MS bands was deemed too coarse for visual identification of shadow boundaries. Any received vector layer was rasterized. Rasterizing error (Bregt et al., 1991) was found to be within 0.1% of the shadow area. The intersection and the union of the shadow layers produced by the individual interpreters were also computed. Apart from the shadow detection on the smaller subscene done by the twelve interpreters (including the first author), the first author also manually digitised shadows on all scenes tested in the automated method.

3.2.2 Automated method

Our method employed thresholds adjusted to an acceptable rate of erroneously labelled shadows, determined using a minimum set of training data. Section 3.2.3 describes a case study in which such data were acquired without user interaction. Two thresholds were applied: one for VHR PAN images and the other for high resolution (HR) MS images. For the former, based on a PAN version of the training data, the darkest objects in a scene were identified using the PAN spectrum. Regarding the latter, while MS training data have coarser resolution, they contain spectral information for a selected spectrum part holding information about land cover types (Belgiu et al., 2014). Moreover the near infrared (NIR) band spectrum is beyond the PAN spectrum. To enable use of a single threshold on the MS data, the dimensionality of the MS bands had to be reduced. A potentially suitable choice for this operation is principal component analysis (PCA). In our approach, the first principal component (PC1) was computed on the covariance matrix of the training data; it was enforced to be positively loaded on NIR. While this may suggest that a single threshold could be applied to the NIR band, PC1 was chosen because it is more discriminative for bare soil, which commonly occurs in imagery acquired early in the growing season and in areas affected by drought. Hence, a threshold was determined on the training area and applied to the area of interest.

The intersection of shadows labelled in two versions of the scene (PAN and MS) allowed the integration of the finer spatial resolution of PAN, resulting in sharper detail, while shadow confirmation from the PC1 allowed shadow detection on vegetated land beyond the PAN band wavelengths. Figure 3.1 presents a flow diagram of the overall procedure.

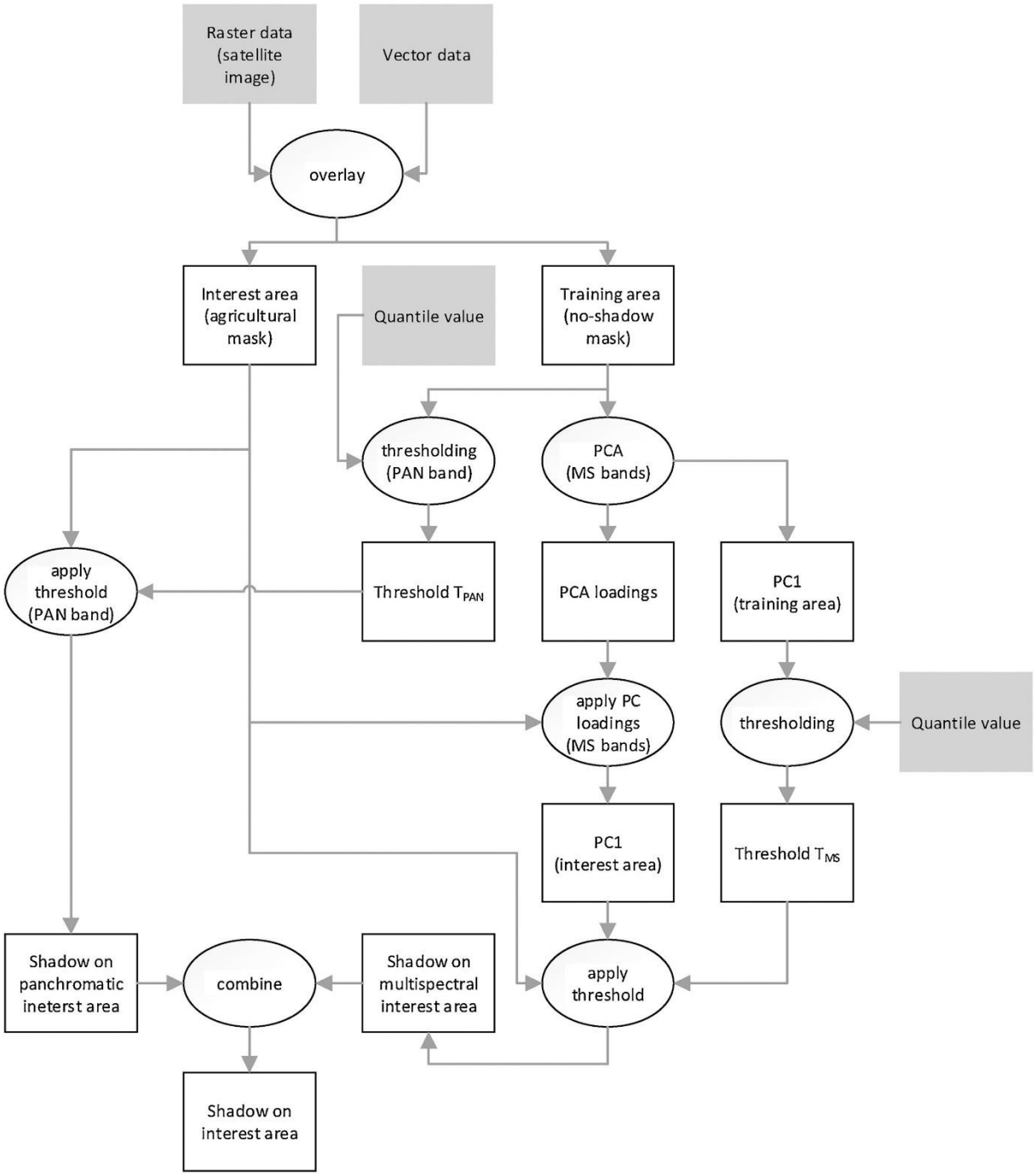


Figure 3.1. Main steps in automated shadow detection. Inputs required from the user are indicated by grey boxes. PAN = panchromatic, MS = multispectral, PCA = principal component analysis, PC1 = first principal component.

3.2.2.1 Determine thresholds

The first step was preparation of the vector layer delineating approximated agricultural land that was deemed shadow free. Training data were obtained by overlaying the imagery and the vector layer. The thresholds were determined from the training data using an acceptable risk level. A relatively low risk level of 5% was applied. This choice was made taking into account the quality of the

used images. Our imagery was acquired under favourable light and viewing angle conditions, and therefore the amount of shadow on the scenes was expected to be minimal. However, the input training data were not perfect. The 5% quantile was considered a suitable compromise.

To obtain a single threshold TPAN for the VHR PAN images, a threshold corresponding to the selected quantile on the histogram of the training pixels from the PAN version of the scene was set. To obtain a single threshold TMS for HR MS images, the threshold corresponding to the chosen quantile was determined using the PC1 scores of the training data. To ensure that high PC1 values corresponded to high brightness, a positive loading to the NIR band was enforced to exclude dark vegetation from the potential shadows in the training areas. If the loading for the NIR band swapped sign, the signs for the other bands were swapped as well. The PC1 loadings computed from the training data were stored since they were to be used on imagery of the area of interest.

3.2.2.2 Apply thresholds

The threshold TPAN was applied to the VHR PAN image of the area of interest. It resulted in a binary image indicating potential shadows at a fine resolution. The threshold TMS was applied to the PC1 scores computed for four HR MS image bands using the PC1 loadings computed from the training data. Application of the threshold TMS to an interest area resulted in a binary image indicating potential shadows with coarser resolution. Finally, the two binary images were combined; that is, their union was our final result.

3.2.3 Case study

World View 2 (WV2) images acquired for the 2015 Control with Remote Sensing campaign (LPIS TG ETS, 2014) were used for our tests. The tests concerned four different zones of approximately 271 ha each in Belgium (Flanders) and Lithuania. Images had a PAN band with 0.5 m spatial resolution and MS bands (NIR, R, G, B) with 2.0 m resolution. The images were delivered with 11 bits resolution (represented in 16 bits format), in Ortho Ready Standard, level 2 A (Digital Globe, 2017), meaning that the pixel values were converted to absolute radiometry and that the products were sensor corrected, and were projected to a constant base elevation, which is calculated on the average terrain elevation per order polygon (Digital Globe, 2017). Each zone was divided into equally sized frames (left and right). Thus, the automated method was also tested on image subsets representing different land cover compositions (indicated by blue and red frames in Figure 3.2). Scenes were selected such as to contain various land cover classes (such as agricultural land, built-up area, trees or forest) as well as water bodies and wet soil, as these are

often confused with shadows (Adeline et al., 2013; Dare, 2005; Li et al., 2005). Areas with dense vegetation and man-made constructions were also pictured.

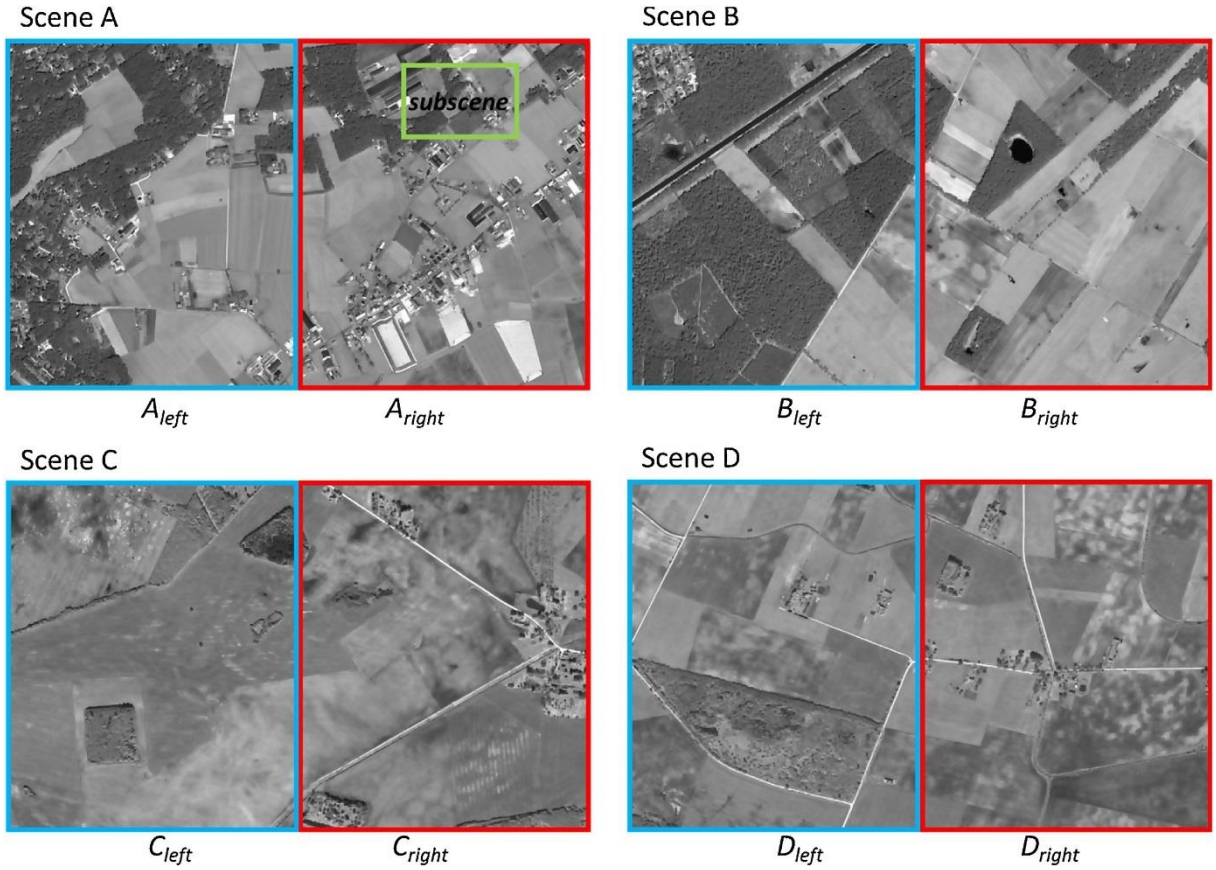


Figure 3.2. Test scenes (presented in PAN band): left part of the scene (blue frame) and right part of the scene (red frame); image subscene used as a reference for the manual method (green frame). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

To assess differences between the vsiual interpretations, a small subscene in Belgium was chosen (11 ha, green frame in Figure 3.2). Because visual interpretation is tedious, we limited the area so that it could be finished within 1–2 h (Table 3.3). The subscene included similar land cover types as present in the full scene.

Figure 3.3 presents a flowchart of the required input data preparation.

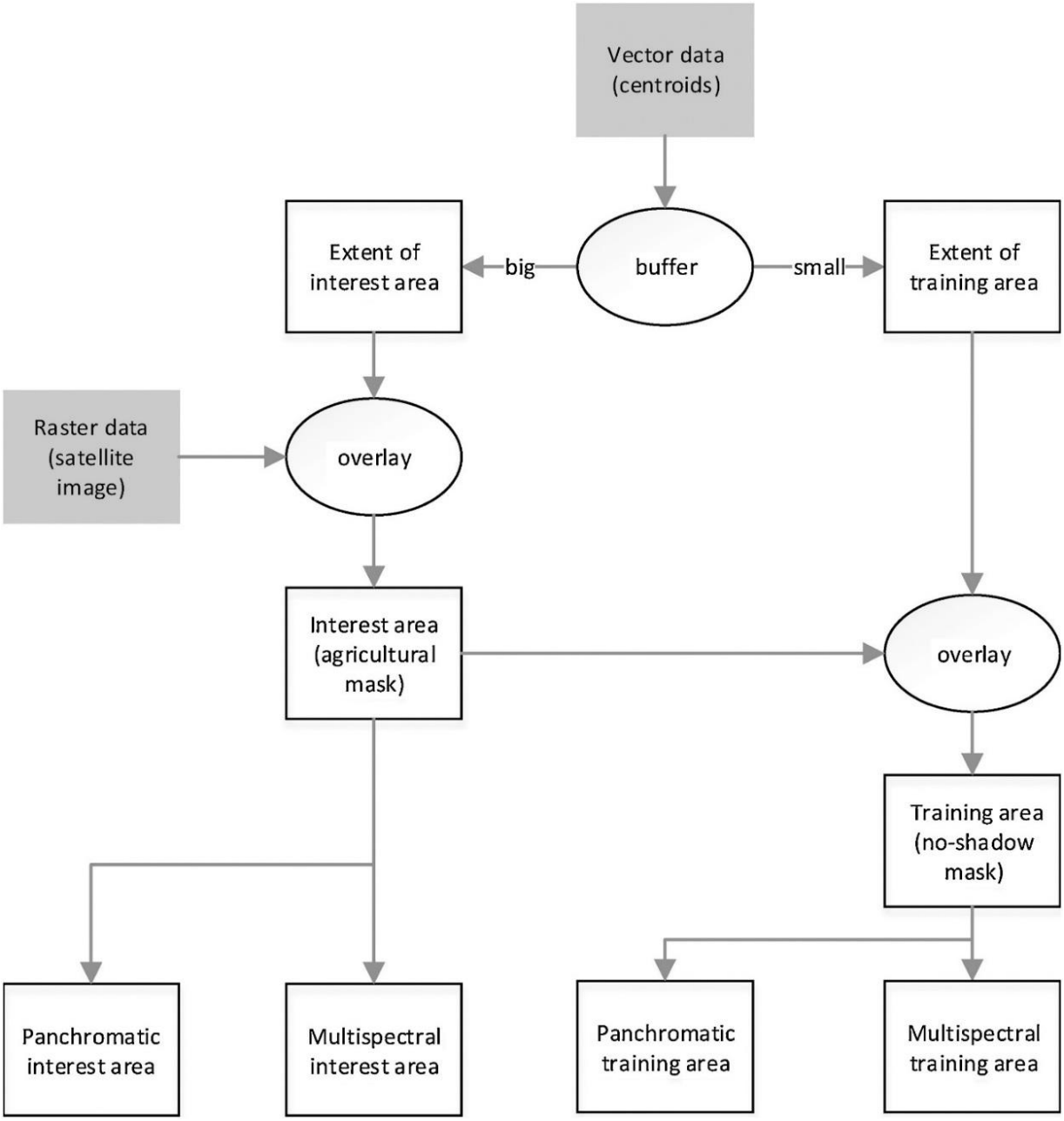


Figure 3.3. Input data preparation. Inputs required from the user are indicated by grey boxes.

To produce a dedicated, virtually shadow-free dataset centroids indicating reference parcels with agricultural land were used. Such centroids are delivered to the EU Joint Research Centre (JRC) by EU Member States indicating reference parcels in the LPIS population for LPIS quality assurance (QA) inspection. Our data was delivered to the JRC in reporting year 2015 (LPIS TG ETS, 2014). Based on a buffer built around the centroids, a dedicated vector input data representing (1) the extent of the training area and (2) the extent of the interest area were created. A buffer was built around point data, with each centroid point having an attribute value stating the area of the agricultural land

it represented. When approximating the buffer size for the training area extent, the possibility of including non-agricultural land inside the training area was minimised and inclusion of field boundaries was avoided, since that is where shadows are likely to occur. When approximating the buffer size for the interest area, two extreme scenarios for the unknown field shape were considered: compact and elongated.

The following values were calculated: (1) half of the radius value, derived from the circle area built around the centroid and equal to the reference area value and (2) distance from the centroid to the most adjacent neighbouring one. From these two values, half of the minimum distance was used as a buffer size for computing circular training areas around the centroids representing compact training area shapes (see examples in Figure 3.4). Twice the maximum distance was used as a buffer distance for computing circular interest areas, again using the centroids as input (e.g., in scene D (Figure 3.4) the interest area covered the entire test image).

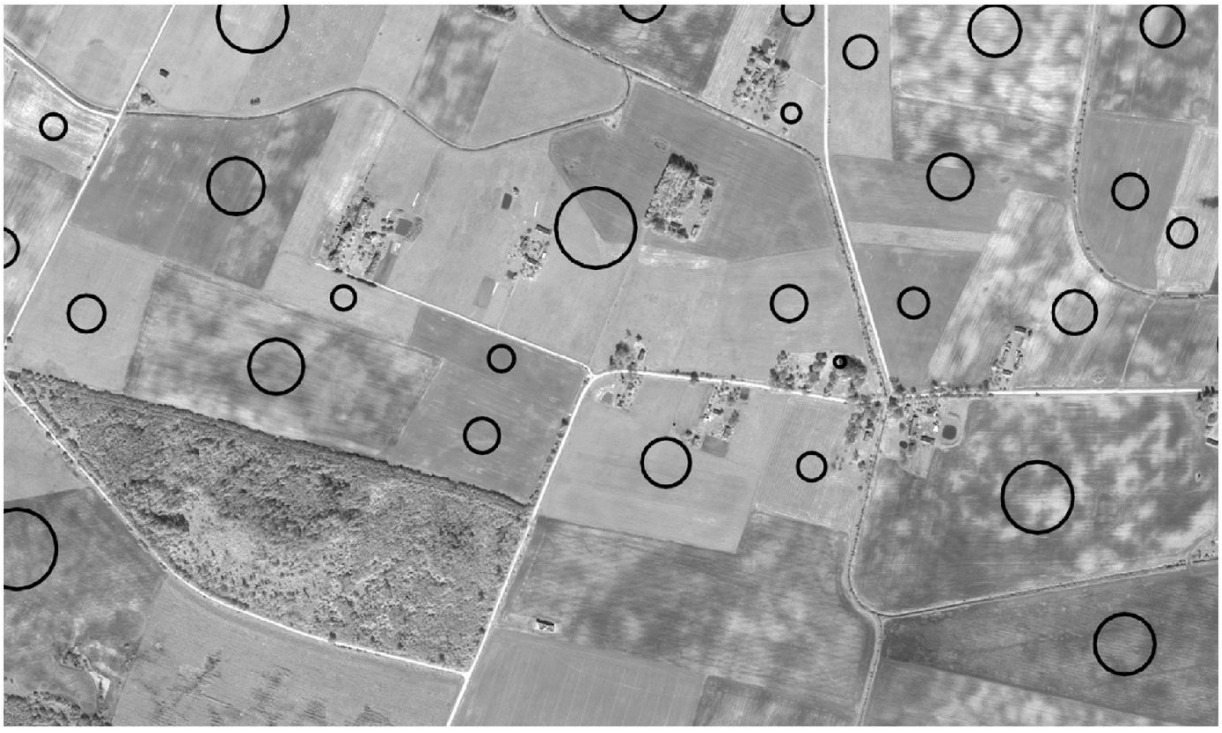


Figure 3.4. Test scene D with training area approximating shadow-free agricultural land (black circles).

Having created the training area and similar to the common alpha level used for hypothesis testing, a 5% alpha level was chosen as an acceptable risk for falsely assigning shadows within the training areas. Note that the 5% acceptance probability of false shadow assignment can be adjusted according to user needs and preferences.

Water bodies were excluded to avoid them being confused with shadows (Adeline et al., 2013; Sarabandi et al., 2004). Identification of water bodies was based on the NDWI index (McFeeters, 2013). Following McFeeters (2013), we set the water index threshold to 0.3. Finally, we removed any no-data occurrences in the interest area.

All steps of data preparation and processing were performed using R software for statistical computing (R Core Team, 2017). The following packages were used: "raster" (Geographic Data Analysis and Modeling, 2017), "rgdal" (Bindings for the "Geospatial" Data Abstraction Library, 2017), "rgeos" (Interface to Geometry Engine - Open Source ('GEOS'), 2017) and "sp" (Classes and Methods for Spatial Data, 2017).

3.2.4 Comparison

The comparison was performed at several levels: (1) for the entire scenes, (2) for half-scenes, (3) for the subscene. For the four entire scenes and for the half-scenes, the results of the automated data-driven method were compared with the results of manual digitisation of shadows (done by the first author). As comparison measures, the total shadow area and the locations of shadow patches were used.

For the subscene, the manual mappings done by the twelve individual interpreters was compared. Moreover, the results of the automated method to the interpreters' collective results were compared. The union of the interpreters' input and the intersection of the interpreters' input were calculated, and these were compared with the automated results for the subscene. Also, locations of shadow patches were visually compared.

3.3 Results

3.3.1 Manual method

Each of the twelve independent interpreters submitted a visual interpretation and digitisation of shadows for the subscene indicated by the green frame in Figure 3.2. The area identified as shadow varied from 0.69% to 5.02% of the subscene (Table 3.1). The interpretations largely disagreed, both in quantity of identified shadows and in the locations and shape of shadow patches (Figure 3.5). Confirming this disagreement is the very small intersection of the shadow cover identified by the different interpreters; this was only 0.24% of the subscene, though the union of the shadow identified by the interpreters amounted to 8.36% of the subscene. The intersection of the shadows digitised by the interpreters was thus less than 3% of their union. Three interpreters did

not digitise shadows cast by high and dense vegetation within the forest land cover. Other interpreters labelled areas with low reflectance, such as dark roofs and water bodies, as shadow.

Table 3.1. Total shadow area identified on the subscene by automated method and interpreters.

Source	Area percentage shadows [% of total area]	Intersection with automated method [% of total area]
automated method	8.27	8.27
intersection of shadows identified by all interpreters	0.24	0.21
union of shadows identified by all interpreters	8.36	3.80
interpreter 1	5.02	2.22
interpreter 2	4.01	2.23
interpreter 3	3.94	2.30
interpreter 4	3.84	1.91
interpreter 5	3.25	2.49
interpreter 6	2.65	1.93
interpreter 7	2.21	1.17
interpreter 8	1.71	1.19
interpreter 9	1.52	1.04
interpreter 10	0.89	0.63
interpreter 11	0.83	0.61
interpreter 12	0.69	0.54

Figure 3.5. Interpreters' shadow visual interpretation on the subscene.

In the visual interpretation of the four entire scenes, done by the first author, shadow varied between 1.63% and 5.68% (Table 3.2). The higher shadow percentage corresponded to areas with dense high vegetation, buildings and uneven terrain (Figure 3.7).

Table 3.2. Comparison of automated and visual shadow detection results.

Scene	Area percentage shadows [% of total area]	
	Automated method	Manual method
A	4.76	4.22
A_{left}	6.03	6.40
A_{right}	4.42	2.04
B	9.40	5.68
B_{left}	13.24	9.55
B_{right}	4.08	1.82
C	3.32	1.75
C_{left}	0.62	1.88
C_{right}	7.56	1.61
D	2.18	1.63

Scene	Area percentage shadows [% of total area]	
	Automated method	Manual method
D_{left}	1.23	2.47
D_{right}	2.26	0.78

The interpreters spend on the effort from 1 up to 2 h to digitise shadows on the subscene (Table 3.3). To manually identify shadows on the entire scenes, the first author needed from 2 up to 4 h, depending on the scene.

Table 3.3. Set-up of manual and automated shadow detection.

	Manual method	Automated method
WorldView2 input image	PAN	PAN, MS
Image setting	up to interpreter discretion	not applicable
Area (ha) of interest (subscene / full scene)	11/–	11/270
Approximate time (minutes) spent on shadow detection (subscene / full scene)	60–120 / 120–240	2/50
Image radiometric resolution	image delivered with 11 bits original resolution (16 bit format); display depends on the chosen GIS software	11 bit original resolution, 16 bit format

3.3.2 Automated method

Automated identification of shadows produced from 2.18% to 9.40% shadow cover per scene (Table 3.2). The higher shadow percentage corresponds to a scene with dense high vegetation, dark roofs and dark soil patches potentially caused by high moisture content. That last will be referred as "wet soil". For scene A, 4.76% of the area was labelled as shadow (Figure 3.6). In the left half of the image (A_{left}), shadow cover was 6.03%. The right half of the image (A_{right}) had a smaller percentage of shadow, 4.42%. For scene B, 9.40% shadow was detected. The left half of this image (B_{left}) produced the highest shadow percentage of all the tested scenes (13.24%). For the right half of this image

(B_{right}), four times less shadow was identified, amounting to 4.08%. In scenes C and D, dominated by agricultural land, the percentage of shadow was 3.32% and 2.18%, respectively. For the subscene, 8.27% of the area was labelled as shadow (see Table 3.1), which is less than the union of the shadow identified by the interpreters. The rightmost column of Table 3.1 shows that for the subscene, 44%–73% of the shadow areas identified by the interpreters overlapped with shadows labelled by the automated procedure. In general, interpreters delineating more shadow area tended to disagree most with the automated procedure (less relative overlap) while shadows identified by interpreters delineating little shadow were largely covered by the automated procedure as well. Approximately 45% of the union of all visual interpretations was also identified by the automated procedure (100% × 3.80/8.36).

The runtime for the automated method was 2 min for the subscene and up to 50 min for the entire scene (Table 3.3) using a PC with an Intel Core 5 CPU at 2.2 GHz and 8.00 GB RAM.

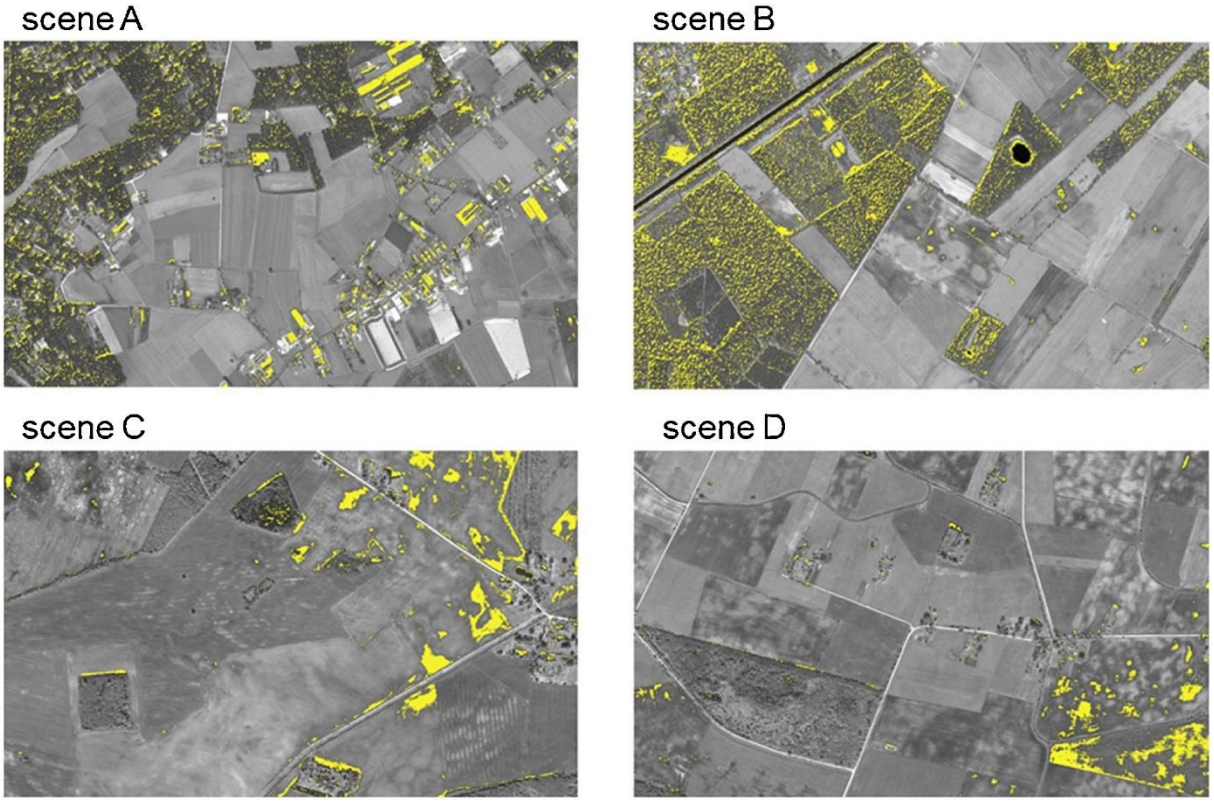


Figure 3.6. Results of the automated shadow detection method (shadows depicted in yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

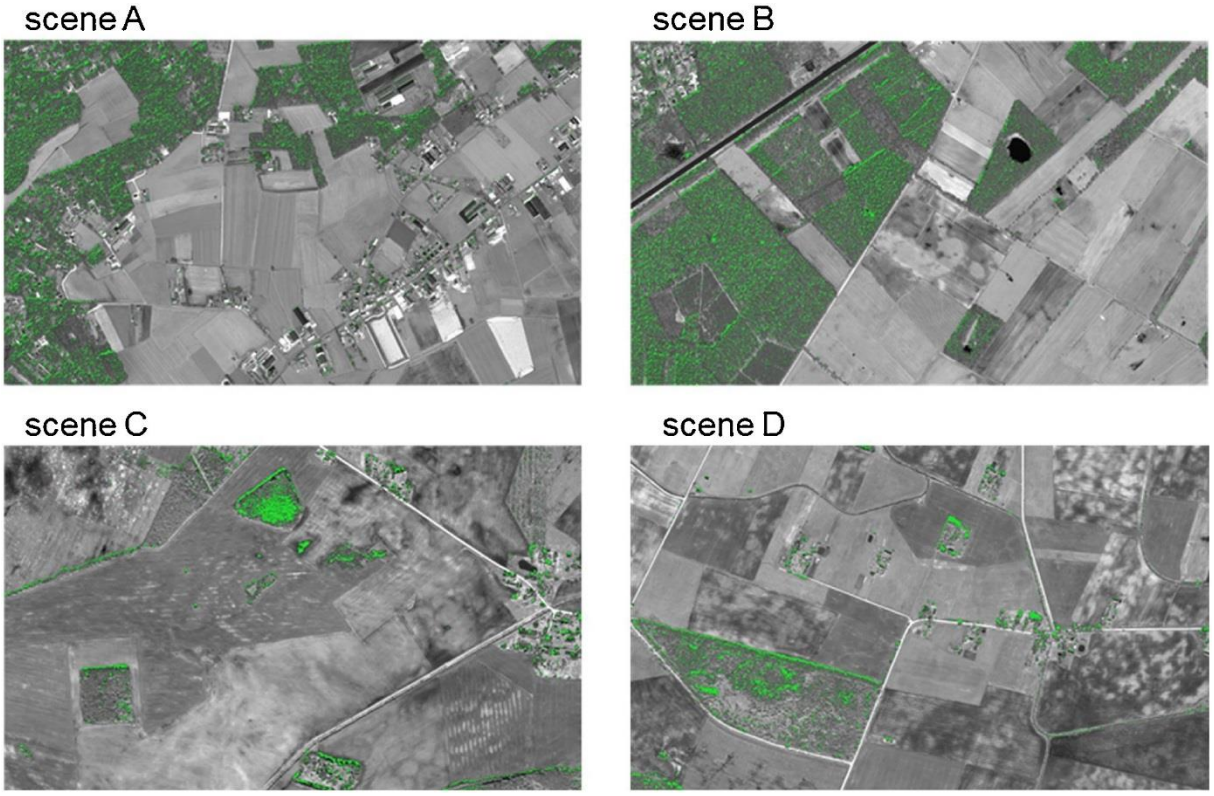


Figure 3.7. Results of manual shadow detection done by the first author (shadows depicted in green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

3.3.3 Comparison

Comparing the automated method with the manual shadow identification for the entire scenes shows that the automated method generally assigned more pixels to the shadow class than the individual who visually interpreted the images (Table 3.1, Table 3.2, Table 3.3). However, in some half scenes (A_{left}, C_{left} and D_{left}), the interpreter labelled more area as shadow than the automated method. Table 3.1 demonstrates that interpretation and mapping of shadows between the interpreters showed substantial disagreement. Hence, an interpreter delineation cannot serve as a single reference map but can only be used for comparing results.

Looking at the shadows identified by the automated and manual method on the entire scenes, it was found that the automated method typically classified more area as shadow in water bodies, wet soils, dark roofs and high vegetation. The automated method typically indicated less shadow than the interpreter on small patches with high and fragmented vegetation (Figure 3.6, Figure 3.7).

In the subscene, the largest differences in shadow designation were concentrated in forest areas with closed high vegetation and in wet soil areas

(rectangle and oval, respectively, in Figure 3.8). In areas with closed high vegetation, the union of visual interpretations was larger than the shadow area labelled by the automated procedure. Wet soil was the only land cover type indicated as a shadow area in the automated method and not by the interpreters. Both methods labelled a shadowed side of a dark roof (arrow in Figure 3.8) as a shadow.

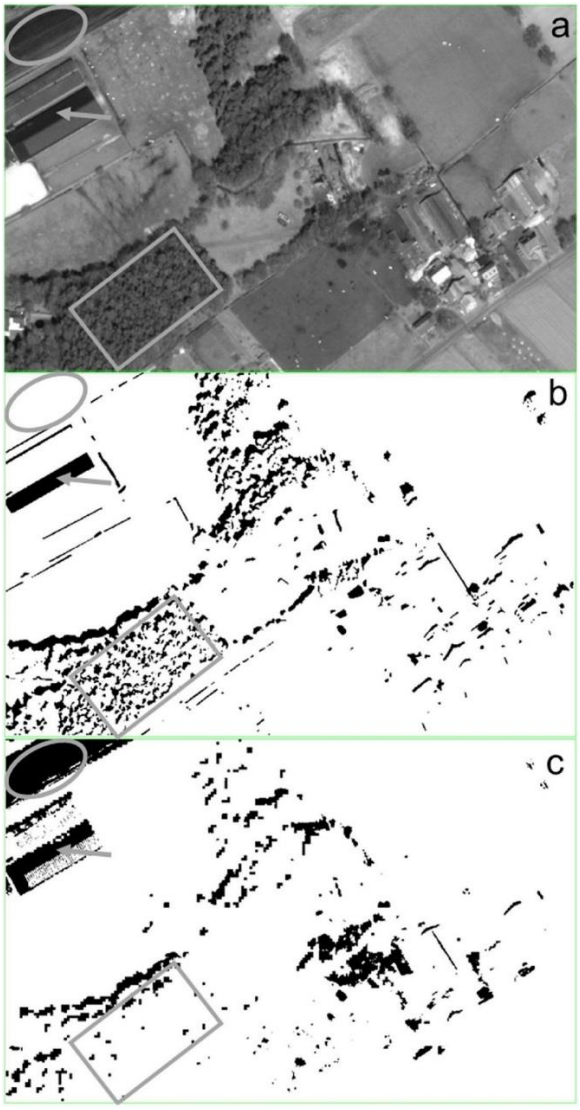


Figure 3.8. Comparison of manual and automated method of shadow detection on image subscene. a: PAN image; b: union of shadows identified by all interpreters (black); c: automated method (shadows in black). Wet soil is indicated by oval; dark roof is indicated by arrow and closed high vegetation is indicated by rectangle.

In general, locations where the visual interpretation indicated more shadow than the automated method corresponded to small shadow shapes and boundary areas of shadow patches (e.g., rectangle in Figure 3.8).

The automated method performed faster than the manual one (see Table 3.3).

3.4 Discussion

3.4.1 Manual method

For the manual method, it was found that the decision as to whether a darker patch was indeed a shadow was subjective, and interpreters differed in where they placed the boundaries of potential shadow patches. The shape and the geometrical placement of shadow boundaries were not coherent in the interpreters' results. While for the man-made constructions the interpreters mostly agreed on shadow boundaries, their interpretation and identification of the shape of shadows cast by natural vegetation were very diverse (Figure 3.5). Interpreter mapping was also highly varied in dark areas that could be interpreted as water bodies or as shadows, even though image settings, use of auxiliary data and software were chosen by each interpreter independently. Hence, interpreters' decisions appear to have been indirectly influenced by the settings chosen for image enhancement and by their analysis of auxiliary data, such as Google Earth imagery. It is possible, that the time-consuming and laborious task of visual interpretation and digitisation of shadows on an 11 ha scene discouraged some interpreters from extensive use of auxiliary data and from image enhancement manipulations during the task. Whatever the case, interpretation and manual mapping appear to be highly subjective and dependent on the expertise level of the digitising person. There is furthermore a need to agree upon the definition of shadow on agricultural land within the context of monitoring programs and to distinguish between umbra and penumbra (Arévalo et al., 2008). Some interpreters have already mentioned the need to better define shadow types (Liu and Yamazaki, 2012). Future advances could focus on defining what kinds of shadow may hamper area-based measurements of reference parcels representing agricultural land and how to treat shadows within dense vegetation.

3.4.2 Automated method

For the automated method, results below 15% of the shadow on the input images were likely caused by the land cover types in the scene (Figure 3.2, Table 3.2). In scene A (Figure 3.6), the right half of the image (A_{right}) had a smaller percentage of shadow than the left half (A_{left}), yet areas were labelled as shadow that were actually wet soil and building roofs. For scene B, looking at only the left half of the image (B_{left}), a high shadow percentage was assigned to the area with forest. The right half of this scene (B_{right}) had the highest percentage of area masked out as water bodies. Despite the exclusion of water bodies, a potential water area close to the shore, possibly undergoing eutrophication, was identified as shadow (Figure 3.6). For scenes C and D, dominated by agricultural land, differences between the shadow percentage identified on the

left and right parts of the images came from wet soil areas labelled as shadow areas (east in each scene, Figure 3.6).

The results of shadow detection by the automated method were also dependent on the acquisition date and view angle of input satellite imagery, which diminished the potential shadow presence (LPIS TG ETS, 2014). Nevertheless, automated shadow detection methods are typically incapable of distinguishing between umbra and penumbra (Arévalo et al., 2008). In our automated method, the result depended on the selected risk alpha level used for the thresholding. Note that for the automated method the obtained thresholds (T_{PAN}, T_{MS}) varied for different training areas. Therefore the threshold determined for an entire scene may not be equal to the threshold determined for a subset of the scene. User selection of risk alpha level allows customisation of the automated method and its adjustment in accordance with the input data and image quality. In our study, the automated method typically labelled larger areas as shadow than the manual method.

The automated method seems promising for the screening of large image scenes and datasets and as a first assessment of image usability for agriculture monitoring programs. It is reproducible, fast and gives results within the union of interpreters' mapping of shadows. Our case study focused on agricultural land within the EU, but it could be applied anywhere, if virtually shadow-free training pixels can be obtained from existing sources and if the focus is on agricultural or natural areas. Unlike many analyses of urban zones (Adeline et al., 2013; Dare, 2005; Li et al., 2005; Sarabandi et al., 2004; Tsai, 2006; Yamazaki et al., 2009), the method can accommodate various changes and be tailored to users' needs, such as in definition of training and interest areas, input quantile value settings and input image type and quality. The tailoring would depend on the type of available input data for creating training areas and the risk of training areas not being shadow free.

3.4.3 Comparison

When comparing automated and manual methods, it can be observed that locations where the automated method indicated more shadow than the manual method typically corresponded to water bodies, wet soils, dark roofs and high vegetation. Locations where the automated method indicated less shadow typically corresponded to small patches identified by interpreters and associated with high and fragmented vegetation (Figure 3.6, Figure 3.7). This suggests that an issue for future improvement of the automated method could be detection of humid surfaces, like wet soil. Small shadow patches are underestimated by the automated method due to the coarser resolution of MS bands.

Similarly, when comparing automated and manual methods for a subscene, the main differences in shadow designation were found in forest areas with closed high vegetation and in wet soil (see, respectively, the rectangle and oval in Figure 3.8). While for closed high vegetation, more shadow was interpreted in the interpreter's maximum union area, the area likely to be wet soil was labelled as a shadow only in the automated procedure. Wet soil was the only land cover type indicated as a shadow area by the automated method and not by the interpreters. Both methods labelled a dark roof as a shadow (arrow in Figure 3.8).

In general, the visual interpretation indicated more small shadow shapes and boundary areas of shadow patches than the automated method (e.g., see the rectangle in Figure 3.8). This is explained by the finer resolution of the imagery available to the visual interpreters. Especially for areas with dense high vegetation, small shadow detection omissions were observed. These omissions and the shadow patch shape coarseness can be attributed to the pixel size of the MS input image. Note that the implications of these missed small shadow patches for visual interpretation of imagery within the context of LPIS are negligible, since they hardly affect agricultural land delineation.

The shadow area identified by the automated procedure was larger than any of the areas identified by individual interpreters but the union of the latter areas was very close to that of the automated procedure (Table 3.1). Assuming that the interpreters jointly identified all shadow areas (with false inclusions) and that the intersection of all the joint visual interpretations with the automatically detected shadows represents true relevant shadowed area, the falsely included shadow area amounts to 4.56% (8.36%–3.80%) of the study area. This is below the assigned 5% acceptance probability of false shadow assignment and, hence, to be expected.

The runtime of the automated method was faster than manual mapping, and its speed could be further improved by using dedicated compiled software rather than a script in R.

3.5 Conclusions

Visual interpretation of shadows by humans is laborious and subjective and therefore hardly reproducible. The intersection of shadowed areas delineated by interpreters in our study amounted to less than 3% of their union, demonstrating that human' interpretations strongly disagreed. A relatively simple, automated, reproducible shadow detection method for satellite imagery

using readily available auxiliary training data was proposed. With a risk alpha level set at 5% the automated method labelled only slightly more pixels as shadow than manual shadow interpretation.

Shadow detection is a complex process that can be tackled from various angles. The presented research and results show that automated methods can enhance or even replace the visual interpretation of shadow detection. Our comparison of visual and automated results found that visual interpretation indicated more shadow in small patches and on shadow boundaries. The total area of shadow designated by the automated method was between that of the intersection and the union of the area identified by the interpreters. The automated method could be considered "another interpreter", which distinguishes itself in being faster, yielding reproducible results and not requiring laborious interpretation effort and skill that come only with years of practice and expertise.

The lack of agreement between the interpreters on the total shadow area and shadow patch location could hinder applications. Possible improvements and further research could focus on further specification of the definition of shadow on agricultural land in the context of monitoring programs. Such definition is important for both visual interpretation and for automated methods. Moreover, automated methods would benefit from improved identification of water bodies, especially close to the shore, and wet soil areas. This risk-based method will be of interest for agriculture monitoring processes such as LPIS, where recent legislation encourages use of automated image classification methods for checking agricultural activities.

Acknowledgments

We would like to thank Wim Devos and Dominique Fasbender from the European Commission's Joint Research Centre (JRC), Sustainable Resources Directorate, Food Security Unit for providing help in preparing the data. We also thank independent interpreters Barbara Barzycka, Giancarlo Carrai, Nilhan Çiftçi Sarılar, Paola Codipietro, Gizem Eren, Beata Hejmanowska, Paolo Isoardi, Pavel Milenov, Luca Sciarri, Saverio Stoppioni and Öngün Şumnulu Esirtgen for shadow interpretation and digitisation.

Chapter 4

Producing consistent visually interpreted reference data: Learning from feedback

Agnieszka Tarko, Nandin-Erdene Tsendbazar, Sytze de Bruin, Arnold K. Bregt

This chapter is based on:

Tarko, A., Tsendbazar, N.E., Bruin, S. de, Bregt, A.K., 2019. Producing consistent visually interpreted land cover reference data: Learning from feedback.

Submitted.

Abstract

Reference data for large-scale land cover map are commonly acquired by visual interpretation of remotely sensed data. To assure consistency, multiple images are used, interpreters are trained, sites are interpreted by several individuals, or the procedure includes a review. But little is known about important factors influencing the quality of visually interpreted data. We assessed the effect of multiple variables on land cover class agreement between interpreters and reviewers. Our analyses concerned data collected for validation of a global land cover map within the Copernicus Global Land Service project. Four cycles of visual interpretation were conducted, each was followed by review and feedback. Each interpreted site element was labelled according to dominant land cover type. We assessed relationships between the number of interpretation updates following feedback and the variables grouped in personal, training, and environmental categories. Variable importance was assessed using random forest regression. Personal variable interpreter identifier and training variable timestamp were found the strongest predictors of update counts, while the environmental variables complexity and image availability had least impact. Feedback loops reduced updating and hence improved consistency of the interpretations. Implementing feedback loops into the visually interpreted data collection increases the consistency of acquired land cover reference data.

Keywords

land cover mapping; learning curve; validation; visual interpretation

4.1 Introduction

Global land cover and land use maps are important for various planning and management activities (Lillesand et al., 2008; Zhao et al., 2014). For map validation and calibration, a reference dataset of greater quality than the map itself is needed. Genuine ground truth would supply such high-quality data, but populating a global dataset with a sufficiently large sample of field measurements is extremely costly. Visual interpretation of high-resolution imagery is a feasible alternative acquisition method.

The reference data collected by means of visual interpretations of remotely sensed data, even when delivered by well-trained professionals, are subject to interpreters' variation. Due to their perception of different land cover types, interpreters may largely disagree on category labels they assign to sampling units based on visual interpretation of imagery. For example, in an experiment set up by Powell et al. (2004), a group of five trained interpreters produced reference data by visual interpretation of aerial videography. The assigned land cover type differed for almost 30% of the sample units. Tarko et al. (2018) compared shadow areas interpreted by 12 individual interpreters and found that the intersection of the shadows digitised by the interpreters was less than 3% of their union. Such disagreement among interpreters is indicative of labelling error, which may have a substantial impact on the later uses of the reference dataset. McRoberts et al. (2018) showed that interpretation error induces bias into the stratified estimator of forest proportion and recommend to use input from at least three experienced interpreters to mitigate this effect. Sample data interpreted by multiple interpreters boosts the accuracy of visually interpreted datasets (McRoberts et al., 2018). In addition to collecting reference data by trained individuals, vast number of land cover interpretations can be obtained from volunteered geographic information (VGI). To overcome the issue of unknown quality of such data, the use of control locations with known land cover were used (Comber et al., 2013). However, there are no concrete methods for implementing VGI data or utilising information about the quality of individual contributors (See et al., 2015).

Another way forward for increasing the consistency of visually interpreted data is to include a review in the data acquisition process. This approach was used by Zhao et al. (2014), who created a validation dataset for a global land cover map. Samples were collected with the help of experts, later checked by those experts from the group with 'outstanding skills in image interpretation' and finally checked, and if necessary adjusted, by the most experienced interpreter. To achieve satisfactory accuracy of dataset, as much effort as two rounds of

review were implemented, but no feedback was provided to the experts during the data collection.

In the domain of education, learning, and instruction, feedback is considered to be a fundamental principle for efficient learning. It is defined as post-response information provided to learners to inform them of their performance (Narciss, 2008). Feedback loops are considered efficient in various research fields, and it is a basic concept in the education science where a feedback loop is needed to adjust the actions of teachers to ensure that a student learns (Boud and Molloy, 2013). Feedback loops are also efficient in the field of automated interpretation of images. An example of active machine learning algorithms benefitting from interpreter feedback is presented in Tuia and Munoz-Mari (2012). In the domain of medical image interpretation, where the misinterpretation of clinical exams is a delicate issue, a good training process is of high importance. Da Silva et al. (2019) proposed a training platform where the application compared the image analysis performed by a student with the teacher's and provided feedback to the user. The measures of teaching efficiency were left for the future work, but the platform usability assessment done by the students was positive.

Similar to the examples above, collecting global land cover reference data by visual interpretation can be expected to benefit from feedback loops. To assess the effectiveness of feedback provided, individual learning curves can be characterised. Learning curves are mathematical models to model skill acquisition, representing the relationship between practice and the associated changes in behaviour (Lallé et al., 2016; Speelman and Kirsner, 2005).

Given that the land cover visual interpretation may differ between the interpreters, the aim of this work is to assess whether feedback loops can improve consistency of reference data for global land cover maps. More consistent land cover reference data can be achieved when there is more agreement between the interpreters and reviewers on land cover visual interpretations. In this paper, we assess the explanatory power of variables related to image interpretation such as interpreter identifier, feedback stage, or location of the sample, in predicting the agreement level between the interpreters and the reviewers regarding visual interpretations of land cover.

Our assessment concerns acquisition of a validation dataset for the Copernicus Global Land Service (CGLS) Dynamic Land Cover project, which is similar to the work of Zhao et al. (2014), involved a review procedure. Moreover, feedback loops on individual interpretations were provided. Feedback loops were expected to induce a positive learning effect. The CGLS Dynamic Land

Cover project is a global land cover mapping effort. It is a component of the Land Monitoring Core Service of Copernicus, the European flagship programme on Earth Observation (CGLS, 2019). Its global land cover maps are generated by a supervised classification aided by random forest (RF) techniques using training data collected through visual interpretation (Smets et al., 2017). Validation is performed according to Committee on Earth Observation Satellites – Land Product Validation Subgroup (CEOS-LPV protocols, CEOS, 2019), and the data follow the design of a multi-purpose validation dataset, aiming to be applicable for multiple map assessments (Tsendbazar et al., 2018).

4.2 Methods

4.2.1 Experimental setting

To collect a global land cover reference dataset, sample sites were selected using a global stratification by Olofsson et al. (2012), which is based on Köppen bioclimatic zones (Peel et al., 2007) and human population density. Tsendbazar et al. (2018) provide details on the used sampling design. The validation sample consisted of 15743 sites of approximately 1 ha. The sites were divided between regional interpreters who then interpreted and mapped sites appointed to them. The sample site is composed of 100 equally sized square elements. Interpreters assigned a dominant land cover class to each of these elements (Figure 4.1). The sample size handled by individual interpreters ranged between 130 and 1194, with an average of 685 sites. Sample sites were offered in random order, so that the individual interpreted different land cover types over the course of the validation task.

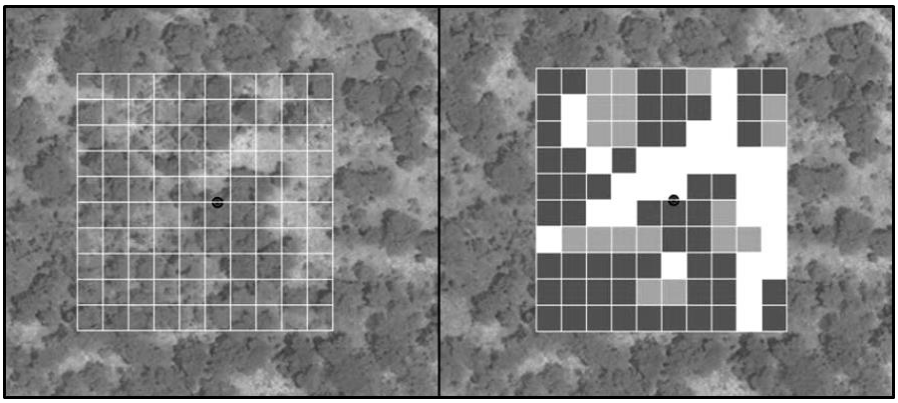


Figure 4.1. Example of a sample site. Left – the sample site (approximate size 1 ha) comprising of 100 equally sized square elements (approximate size of an element 10 m by 10 m). Right – interpreted sample site with three different land cover classes assigned to every block (white, grey, and black indicate different dominating land cover classes at element level). Source: Tsendbazar et al., 2018.

During the data collection process, four review cycles were conducted by two global land cover reviewers who provided feedback on each interpretation to the regional interpreters. In case of disagreement on the interpretation, the regional interpreters either rebutted the feedback or modified their interpretations where necessary. After finishing a feedback loop and potential modifications by the regional interpreters, the interpreters proceeded to interpret and label the next batch. For the majority of interpreters, the feedback loops were designed to first provide a quick feedback on a batch of 10–20 sample sites within few days after submission by the interpreters. Next, approximately 50 sample sites were reviewed followed by a batch of some 100 sample sites, and finally, the remaining sample sites were reviewed and feedback was passed to the regional interpreters. Data collection and the review process are schematised in Figure 4.2.

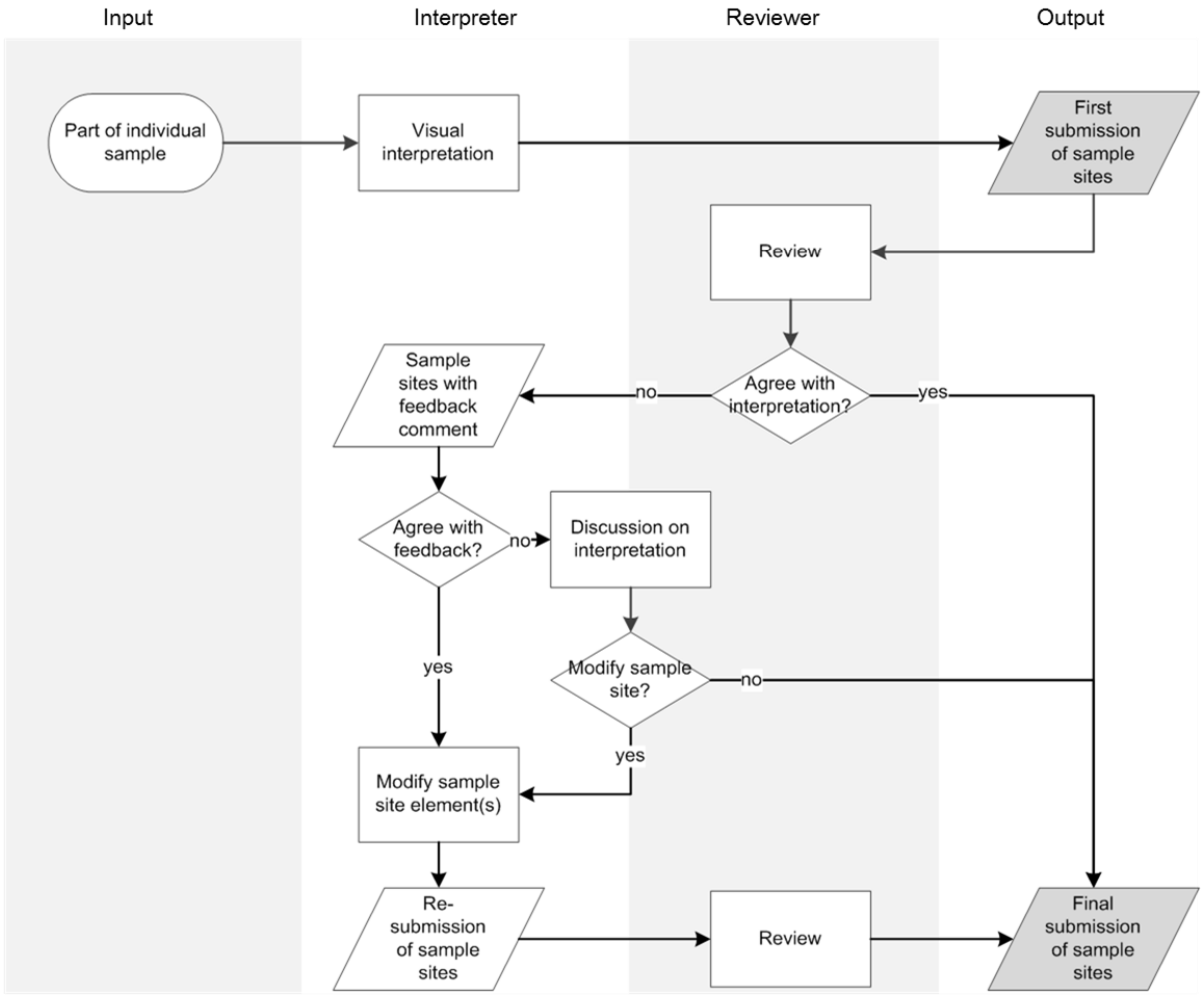


Figure 4.2. Flowchart presenting the simplified process of sample collection, review, and feedback in one of the four loops; flowchart shapes with grey background indicate compared data.

If the regional interpreters (or, in exceptional cases, a reviewer) modified land cover type for at least one of the 100 elements of a sample site, the entire sample site was considered to be re-submitted. By comparing counts of elements assigned to land cover types at the first submission and the final submission of given sample site, updated sample sites were identified (Figure 4.3). In what follows, such a sample site is referred to as an "updated sample site". Note that not every sample site with modification of element results in an updated sample site, for example re-submitted sample site, where the land cover assigned to elements has been modified, but the counts of elements assigned to land cover type are the same.

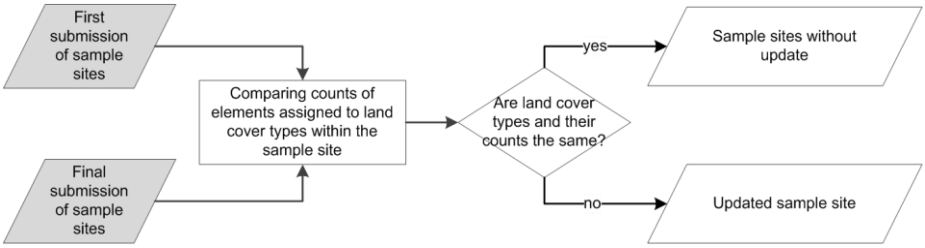


Figure 4.3. Flowchart for identifying "updated sample sites".

In a post-processing step, the proportions of land cover types at 1 ha site level were translated into the CGLS legend categories (see class definitions in CGLS (2019) and Tsendbazar et al. (2018)). Sample sites with a CGLS legend update were identified by comparing the CGLS legend categories assigned at the first and the final submission (Figure 4.4). Note that not every updated sample site results in a change in CGLS legend category.

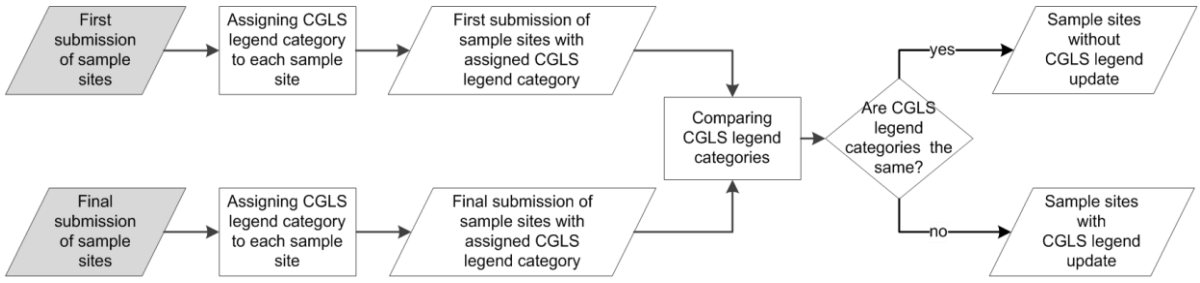
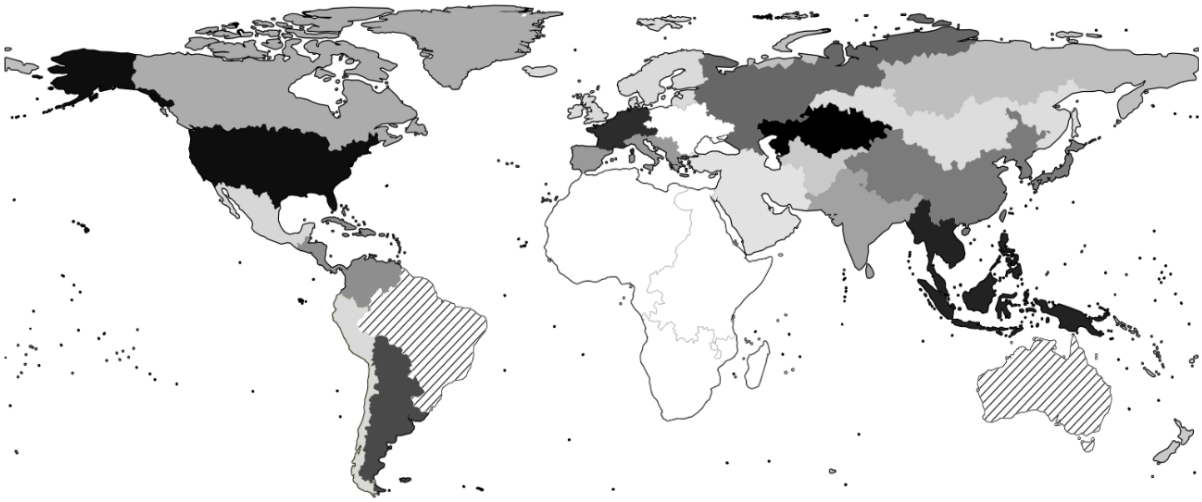


Figure 4.4. Flowchart presenting the process of identifying sample sites with update in CGLS legend category.

Data acquisition involved 27 regional interpreters distributed over 25 regions. Following the finding that volunteers interpreting land cover perform better in case of samples near their familiar places or samples with their familiar climate type (Zhao et al., 2017), experienced interpreters involved in our experiment were selected based on their region of expertise. In two regions, data collection was done by two interpreters to handle the large sample size; the other regions had one interpreter each (Figure 4.5(a)). All interpreters were experienced in satellite-based land cover analysis and image interpretation. All of them were provided with a mapping tutorial explaining the interface for data collection, the land cover interpretation specific for the project, and the interpretation keys. Since the learning curves of the interpreters most likely changed already after getting acquainting with the tutorial, the starting point of our analysis coincides with the moment the tutorial was finished. Collection of the first few points was organised as an on-line training exercise that was tailored to each interpreter's needs. Three interpreters mapping three regions in Africa had prior knowledge and experience with the project because they had contributed to a similar task before (Tsendbazar et al., 2018). The results produced by those interpreters were excluded from the experiment, as their learning curves were expected to be different from the interpreters who took the activity for the first time. For similar reasons, data of one interpreter mapping Eastern Europe was excluded from the analysis (Figure 4.5(a)). In total, input of 23 interpreters was analysed for the purpose of this paper. Figure 4.5(b) shows the spatial distribution of sample sites.

The CGLS land cover validation data were collected using a dedicated branch on the Geo-Wiki Engagement Platform (http://www.geo-wiki.org). Figure 4.6 shows a screen shot of the validation data collection interface. Through the interface, several remote sensing images were interpreted, and the prevalent land cover was assigned to each element. Land cover types to be assigned are listed in the rightmost panel of Figure 4.6. Interpreters could use several data layers, i.e.:

- openly-available very-high-resolution Google and / or Bing imagery;
- Natural Colour Composite and False Colour Composite Sentinel-2 imagery from 2015;
- time-series imagery from Sentinel 2;
- normalised difference vegetation index (NDVI) time-series from Landsat 7 32-Day, MOD13Q1.005 16-Day Global 250 m, PROBA-V C1 Daily at 100 m; and/or
- map with Köppen-Geiger bioclimatic zones (Olofsson et al., 2012).



(a)

(b)

Figure 4.5. (a) Validation regions. Grey tones indicate regions interpreted by single interpreters; hatch patterns indicate regions interpreted by two interpreters; white fills indicate regions outside the scope of this paper's experiment. (b) Distribution of sample sites (grey dots) in the scope of this paper experiment.

Interpreters were also offered functionality to export the sample site to Google Earth, which allowed viewing historical imagery. Whenever possible, Google image was the main data layer to be used. Interpretation targeted to represent land cover in the growing season of 2015.

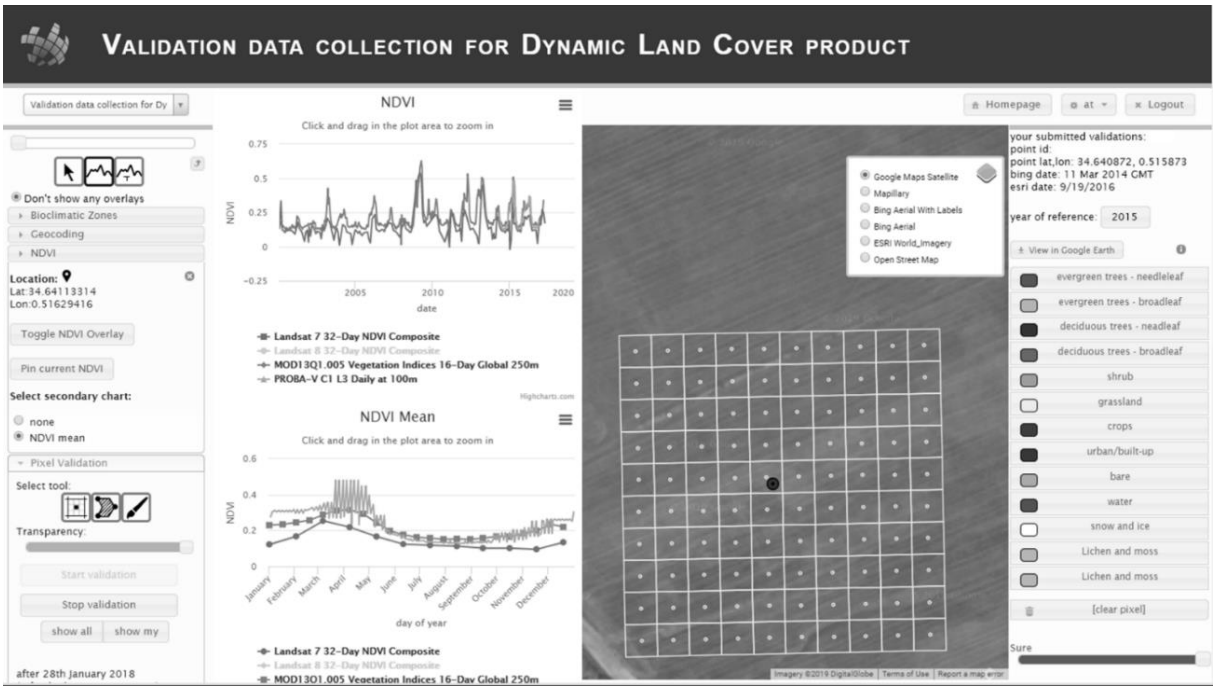


Figure 4.6. Screen shot of Geo-Wiki portal interface for land cover validation. The leftmost panel allows selection of additional data such as NDVI profile or bioclimatic zone; the second panel from the left shows the local NDVI; the third panel from the left displays the sample site with chosen background image; the rightmost panel shows the list of land cover types.

4.2.2 Exploratory analyses

We expected that regional interpreters interpreting the validation samples gained practice over time and that the feedback loops induced the learning effect. We quantified the learning effect with the update level changing in time for each individual. Updates upon feedback were counted and expressed as a percentage relative to the total number of sample sites submitted by the interpreter concerned up to a given moment in time. From here on these percentages are referred to as "momentary percentage of updates" (MPU).

We researched nine factors as potential explanatory variables, clustered in three categories (training, personal, and environmental) and listed in Table 4.1. Note that interpretation duration was calculated under the assumption that a submission gap longer than 30 min corresponded to a break taken by the interpreter. Interpretation duration could not be computed for the first submission after any break. As a consequence, 1152 out of the 15743 sample sites lacked data of interpretation duration.

To assess the relationship between interpreter identifier and MPU, we investigated individual learning curves of the interpreters as well as a collective

learning curve (aggregated over all interpreters). Learning curve is expressed as a graph indicating normalised timestamp in the x-axis and MPU in the y-axis.

To approximate interpreter's proficiency in land cover interpretation, we asked the interpreters about their years of experience with land cover, land use, and vegetation cover mapping in a form of a survey. Possible responses were grouped in five ordinal categories:

- up to 2 years;
- from 2 up to 4 years;
- from 4 up to 6 years;
- from 6 up to 10 years;
- 10 and more years of experience.

Image availability was assessed using data from the work of Lesiv et al. (2018), which presents the availability of Google Earth imagery (with resolution <5 m) across the world's land surface for different growing seasons. Bing images were not included in their seasonal analysis. The world is represented by a 1° grid holding information concerning seasons on available imagery in four ordinal categories:

- no information on seasons;
- images taken only in non-growing season;
- images only in growing season;
- images from growing and non-growing seasons.

Through overlay, we determined the availability of Google images in growing seasons for each sample site.

The influence of each factor on the MPU was assessed using scatter plots, bar graphs, box plots (McGill et al., 1978), and Spearman's rank-order correlation. Factors can be correlated because some of them represent similar attributes, such as timestamp and feedback stage. As a diagnostic for RF analysis we used a correlation matrix. For obvious reasons, categorical factors (land cover class and interpreter identifier) were excluded from the correlation analysis.

All plots were created using R software for statistical computing (R Core Team, 2017) using the "graphics" packages for box plots (R Core Team, 2017), the "plotly" package for scatter and bar plots (Sievert, 2018), and the "corrplot" package for correlation matrix (Wei and Simko, 2017).

Table 4.1. Selected factors potentially influencing the MPU.

Category	Factor	Description	Range
Personal	Interpreter identifier	Individual identification of the interpreter	23 interpreters, id labels from 1 to 23
	Experience	Ordinal categorisation of years of experience in land cover / land use visual interpretation of the interpreters	5 ordinal categories: up to 2 years; 2-3; 4-5; 6-9; 10 and more
	Interpretation duration	Time used to submit the sample site by the interpreter	From 0 to 30 minutes
Training	Feedback stage	Ordinal categorisation of the review cycle at which the sample site was mapped	4 ordinal categories: from first to fourth stage (review cycle)
	Timestamp	Time (seconds) between the first collected sample site (time 0) and the submission of any other sample site. Registered for each interpreter, for first submission of given sample site	From 0 to 10262630 seconds (16 weeks 6 days 18 hours 44 minutes)
Environmental	Complexity	Number of different land cover types identified and mapped within the sample site final submission	From 1 to 6
	Image availability	Four-level ordinal categorisation explained above (section 4.2.2)	4 categories: no information on season, non-growing season only, growing season only, information on both seasons
	Land cover	Final land cover assigned to the sample site according to the CGLS legend category	9 categories: bare, closed forest, crop, grass, open forest, shrub, snow and ice, urban, water
	Location	Longitude and latitude of the sample site	84°N - 56°S, 180°W - 180°E

4.2.3 Modelling the learning effect

RF regression analysis was chosen to identify importance of factors for describing the learning effect. The input factors in Table 4.1 were used as explanatory variables. Random forest regression analysis was chosen because tree-based models can handle correlated input data, non-linear relationships, and mixtures of categorical and numerical data types. Moreover a RF model is non-parametric, accounts for interactions, and is robust against overfitting (Breiman, 2002, 2001).

Since RF cannot handle missing predictor values, we analysed two models:

- a model using all (ten) explanatory variables but excluding sample sites without data on interpretation duration (14591 sites were used);
- a model using all sites (15743 sites) but without the interpretation duration factor (nine explanatory variables were used).

First model allows importance identification of all factors while the second model uses all available input sites. The two models are complementary.

The parameter settings in the RF regression analysis were as follows: 500 trees, three variables tried on each split and a minimum of five observations in the terminal nodes. From the model we obtained:

- mean square difference (MSD, sum of squared residuals divided by the number of sample sites in the dataset);
- percentage of variance explained for the entire validation dataset (formula: 1 – MSD / variance of the dataset);
- variable importance (reported as % increase of MSD). Variable importance was estimated with out-of-bag cross-validation as a result of variable being permuted.

To assess the stability of the RF results, we ran the models 15 times and reported average values of MSD, percentage of variance explained for the entire validation dataset, and variable importance, as well as the range (smallest and largest value) obtained from the 15 iterations for each value. Goodness of fit is indicated by the percentage of variance explained and MSD, while variable importance was assessed by the percentage increase of MSD.

The RF regression analysis was performed using R software (R Core Team, 2017) using the "randomForest" package (Liaw and Wiener, 2002).

4.3 Results

4.3.1 Exploratory analysis

Figure 4.7 shows the correlation matrix of selected factors that were deemed to influence the MPU. As expected, timestamp and feedback stage are strongly correlated, which can be explained by the second factor being a discrete representation of the first one. Note also the observed positive correlation between interpretation duration and complexity owing to visual interpretation of complex scenes being usually more time consuming. Location factors (longitude and latitude) showing negative correlation with interpreter's experience are considered as random effect of the choice of regional interpreters.

	Longitude	Latitude	Interpretation duration	Image avaliability	Experience	Feedback stage	Timestamp	Complexity
Longitude	1	-0.25	-0.04	-0.05	-0.23	0.04	-0.01	0.01
Latitude		1	0.2	-0.17	-0.04	-0.07	-0.12	-0.03
Interpretation duration			1	0.03	0.11	-0.15	-0.19	0.35
Image avaliability				1	0.11	-0.07	-0.03	0.18
Experience					1	-0.01	-0.01	0.03
Feedback stage						1	0.69	-0.02
Timestamp							1	0.06
Complexity								1

Figure 4.7. Correlation matrix of factors potentially influencing the momentary percentage of update.

4.3.2 Personal factors

Figure 4.8 shows selected learning curves for individual interpreters with normalised timestamp factor on the x-axes. MPU varied in time and per interpreter and changed from 0 up to 100 for different interpreters at different moments during the mapping process. For Figure 4.8(a), the curves indicate a general downward trend in time; those correspond to interpreters who learned

from the feedback loop. These curves represent positive learning effects. Positive learning effects were observed for the majority of interpreters who were characterised by high MPU at the beginning of the task and lower MPU towards the end of the data collection process. In Figure 4.8(b), the curves show upward MPU trends, representing interpreters to whom the feedback did not bring the expected learning effect. Learning curves strongly differed between individual interpreters (Figure 4.8). Moreover, learning effects also changed over time for individual interpreters (see Figure 4.8). When calculating the percentage of updated sample sites per feedback loop for each interpreter, only three of them reached the highest update percentage in the third or fourth loop, meaning that the positive learning effect is not confirmed for those three individuals.

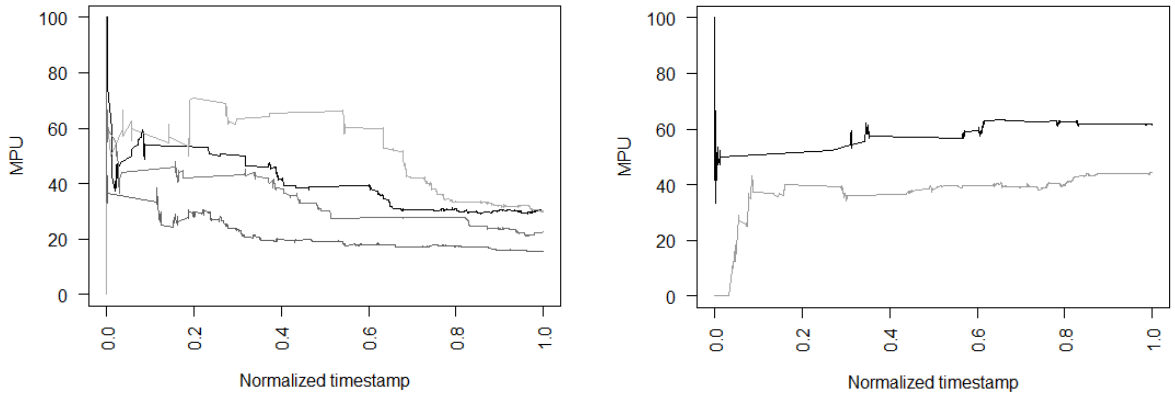


Figure 4.8. (a) Exemplary learning curves of interpreters (indicated by different grey shades) with positive learning effect. (b) Exemplary learning curves of interpreters without positive learning effect (indicated by different grey shades).

Figure 4.9 shows the aggregated learning curve over all regional interpreters. The solid black line with the downward trend means that there was a positive learning effect over the entire group of interpreters on average because the MPU dropped in time and finally reached 30% of updated sample sites. Translated into the CGLS legend category at the sample site level, the final update percentage on CGLS legend category is 9% (solid grey line in Figure 4.9).

The dashed lines in Figure 4.9 show the update percentage relative to the total sample. The lightest grey line shows that the data collection increases in time, and the exponential-like shape of the plot indicates that data collection was more intensive during the last stretches of the project. The darkest line

indicating the percentage of updated sample sites shows a stable increase over time, with slightly steeper slope of the plot from the 0.8 of normalised timestamp of the collection task. Similarly, for the percentage of sample sites with CGLS legend update, the percentage increase plot seems linear (medium-grey colour).

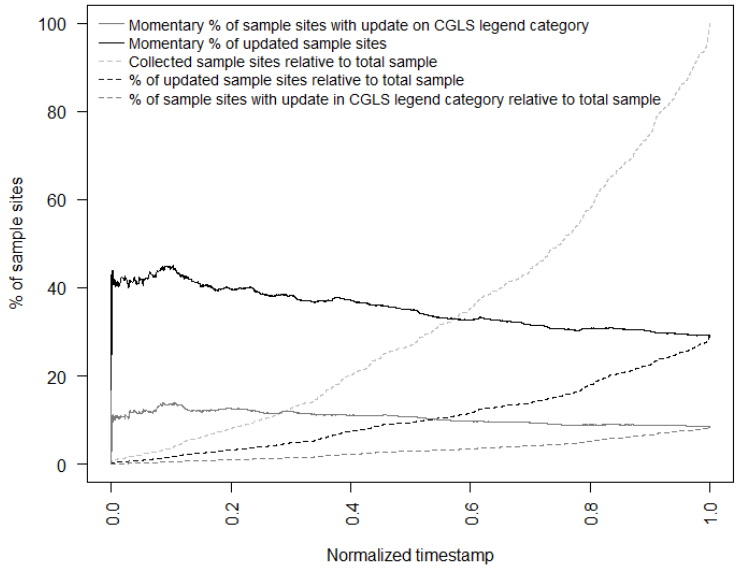


Figure 4.9. Learning curves aggregated over all regional interpreters.

Figure 4.10 shows the distribution of percentage of updated sample sites per experience category. The update percentage is expressed relative to an interpreter's individual sample size, and the category is represented as years of experience in land cover / land use visual interpretation. Regional interpreters participating in the land cover reference data collection were evenly distributed concerning years of experience (three interpreters with the least experience category and five interpreters in each of the other categories). The lowest mean value of the update percentage for the individual interpreters was for the group with four to six years' expertise, and the highest mean value concerned interpreters with the longest experience. Less experienced interpreters (less than six years of experience) tended to have similar update rates, while interpreters with more than six years of experience varied considerably in terms of update rates. The percentage of updated sample sites substantially varied between individual regional interpreters: the lowest update percentage was 12%, the highest 62%, and the mean 30% (Figure 4.10).

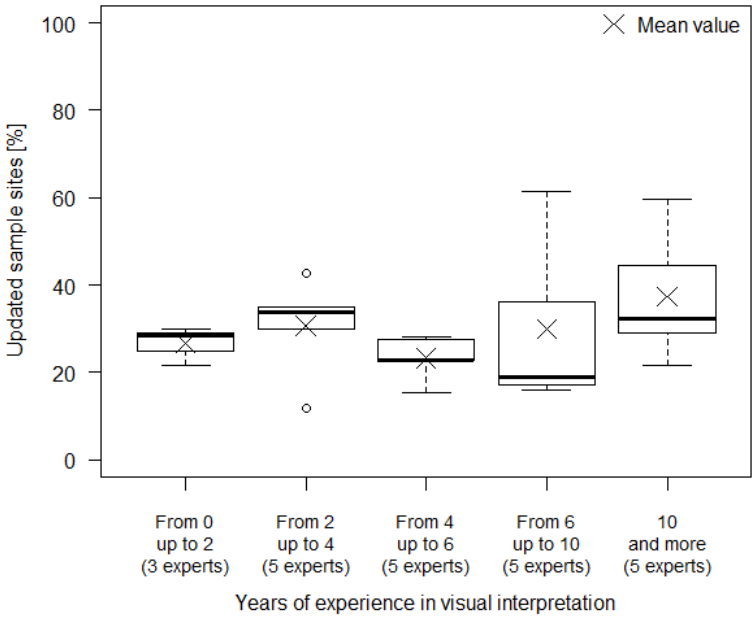


Figure 4.10. Distribution of updated sample sites per interpreters' experience category.

4.3.3 Training factors

The box plot in Figure 4.11 shows percentage of updated sample sites grouped by feedback stage. The mean and the median of update percentage decreased over subsequent feedback stages. The spread of update percentages for individual feedback stages is caused by the large variation among the interpreters.

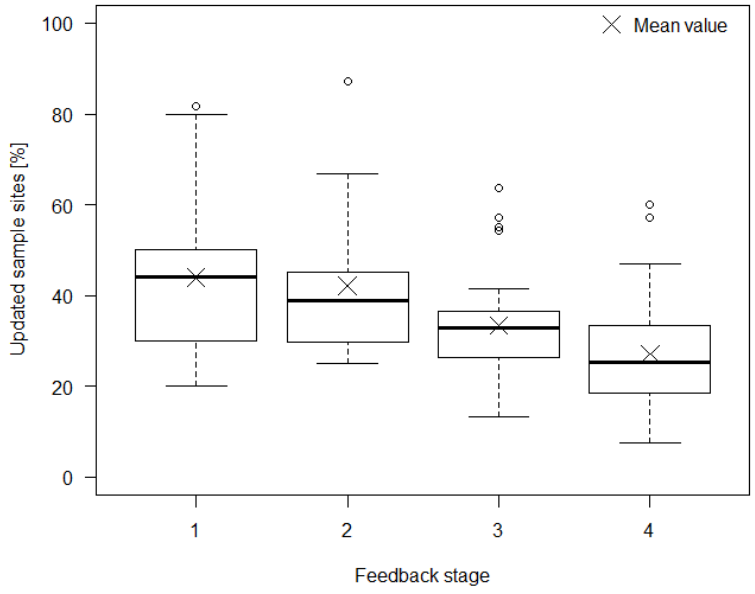


Figure 4.11. Distribution of updated sample sites per feedback stage.

4.3.4 Environmental factors

The exploratory analysis of relationships between environmental factors and interpretation updates are shown in Figure 4.12. Figure 4.12(a) concerns land cover complexity expressed by the number of land cover types within a sample site. The majority of the sample sites (~89%) did not have more than three different land cover classes. The update percentage increased with the increasing number of land cover classes up to five (Figure 4.12(a)). Note that fewer than 4% of all sample sites had five or more different land cover classes, and therefore the categories with highest number of land cover may not be representative for drawing conclusions on update percentage.

Figure 4.12(b) and 4.12(d) show the total sample categorised by the final CGLS legend. Figure 4.12(b) illustrates that the majority of sample sites (63%) had forest (closed and open) or grassland as a final CGLS legend category. The urban land cover had only 3% of sample sites from the total sample, but the update percentage was the highest from all CGLS legend categories (44%). The lowest update percentages were for the classes "water" and "snow and ice" (12% and 11%, respectively).

Figure 4.12(d) shows the distribution of percentage of updated sample sites for individual interpreters against the final CGLS legend category. Closed forest, open forest, and grassland land cover had the biggest dispersion of the update percentage among the interpreters and the highest mean update percentage values.

Figure 4.12(c) shows the total sample categorised by the image type available for mapping and the distribution of percentage of updated sample sites with the same image availability, calculated for each interpreter. For more than half (59%) of total sample, images with at least growing season were available. The percentage of updated sample sites relative to all sample sites with given image availability varied between the interpreters: most for the updated sample sites with images available only in non-growing season (from 10% to 90%) and least for the updated sample sites with images available only in growing season (43% difference).

The location factor showed negative correlation with interpreter's experience, but less strong negative correlation with image availability and timestamp (Figure 4.7).

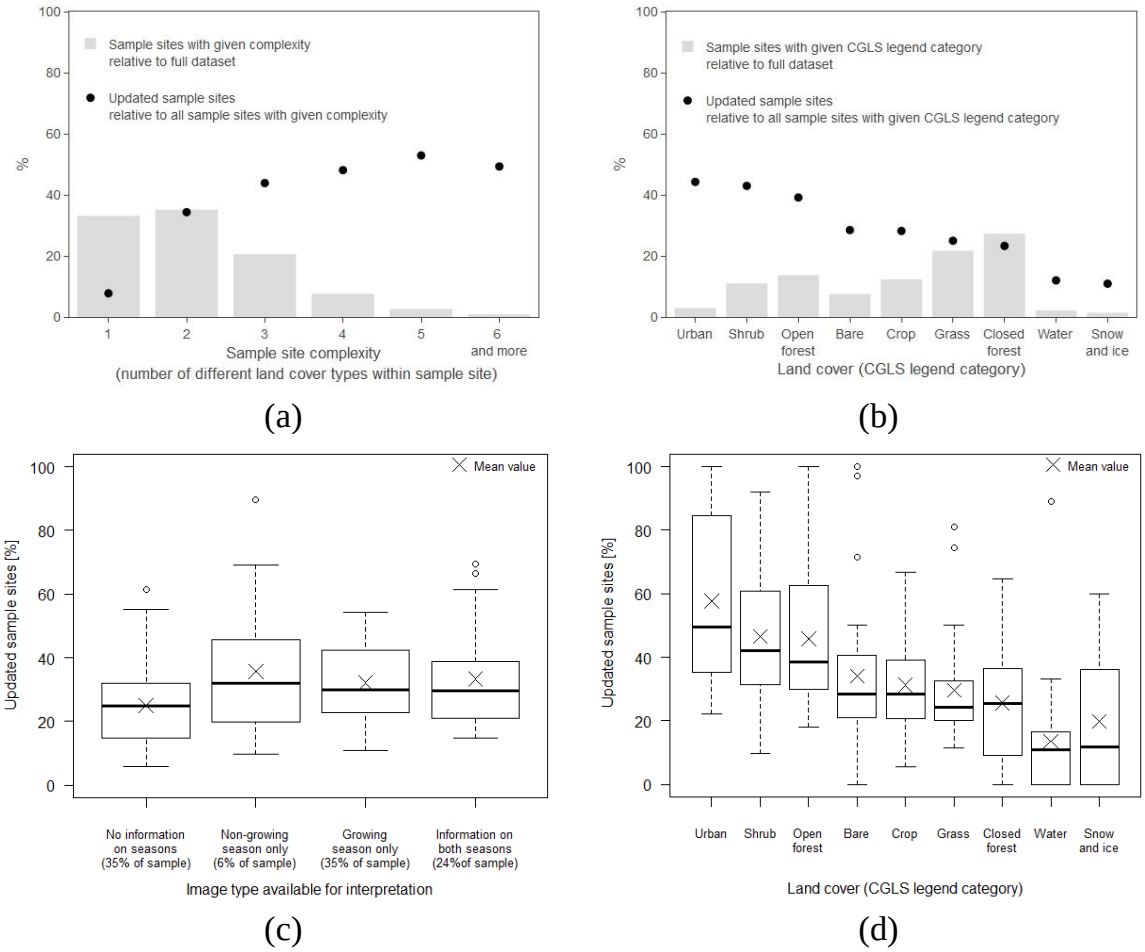


Figure 4.12. Environmental category analysis: (a) updated sample sites per given complexity (black dots) and percentage of sample sites with given complexity relative to total sample (grey bars); (b) updated sample sites per final land cover class (black dots) and percentage of sample sites with given CGLS legend category relative to total sample (grey bars); (c) distribution of updated sample sites per image type available for mapping; (d) distribution of updated sample sites per CGLS legend category.

4.3.5 Random forest

In Table 4.2, we report the percentage of variance explained and MSD results of the RF regression model. Table 4.2 shows that the fit is high for both versions of the model. The mean from 15 runs explained 98.0% and 96.5% of variance of the MPU for the individual interpreters in first and second model, respectively, and the range was less than 1% in both cases. The MSD value was higher for the second version on the model (5.5%) and almost double compared with the first model version.

Table 4.2. Goodness of fit statistics of the RF regression model based on 15 iterations.

Model version / Value	(1) Dataset subset	(2) Full dataset
	Mean (range)	Mean (range)
Variance explained, %	98.0 (97.9–98.0)	96.5 (96.4–96.6)
MSD, %	3.2 (3.1–3.2)	5.5 (5.5–5.7)

For both model versions, the order of mean importance value was the same for the first three factors: interpreter identifier, timestamp, and feedback stage from the personal and training categories. From those, the first two factors were ranked the same in all single model runs. In Table 4.3, we reported the mean importance of the input factors and in parentheses their range in 15 runs. The most important variable for both models and in all runs was the interpreter identifier, with 76.2% mean importance in first model and 80.0% mean importance in second model. The second-most important factor was the timestamp and the third-most important factor was the feedback stage. In the first model, the range of the feedback stage importance was overlapping with the next in order – experience factor range; therefore, in two single runs, the order of feedback stage and experience factors was swapped. The two least-important factors were complexity and image availability, with swapped order between the model versions and between the runs within the model. Their mean importance was between 11.6% and 13.2%, with the ranges from 9.1% to 15.1%.

Table 4.3. Importance of the RF explanatory variables based on 15 iterations.

Model version / Factor	(1) Dataset subset	(2) Full dataset
	Mean importance, % (range)	Mean importance, % (range)
Interpreter identifier	76.2 (73.1–80.4)	80.0 (77.3–84.6)
Timestamp	65.8 (61.9–68.9)	69.6 (66.9–72.1)
Feedback stage	34.3 (30.5–37.4)	35.9 (33.1–38.6)
Experience	32.1 (30.9–33.8)	31.3 (29.3–33.1)
Location (latitude)	30.8 (29.7–32.3)	31.4 (30.0–32.4)
Location (longitude)	26.6 (24.9–29.3)	27.5 (25.7–29.9)
Land cover	22.6 (21.1–24.9)	19.1 (17.5–21.5)
Interpretation duration	19.0 (16.0–21.8)	–
Complexity	13.2 (11.4–14.1)	11.6 (9.6–12.6)
Image availability	12.6 (10.9–15.1)	12.0 (9.1–14.0)

To assess the importance of feedback, we run once RF regression with parameter settings as above, but without timestamp variable. The model fit was high, at 92.2%, with MSD of 12.1%. Regarding the importance of explanatory variables, by far the most important factor was feedback stage (228.2%), followed by interpreter identifier (78.9%), land cover (32.1%), and latitude, (30.1%). The least important was image availability (14.1%).

4.4 Discussion

4.4.1 Interpreter identifier and training factors

We assessed basic factors influencing learning effect represented by MPU. The most important factors were interpreter identifier, timestamp, and feedback stage (Table 4.3). Timestamp and feedback stage were strongly correlated (Figure 4.7), as the latter can be considered a discrete representation of the first one. In the RF regression model, timestamp has a finer granularity than feedback stage, which may explain its higher importance rating compared with the four-level feedback stage (Table 4.3). Despite its coarser granularity, feedback stage immediately follows timestamp in the importance ranking (Table 4.3). This implies that it adds information to the timestamp variable. Assessing the model without the timestamp factor, feedback stage comes in first place as the most important explanatory variable influencing the MPU. Feedback adds to the fact that, with time, interpreters gained more knowledge on the project and confidence using the software through autonomous learning or "learning by doing" (Schank et al., 1999).

Interpreter identifier and timestamp, together with the MPU, are presented as individual learning curves, and in our study a decrease of MPU for individuals indicated a positive learning effect of the regional interpreters (Figure 4.8(a)). The biggest drops in the MPU for various interpreters were in different moments of the normalised time (Figure 4.8(a)). All curves were distinct, emphasising the interpersonal differences between the interpreters. Despite regular review and feedback loops, the positive learning effect is not confirmed for three interpreters out of 23 (Figure 4.8(b)). The reasons of this finding are not clear to the authors.

The interpreter identifier is a categorical factor, with 23 distinct values. Since in the RF method the variable importance measures for categorical predictor variables are affected by the number of categories (Strobl et al., 2007), we repeated variables assessment with the "cforest" function from the "party" package (Hothorn et al., 2006; Strobl et al., 2008, 2007). This function provides unbiased variable selection in the individual classification trees (Strobl et al.,

2007). The importance of the order of factors was identical to the order reported in Table 4.3, confirming our earlier results. Despite the many levels of the interpreter identifier, its importance was prevalent, meaning that this remains the most important factor influencing the positive learning effect of the interpreters.

The group of interpreters collected less intensively at the beginning of the task and collected many more sample sites towards the end of the mapping task: Figure 4.9 shows that only 30% of the sample sites were collected half way during the assignment. This might be also partially a result of more frequent feedback loops at the beginning of data collection. However, the last feedback loop had the lowest update percentage (Figure 4.11). A regular review without feedback is one of the ways to increase the consistency of collected dataset (Zhao et al., 2014). In the work of Zhao et al. (2014), sample sites collected by interpreters were checked by one reviewer and adjusted when necessary. Such a procedure can be prone to the subjectivity of the reviewer's final assignment of land cover. In our data collection design, feedback on all sample sites was implemented and provided to the regional interpreters. In case of disagreement, interpreters had a possibility to rebut the reviewer's feedback and therefore to reduce the reviewer's subjectivity of land cover interpretation. The mean and the median of update percentage for individual interpreters was decreasing in the subsequent feedback stages (Figure 4.11), meaning that the interpreters and the reviewers agreed more often on the sample site interpretation at the later stages of the data collection process.

In the experiment of Powell et al. (2004), five trained interpreters produced reference data by visual interpretation of aerial videography, where the assigned land cover type differed for almost 30% of the sample units. In our study, the MPU at the end of our experiment showed that 9% of sample sites were updated regarding CGLS legend category (Figure 4.9). This update percentage highlights that fewer updates were required thanks to the feedback stages implemented in this study.

4.4.2 Personal factors

Personal factors influenced the learning effect of the individuals. This result is similar to a study done by Van Coillie et al. (2014) where a web-based digitisation exercise performance was mainly determined by interpersonal differences.

The number of years of experience in visual interpretation was previously used as a measure of interpreter expertise (Mincer, 1974). Our results (Table 4.3)

suggest that the interpreter identifier is twice as important as the number of years of experience. This finding indicates that there are large differences between interpreters, which are not captured by years of experience.

In visual interpretation projects with many actors, it is challenging to engage a uniform group of interpreters with similar interpretation skills, regional expertise, and experience. In our research, interpreters had different years of experience and their percentage of updated sample sites varied, even for individuals within the same interpreter's experience category (Figure 4.10). In our experiment all interpreters had remote sensing background, previous experience in land cover classification and knowledge on the region of their expertise. In the absence of detailed information about the experience of interpreters, we chose number of years of experience in land interpretation as a feasible indicator of individual experience. The number of years of experience may be considered an insufficient indicator of interpretation expertise as it does not cover the intensity of work nor regional knowledge, for example. It would be worthwhile exploring alternative indicators (e.g., experience only in image interpretation) if richer data about the interpreters is available.

4.4.3 Environmental factors

The complexity factor was positively correlated with the interpretation duration (Figure 4.7 and Figure 4.12(a)), meaning that more land cover classes within a sample site coincided with an increase in time needed to interpret a sample site. Although complexity had little impact on the learning effect (Table 4.3), knowledge on the level of complexity for a mapped area can facilitate task planning: visual interpretation is likely to take more time for sample sites with complex land cover.

Image availability (see section 4.2.2) was found to be the least important explanatory factor (Table 4.3). In contrast, a study of Zhao et al. (2017) found that with increased VHR image availability, more volunteering interpreters agreed on the majority land cover type, which implied higher reliability. In our research image availability did have an influence on MPU, although other factors were found to be more important. Moreover, we did not investigate whether interpreters have used all available imagery and ancillary data.

It could be valuable to assess the extent in which data were really used by the interpreters. Additional detailed characteristics of all available images (such as spectral, temporal, and spatial resolution) and other input data such as NDVI information or Google Street View can be an important tool in the absence of ground truth observations. Integration of various imagery and ancillary data is

a current direction in land cover / land use data collection platforms. For example, a dedicated branch of the Geo-Wiki Engagement Platform (http://www.geo-wiki.org) used in this experiment, next to the collection of Bing and Google images, Sentinel 2 imagery, and NDVI profiles, offered functionality to export sample site shape to a Google Earth programme to review historical imagery and Google Street View. Another example is Collect Earth, an open source tool for environmental monitoring enabling data collection through Google Earth in conjunction with Bing Maps and Google Earth Engine (http://www.openforis.org).

Location of the interpreted sample site is less important than the feedback stage, yet latitude is more important than longitude (Table 4.3). A potential explanation is that latitude is roughly followed by the climate zones, which in this research were taken into account in sample sites selection by stratified random sampling considering Köppen bioclimatic zones (see section 4.2.1). There are more consistent variations in the bioclimatic zones along the latitudes rather than the longitudes, and bioclimatic zones could reflect landscape types. The influence of bioclimatic zones could be investigated further to identify MPU hot spot areas.

4.4.4 Research method

In case of absence of land classification performed on the ground, reference data used for developing and validating large-scale land change maps are commonly acquired by visual interpretation. Interpretation involves remotely sensed images with higher resolution than those used for map creation and is considered of greater accuracy than the map (Olofsson et al., 2014). Since visual interpretation is subjective which introduces a source of uncertainty (Jia et al., 2016; Pengra et al., 2019; Powell et al., 2004), various methods of boosting data consistency can be implemented, such as field visits (if resources are available), having sites labelled by multiple interpreters, or a review procedure. In our research, a review with feedback loops was implemented, and we assessed the effect of multiple variables influencing agreement between the interpreters and reviewers on visual interpretations. To assess the magnitude of the reference data consistency improvement, we recommend a comparative study setup including a control group performing visual interpretation but not receiving a feedback.

Having confirmed the disagreement between individuals in land cover interpretation, to obtain the reference data with boosted accuracy, McRoberts et al. (2018) and Powell et al. (2004) suggest having sites labelled by multiple interpreters providing the majority interpretation. Such an approach can be

challenging to implement for a large-scale global reference datasets that involve many interpreters from different regions of the world. The two approaches – multiple interpreters delivering majority land cover class and a single interpreter collecting land cover data whose work is reviewed and feedback is provided – are considered complementary.

4.5 Conclusions

Land cover reference data acquired by visual interpretation are affected by interpreter subjectivity. One way to assure a consistent land cover reference dataset is to include a review step in the acquisition process. In our experiment concerning global land cover reference data acquisition, we researched the rate of land cover updates following reviewers' feedback on visual interpretations performed by 23 regional interpreters. The number of updates following feedback differed substantially between interpreters. Despite those differences, feedback loops induced a positive learning effect in land cover visual interpretation for 20 of the 23 interpreters. Those interpreters delivered more consistent land cover interpretations, which resulted in a reliable land cover validation dataset.

The most important factors influencing the learning effect were those from the personal and training categories: interpreter identifier, timestamp, and feedback stage while the least important factors were from the environmental category, being complexity of the sample site and image availability. We observed a positive learning effect upon consecutive feedback loops. Interpreter identifier and timestamp, together with the momentary percentage of update, can be expressed as individual learning curves. The majority of individual curves showed a positive learning effect.

Collection of reference data through visual interpretation performed by interpreters benefits from a feedback loop, which increases the consistency and reliability of the collected dataset. Within a reference data collection project, factors such as interpersonal differences between the interpreters or autonomous learning of interpreters cannot be fully controlled, while review and feedback can be planned and customised to optimise the project results.

Acknowledgements

This work was supported by the European Commission – Copernicus program, Global Land Service. The authors thank the regional interpreters for their contribution to collecting the validation dataset.

Chapter 5

Influence of image availability and change processes on consistency of land transformation interpretations

Agnieszka Tarko, Sytze de Bruin, Nandin-Erdene Tsendbazar, Arnold K. Bregt

This chapter is based on:

Tarko, A., Bruin, S. de, Tsendbazar, N.E., Bregt, A.K., 2019. Influence of image availability and change processes on consistency of land transformation interpretations.

Accepted for publication in Int. J. Appl. Earth Obs. Geoinf.

Abstract

Large-scale land change maps are essential to support policies addressing land transformations. Development and validation of large-scale land change maps use reference data that are commonly acquired by visual interpretation of remotely sensed images. However, visual interpretation itself is prone to error. Little is known about factors influencing the quality and consistency of changes detected by visual interpretation. This paper reports on an experiment assessing the effect of the number of very high resolution images and land change process types on the consistency of visual interpretations. The experiment involved 48 sites scattered over Europe for which 18 individuals interpreted very high resolution images provided via Google Earth. Land change process type was found to have a significant impact on the consistency of visual interpretations, while the marginal effect of the number of images was not significant. Absence of change on non-agricultural land was interpreted with high consistency. On the contrary, agricultural land abandonment and reforestation were the least agreed upon. We conclude that for increased efficiency, resources allocated to acquire reference data by visual interpretation should be adjusted based on the expected type of land change. Interpretation of agricultural land abandonment, reforestation and agricultural land expansion require most efforts.

Keywords

visual interpretation; land cover; land use; agriculture; land change

5.1 Introduction

Land change is both a cause and consequence of global environmental change (Foley et al., 2005; Song et al., 2018). Understanding this change is essential for managing the natural capital of the earth and, therefore, land change maps have an important application in research, management and policy at global and regional scales (Pengra et al., 2019). To detect land changes, temporal and spatial data are acquired using remote sensing. Recurrently monitoring the global surface, satellite observations contribute substantially to our understanding of the land change extent (Song et al., 2018). For example near-real time monitoring of changes on agricultural land has become essential for the European Union common agricultural policy. This policy aims to ensure food security, the sustainable use of natural resources and balanced development of Europe's rural areas (DG AGRI, 2018).

Land change monitoring is aided by increasing numbers of open-access wall-to-wall maps of land cover, land use and land change. Examples are:

- the pan-European land-change layer characterising changes in land cover and land-use (CORINE, 2019);
- an ongoing programme of yearly global land cover maps at 100m spatial resolution within the Copernicus Global Land Service: Dynamic Land Cover project (CGLS, 2019);
- a Global Forest Change map characterising forest extent and change (Global Forest Change, 2019; Hansen et al., 2013);
- the U.S. Geological Survey Land Change Monitoring, Assessment, and Projection land cover change global map (USGS LCMAP) (USGS, 2019; Young, 2017);
- the Chinese Earth land cover map GlobeLand30 (GLOBELAND30, 2019; Jun et al., 2014).

Large-scale maps of land cover, land use or land change are produced by automatic or semi-automatic approaches; examples are the Forest Change map (Hansen et al., 2013) and the GlobeLand30 land cover map (Chen et al., 2015). Deilami et al. (2015) and Devi and Jiji (2015) provide insights into advances in land change detection techniques.

Training and validation of land change maps rely on higher quality reference data. Collecting reference data (both for training and validation) for large-scale land change maps through field visits typically is too costly or otherwise infeasible. Instead, visual interpretation of sampled areas on images with higher resolution can be performed. Such approach resembles the validation practice

adopted for GlobeLand30 (Chen et al., 2015), the Forest Change map (Hansen et al., 2013) and training data collection by Copernicus Global Land Service (CGLS, 2019). Analysing data of higher resolution by means of visual interpretation is assumed to deliver data of higher quality than the maps being generated or assessed (Olofsson et al., 2014). However, visual interpretation is difficult to replicate because different interpreters may disagree in their interpretations (Jia et al., 2016; McRoberts et al., 2018; Pengra et al., 2019; Powell et al., 2004; Tarko et al., 2019). Similarly, when detecting changes, different interpreters produce different results (Coppin et al., 2004). Such disagreement induces bias in the accuracy or area estimations, for example, in the stratified estimator of forest proportion (McRoberts et al. 2018). To increase consistency of visually interpreted reference data, various methods have been proposed and implemented. Zhao et al. (2014) included a review of acquired data, Tsendbazar et al. (2018) and Tarko et al. (2019) implemented a review with feedback, Pengra et al. (2019) trained interpreters before the assignment, McRoberts et al. (2018) recommend to use input from at least three experienced interpreters and Powell et al. (2004) applied agreement of multiple interpreters. Whatever the way forward to increase the consistency of visually interpreted reference dataset, it is important to understand what influences the agreement between interpreters on land cover, land use and change detection.

Visual interpretation of a remotely sensed image received some attention at the beginning of the discipline's development, but current focus has shifted towards automated methods of image analysis (Bianchetti and MacEachren, 2015). Only a few studies have aimed to increase understanding of the causes of inconsistency in visual image interpretations. For example Gardin et al. (2011) designed a web-based assessment to determine human factors that influence operator performance. Operator's performance was mainly determined by non-cognitive and cognitive personality factors. They found that some geographic objects were more difficult to accurately digitise than others (e.g. lamp posts were more difficult to digitise than trees) while there was also a gradual decline in performance over time (Van Coillie et al., 2014). In recent study of Pengra et al. (2019), duplicated interpretations of randomly selected reference pixels were assessed. Even though the context of the paper was land cover change, the reported results concerned labelling of land cover for different reference years. The authors focused therefore on land cover class assignment and assessed whether land cover type have an impact on consistency between interpretations. Results showed that agreement between interpreters varied per land cover class and per region. Land cover classification from different time frames allows to detect changes by post-classification comparison and to derive land cover change matrices (Coppin et al., 2004).

Another way to characterise changes focuses on classification of specific types of land transformation (hereafter called land change process) rather than on land cover type (Comber and Wulder, 2019; Lesiv et al., 2018b). Similar to the influence of land cover types on interpretation consistency (Pengra et al., 2019), land change process type could have an influence on the interpretation consistency. However, this influence has not been addressed in previous studies.

Lesiv et al. (2018) made an inventory of the current availability of very high resolution (VHR) satellite images from Google Earth (GE) and Microsoft Bing Maps across the globe. Authors emphasised the importance of image availability for monitoring changes in cropland areas with visual interpretation. Pengra et al. (2019) found that larger availability of images slightly improved the quality of image interpretation for land cover. However the influence of number of available images on the consistency of land change interpretations on agricultural land was previously not assessed.

The aim of this work is to assess whether image availability, land change processes or both influence consistency of multiple interpretations of land transformation by different individuals. Within the context of the European Union's Common Agricultural Policy (CAP), we report on an experiment involving agricultural land changes in Europe.

5.2 Methods

We designed an experiment to assess the impact of the number of VHR images available in GE and land change processes on the agreement on visual interpretation of agricultural land change detection. The experiment was conducted using GE as the interface. Based on FAO (2019), agricultural land was defined as an area for which land use is devoted to agricultural food production, with one of the following land types: arable land (non-irrigated, permanently irrigated, rice fields), permanent crops (vineyards, fruit trees, berry plantations, olive groves), and permanent pastures (sown). The interest area was Europe, thus the available reference for our experiment was CORINE 2012-2018 LC change map (CLC CH, CORINE, 2019).

5.2.1 Experimental setting

Land transformation and land change process types were derived from the land cover types from CLC CH legend: first we established three main types of land transformations, as listed in Table 5.1, first column. Next, within land transformation types, we defined eight categories of land change processes

involving agricultural land (listed in Table 5.1, second column). Finally, we used CLC CH legend Level 1 to define land types that belong to each change process. In Table 5.1, in third and fourth columns the land types are listed.

To analyse the impact of the number of images available, we created ordinal intervals representing the most typical scenarios: a minimal input data requirement for change detection (2 - 3 images), the most common case of image availability in GE (4 - 8 images), and an exceptional case with large image availability (≥ 9).

Table 5.1 lists the levels of the main factors considered:

- eight levels for the categorical land change process variable listed in the second column;
- three levels for the ordinal number of images (three rightmost columns).

Accordingly, there are 24 cells for the factorial interactions.

To create strata representing change on agricultural land we used 9,344 CLC CH polygons corresponding to land transformation from non-agricultural land into agricultural land and 25,860 polygons corresponding to land transformation from agricultural land into non-agricultural land (Table 5.1, first column). From those we selected 18 polygons for each land transformation. The selection was made maintaining a diverse representation of land cover types (Table 5.1, third and fourth columns) and assuring six polygons for each land change process. We selected polygons until the following conditions were met: visual verification of GE images confirmed the presence of designed process and the desired number of images was available. From each polygon one point was sampled. Additionally, 12 points not identified by the CLC CH (but within the interest area) were randomly selected to represent "no change". Points were selected until visual interpretation confirmed the presence of designed land type and the desired number of available images.

We created circles with 10m radius around each point (in QGIS) using buffers in the Web Mercator projection (EPSG: 3857). Sites were next exported in random order to KML format using the altitude mode for placing the feature on the ground surface (https://developers.google.com/kml/documentation/altitudemode).

Table 5.1. Experiment design: site's id numbers for each interaction (cell).

Land transformation	Land change process	Land type		Number of images*		
		From	To	2 - 3	4 - 8	≥ 9
Change from non-agricultural land into agricultural land	Agricultural land expansion	Natural land: forest, bare, shrub, natural grassland	Agricultural land	1, 2	3, 4	5, 6
	Ruralisation	Urban	Agricultural land	7, 8	9, 10	11, 12
	Reduction of water surface	Water	Agricultural land	13, 14	15, 16	17, 18
Change from agricultural land into non-agricultural land	Agricultural land abandonment and reforestation	Agricultural land	Natural land: forest, bare, shrub, natural grassland	19, 20	21, 22	23, 24
	Urbanisation	Agricultural land	Urban	25, 26	27, 28	29, 30
	Expansion of water surface	Agricultural land	Water	31, 32	33, 34	35, 36
No change	No change on agricultural land	Agricultural land	Agricultural land	37, 38	39, 40	41, 42
	No change on non-agricultural land	Non-agricultural land	Non-agricultural land	43, 44	45, 46	47, 48

* Numbers in the table refer to site id numbers.

In total, 48 sites were selected, two for each interaction (cell), so that a balanced factorial design was obtained. To avoid the mental fatigue effect (Van Coillie et al., 2014), we aimed for a maximum of one hour for interpreting all sites, thus the limited number of sites.

We offered three response options to the interpreters, as listed in the first column of Table 5.1. During the experiment, the interpreters worked independently. They were asked to interpret available GE historical imagery for each site and to decide if there was an agriculture-related change in the period 2010-2018. They could set their favourite zoom level and were allowed a single response option per site. If only part of the site changed, interpreters had to use the majority rule (i.e. sites were to be labelled based on the largest change/no-change area within the circle). In June 2019, 18 students of Wageningen University attending the final course of the first year Master Geo-

Information Science (Remote Sensing and GIS Integration) participated in the experiment. This group of interpreters was deemed homogeneous in terms of visual interpretation experience.

Since the analysis involved interpretation consistency rather than the land transformation label itself, the latter were ranked on frequency of occurrence and the analysis was conducted on the rank number rather than the land transformation label. For example, if six interpreters labelled a site as "Change from non-agricultural land into agricultural land", two assigned "Change from agricultural land into non-agricultural land" while ten decided "No change", these categories were ranked and labelled as 2, 3 and 1, respectively.

Commonly used measures of statistical heterogeneity are Gini's index and Shannon's entropy (Eliazar and Sokolov, 2010). We used Shannon entropy (Shannon, 1948; Tarko et al., 2015) as a response variable to measure the dispersion of the ranked answers: the higher the entropy, the higher the disorder, meaning less consistent interpretations. Shannon entropy (further called entropy) was calculated using base two for the logarithm.

Response variable was calculated over aggregated ranked answers for each factorial interaction (cell) and each level of main factors (land change processes and image availability).

5.2.2 Statistical testing

To assess both marginal (main) effects and interactions of the two factors, we conducted multiple comparison tests. Conventional two-way ANOVA testing of the dependent variables was not possible because after computing entropy for each factorial interaction there would be no remaining degrees of freedom. Moreover, our dependent variables were derived from ranked categorical answers obtained by sampling; uncertainty in the dependent variables would have to be accounted for. Alternatively, we relied on pairwise permutation tests of main factors as well as all factorial interactions.

We chose $\alpha = 0.05$ and computed $P(d_{sim} > |d_{obs}|)$, where d_{sim} is the simulated (Monte Carlo) test statistic (i.e., difference between entropy for two elements in the pairwise comparison) and $|d_{obs}|$ is the absolute value of the observed difference between the two. Aggregated ranked answers were permuted 100,000 times over the pairwise compared elements, without replacement. Response entropies were calculated and $P(d_{sim} > |d_{obs}|)$ was computed (Phipson and Smyth, 2010). To compensate the chance of rare events due to multiple comparisons we applied the Bonferroni correction to α (Bonferroni,

1936). Statistically significant differences ($P < \alpha_B$; α_B being the Bonferroni adjusted α) were indicated by different letters assigned to all pair-tests (Radhika et al., 2008). To facilitate interpretation, interaction effects were only calculated for cells having the same marginal effect. Permutation tests were calculated using R software for statistical computing (R Core Team, 2017).

5.3 Results

5.3.1 Exploratory analysis

There were no two interpreters agreeing on the interpretations of all 48 sites (not shown), but for twelve sites (Figure 5.1, sites 7, 9, 11, 37, 39, 42-46 and 48) all interpreters assigned the same land transformation type. Twenty-three sites (e.g. sites 2, 20, 40) were assigned two land transformation types while the remaining site received all three types. The sites having most inconsistent results are sites 17 and 38, which each had 50 percent assigned to two types. For 19 sites at least one land transformation type was indicated by only one interpreter (e.g. sites 13, 15, 29, 31, 47). It is worth mentioning that ten of these cases concern the same single interpreter (sites 4, 6, 21, 30, 31, 33, 36, 40, 41, 47).

For six sites (sites 16, 19, 21, 22, 24, 28), the land transformation types assigned by the majority of the interpreters disagreed with the types based on the CLC CH. While according to CLC CH site 16 belonged to "change form non-agricultural land into agricultural land" and sites 19, 21, 22, 24, 28 to "change from agricultural land into non-agricultural land", the majority of the interpreters interpreted them as "no change".

Overall consistency for the interpretations, calculated as the majority percentage of ranked interpretations over all sites, was 84%.

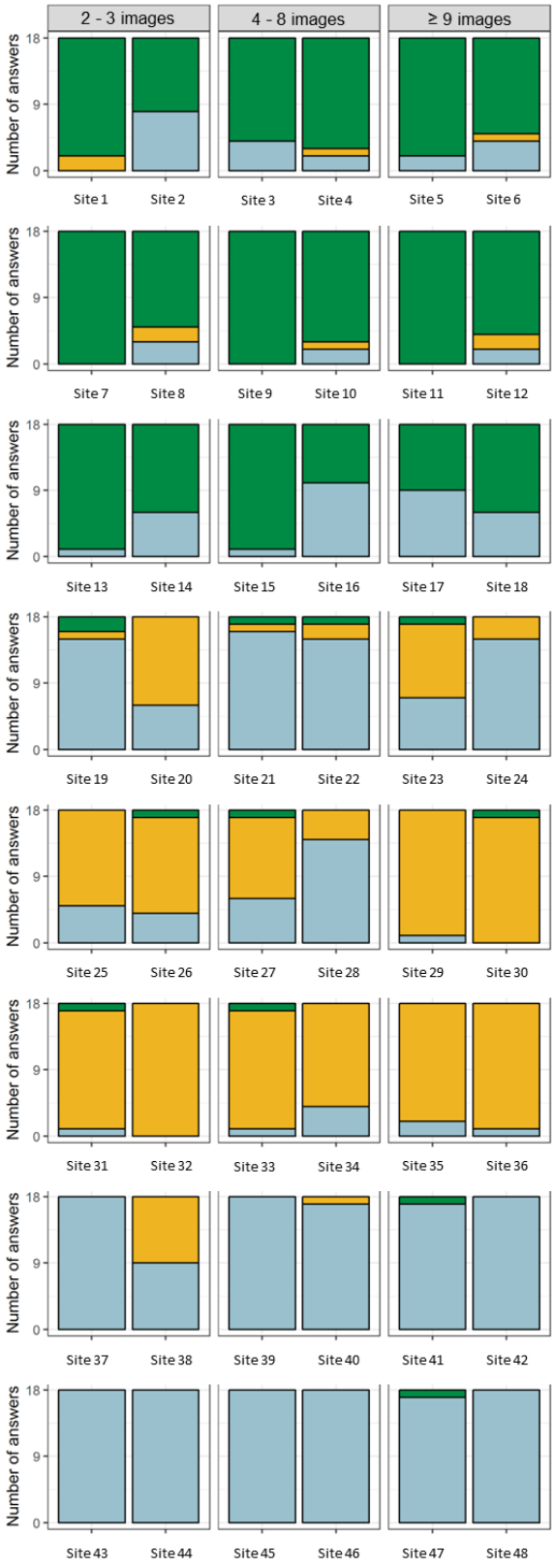
Land transformation (CLC CH)	Land change process (CLC CH)	Interpretations
Change from non-agricultural land into agricultural land	Agricultural land expansion	
	Ruralisation	
	Reduction of water surface	
Change from agricultural land into non-agricultural land	Agricultural land abandonment and reforestation	
	Urbanisation	
	Expansion of water surface	
No change	No change on agricultural land	
	No change on non-agricultural land	

Figure 5.1. Land transformation types assigned to each site by the interpreters. Colours refer to the land transformation types according to the experimental design (derived from CLC CH) and those assigned by the interpreters.

5.3.2 Statistical testing

The entropies for each level of the marginal effects of the number of images and land change process type, as well as factorial interactions are shown in Table 5.2. Land change process type was found to significantly influence entropy. The most consistently interpreted "no change on non-agricultural land" process (according to CLC CH) was different from all other processes except "no change on agricultural land" and "expansion of water surface". The process type "reduction of water surface" had one of the highest entropies caused by interpretation inconsistency between two labels (cf. Figure 5.1), which caused it to be statistically different from both "no change" classes in the two bottom rows of Table 5.2. No further marginal effects were found, including absence of marginal effects for the number of images.

Table 5.2. Shannon entropy for main factors (grey cells) and interactions (white cells). Different upper case letters indicate significant differences between levels of main factor land change process. Different lower case letters indicate significant differences between interactions within internally homogeneous groups according to the marginal (main) effect. The number of images had no significant marginal effect on entropy. Within internally homogeneous marginal effect group indicated with letter "B" there were no significant interactions.

			Number of images		
			2 - 3	4 - 8	≥ 9
		Entropy	0.76	0.69	0.68
Land change process	Agricultural land expansion	0.86	0.85	0.83	0.83
		A B			
	Ruralisation	0.61	0.72	0.49	0.61
		A B			
	Reduction of water surface	0.86	0.71	0.81	0.98
		A			a
	Agricultural land abandonment and reforestation	0.93	0.94	0.72	1.02
		A B			
	Urbanisation	0.84	0.98	1.02	0.31
		A B			b
	Expansion of water surface	0.54	0.37	0.76	0.41
		A B C			
	No change on agricultural land	0.47	0.81	0.18	0.18
		B C	c		
	No change on non-agricultural land	0.08	0.00	0.00	0.18
		C	d	d	

Even though the number of images by itself had no marginal effect on interpretation consistency as expressed by entropy, a few significant interactions were found. Entropy of sites categorised by CLC CH as "reduction of water surface" having nine or more images, significantly differed from that of "urbanisation" having nine or more images. The latter case was more consistently interpreted. Very consistently interpreted "no change on non-agricultural land" with eight or less images significantly differed from "no change on agricultural land" with three or less images. The latter case had highest entropy (least consistency) within the "no changes" group.

5.4 Discussion

5.4.1 Consistency in land change visual interpretation

Visual interpretation of sampled areas on images with higher resolution is often used as reference data for large-scale maps, but the interpreters inconsistency needs to be taken into account (Olofsson et al., 2014). Agreement of multiple interpreters on the same sites is one of the ways forward (Powell et al., 2004). However, in our experiment some of the sites, even those with high consistency, had a different prevalent land transformation interpretation than the one from CLC CH (Figure 5.1, sites 16, 19, 21, 22, 24, 28) meaning that either CLC CH map or the majority of the interpretations is incorrect. It further confirms that when classifying reference data based on majority agreement of visual interpretations, the possible erroneous assignment of the class to the site needs to be considered in the assessment of maps (Foody, 2010). To start with, a protocol to estimate and monitor interpreter consistency for visually collected large-scale reference dataset proposed by Pengra et al. (2019) could be implemented.

Our study on the agreement of land change interpretation revealed 84% overall consistency. This result is consistent with figures reported in previous work on subjectivity and inconsistencies in visual interpretations of land cover obtained by different approaches (Pengra et al., 2019; Powell et al., 2004; Tarko et al., 2019).

In an experiment set up by Powell et al. (2004), a group of five trained individuals produced reference data by visual interpretation of aerial videography. The assigned land cover type differed between the interpretations at almost 30% of the sites. Similarly, Tarko et al. (2019) reported that 30% of interpreted sites were fully or partially updated upon review of regional experts' interpretations. Wherein the latter study, land cover interpretation concerned over 15,000 sites collected as reference data for a global land cover map. Pengra

et al. (2019) reported an overall agreement of 88% of duplicate interpretations of almost 3,000 pixels. While the studies of Pengra et al. (2019), Powell et al. (2004) and Tarko et al. (2019), concerned land cover interpretations, our current experiment is about interpretation of land transformations. Our experiment had a smaller number of sites (48) than the experiments mentioned above, although each site in our study was interpreted by a larger number of interpreters (18). The larger number of interpretations per site allowed assessment of interpersonal differences with more precision compared to the precision obtained with duplicate interpretations of the same site (as adopted by Pengra et al. (2019)). For example, the impact of one individual who in our experiment interpreted land transformations for ten test sites differently than the all other interpreters (Figure5.1, sites 4, 6, 21, 30, 31, 33, 36, 40, 41, 47) would be dramatic if only duplicate interpretations would be used. Our limited number of sites is likely to underrepresent diversity in the considered land change processes. Still, land change process turned out to have significant impact on interpretation consistency.

The design of our experiment aimed at minimising the influence of personal factor on interpretation consistency by choosing a homogeneous group in terms of visual interpretation. Despite that, there was the earlier mentioned case of one interpreter who disagreed with all other interpreters. This result indicates the difficulty of choosing a homogeneous group of interpreters. It further confirms the finding of Tarko et al. (2019) who found interpersonal differences to be the most important explanatory factor in visual interpretation of global land cover.

5.4.2 Effect of the land change process types on the consistency of visual interpretations

No change processes on both non-agricultural land and on agricultural land (two bottom rows of Table 5.2) had lowest entropy meaning they were most consistently interpreted. Sites designed as the ones representing presence of change on agricultural land were less consistently interpreted. This demonstrates the difficulty in consistently interpreting land changes. An exception in generally consistently interpreted "no change" processes was cell "no change on agricultural land" with less than four images. This cell had the highest entropy within the no change group and it significantly differed from cells where all interpreters were consistent (process "no change on non-agricultural land" with less than nine images, Table 5.2). This cell's high entropy resulted from inconsistent interpretation of site 38 (Figure 5.1). Despite the small number of available images for site 38, the acquisition dates were favourable for visual interpretation: one image date was close to the beginning

and the other to the end of the period over which changes had to be assessed. It is likely that land type, rather than low number of available images with favourable acquisition dates influenced interpreters' agreement in this case. To address the disagreement on land type between the interpretations, a dedicated training could be provided to the interpreters. Such training should focus on interpretation of land types belonging to transformation processes with least consistency, like natural land (Table 5.1, third and fourth columns). When designing stratified sampling to acquire reference data on land change detection, existing land cover, land use and change maps relative to the researched change timeframe can be used. Based on the expected type of land change, additional resources can be allocated to aid visual interpretation of land change processes which require most efforts.

Confusion between land use and land cover may lead to interpretation errors (Brus et al., 2018). In our experiment, misunderstanding of the concepts of land cover and land use may have caused the difference between CLC CH and the most consistent answer for sites 19, 21 and 24 (Figure 5.1). According to CLC CH, these sites belonged to the process "agricultural land abandonment and reforestation" but interpreters consistently labelled them as "no change". For example, according to the detailed legend of CLC CH, site 19 (Figure 5.1) was transformed from agricultural land into a sport facility (a golf field). A golf field (cultivated and sown grassland) is easily misinterpreted as agricultural land cover, but land use is non-agricultural. For sites 21 and 24 (Figure 5.1), where according to CLC CH the agricultural land was abandoned, natural land such as forest or natural grassland may have been interpreted as orchard and pasture. Similar to findings of Jia et al. (2016) and Pengra et al. (2019), these examples further illustrate that some land classes are more challenging for visual interpretation than others. Interpretation consistency of more challenging classes can be improved by providing training and feedback to the interpreters (Tarko et al., 2019; Zhao et al., 2014).

5.4.3 Effect of the number of images on the consistency of visual interpretations

The absence of a marginal effect of the number of available images (reported in section 5.3.2, Table 5.2) contrasts with results in Pengra et al. (2019) who found that increasing availability of Google Earth and Landsat data coincided with increase of visual interpretation consistency. Pengra et al. (2019) graphed overall interpreter agreement (based on almost 3,000 sites interpreted by 2 individuals over 33 years) with image data density and assessed the agreement of change in years 2001 to 2006 when the increase in image availability was observed. Increasing agreement in time was observed for cropland class,

grassland and shrubs, trees and "disturbed" class. However, the increase in annual overall interpretation agreement improved by only slightly more than 1%. The rather small sample size of our experiment does not allow detecting such small effect of image availability on interpretation consistency. Similar to Pengra et al. (2019), Table 5.2, rows 1, 5, 7 suggest increased consistency when more images are available, but this observed trend was non-significant. Unlike number of available images, land change process type did have a significant impact. Based on personal experience, Olofsson et al. (2014) suggested that for detecting some change types the date of image acquisition is more important than the total number of available images. Jia et al. (2016) highlighted that for land cover types where seasonal pattern aids interpretation, multiple images from one year should be used. Since the number of images had null (this study) or slight (Pengra et al., 2019) influence on visual interpretation consistency, further research could assess the influence of other image characteristics on visual interpretation consistency. For example, the influence of image acquisition date and number of images per season could be assessed.

5.5 Conclusions

Despite the importance of visual interpretation for acquiring reference data, little attention has been paid to factors influencing the quality and consistency of interpreted land changes. This study assessed whether image availability and land change processes influence consistency of multiple visual interpretations using Shannon entropy as the response variable.

Interpretation of change detection substantially differed between interpreters confirming the subjectivity of land change interpretation. Consistency of interpretations of land transformation was influenced by the land change process, while image availability had no significant effect. The process type "no change on non-agricultural land" was the most consistently interpreted, while least interpretation consistency was for "agricultural land abandonment and reforestation" processes.

Within land change processes, difference between land cover and land use concepts influenced the consistency of interpretations, especially processes "agricultural land abandonment and reforestation". Agricultural land transformations involving interpretation of natural land were least agreed upon. To reduce confusion between the concepts of land cover and land use, dedicated training can be provided to the interpreters.

Collection of reference data for land change mapping projects should take into account previous land cover, land use and land change maps. Those maps can be used for designing stratified sampling in a way to allocate more resources to acquire reference data on land transformation processes with least consistency. Furthermore, in addition to the available number of images for interpretation, the acquisition date of images could be considered when collecting reference data. The acquisition date should correspond to the evaluated land change period.

Based on this experiment's findings, it is recommended to pay special attention to land change processes involving agricultural land abandonment, reforestation and agricultural land expansion when planning a visual change detection.

Acknowledgements

We would like to thank students from Wageningen University, Remote Sensing and GIS Integration course for participating in the experiment.

Chapter 6

Synthesis

6.1 Main findings

This thesis aimed to identify factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring. The four research questions addressing the main objective are defined in chapter 1, section 1.5. The key findings for each question are summarised below, based on the results from previous chapters.

6.1.1 What image characteristics are preferred by visual interpreters?

Satellite images are used for updating and for quality assurance of the land monitoring system implemented by EU Member States, LPIS. Correct determination of agricultural land through visual interpretation requires images of sufficient quality (LPISQA, 2012). Geometric image quality is well understood and can be described using statistics such as mean-squared error (Smits et al., 1999). However, little is known about preferred photometric image characteristics such as image brightness and contrast settings, colour composite bands, for visual image interpretation.

Chapter 2 reports on a survey aimed at understanding users' preferred brightness and contrast ranges for images used in visual interpretation for LPIS purposes. Results showed that appreciation of image photometric settings is stable at the individual level, but preferences vary across respondents. For visual interpretation of agricultural land, false colour composite images were found to be preferred over the natural colour images. The results further revealed strong subjectivity in image appreciation. It was recommended to allow interpreters to personalise settings, such as brightness and contrast values, and to choose the displayed band combination from at least four spectral bands. Setting image colour composite display, brightness and contrast require certain level of expertise and experience from the user. Moreover, it was found that there is a need to design image quality standards for areas with cast shadows within LPIS context. Cast shadows covering agricultural land may prevent successful land monitoring. This finding directed the further line of research.

6.1.2 What is the agreement between visual interpretations and automatically detected cast shadows?

As identified in chapter 2, cast shadow may hinder successful visual image interpretation for agricultural areas, for example within the context of LPIS. Cast shadow identification and quantification on an image are therefore important. While there are several automated methods for shadow detection, their performances are often compared to a single visual interpretation result

treated as "the truth"(Adeline et al., 2013; Tsai, 2006). In chapter 3, a simple thresholding method for shadow detection was proposed to address this research question. This method was compared with visual interpretation of cast shadows on agricultural land performed by twelve experienced interpreters. The agreement between visual interpretations and automated detection of cast shadows was assessed.

Results from chapter 3 showed that the cast shadow visual interpretation greatly differed between interpreters. This confirms that visual interpretation is subjective and hardly reproducible. Compared to the automated method, visual interpreters labelled more cast shadow in small patches and on shadow boundaries. The intersection between visually interpreted shadows was small, indicating the lack of agreement among interpretations. The total shadow area labelled by the automated method was in between the intersection and the union of the areas annotated by multiple interpreters. Benefits of the automated procedure are its reproducibility and its independence of visual interpretation skills. Lack of consistency in visual interpretations challenged the traditional evaluation procedure of automatic method, often based on visual interpretation performed by only one interpreter. The issue of subjectivity in visual interpretation directed the further researches.

6.1.3 How can the consistency of land cover reference data acquired by visual image interpretations be improved?

Data acquired by visual interpretation of land cover using VHR images often serves as a reference for large-scale mapping, which implies it is deemed of higher quality than the maps being generated or assessed (Olofsson et al., 2014). An excellent opportunity to research visual interpretation consistency on a large sample is acquisition of reference data for a global land cover dataset. Such data was acquired through the means of visual interpretation within the Copernicus Global Land Service: Dynamic Land Cover project (CGLS, 2019).

In chapter 4, the rate of land cover updates following review and feedback on visual interpretations performed by 23 interpreters. The number of updates following feedback differed substantially between interpreters. Despite those differences, feedback on submitted interpretations induced a positive learning effect in land cover visual interpretation for the majority of the individuals. Including feedback loops in the collection process of visually interpreted data is, therefore, expected to increase the consistency of acquired land cover reference data.

In this research, we further studied factors influencing interpretation consistency. The strongest predictor of interpretation consistency was the identifier of the interpreter. This again confirmed the subjectivity of visual interpretations, outreaching other factors, such as land cover type and image availability. Knowing that image availability and land type are considered as important factors influencing interpretation, the final research within this thesis was designed to assess the influence of those factors on interpretation consistency in conditions where the interpreter's experience is homogeneous.

6.1.4 Which are important factors influencing interpretation consistency of land change reference data?

Accurate identification of land change is of great importance for agricultural land monitoring. Similar to visual interpretation of land cover, interpretation of land changes is subjective and interpreters often produce different results.

Chapter 5 assessed the influence of image availability and land change process type on consistency of land transformation interpretations. It was based on an experiment involving 18 interpreters. Each interpreter labelled 48 sites using VHR images. Sites were scattered over Europe. The group of interpreters taking part in the experiment was deemed homogeneous in terms of visual interpretation experience. This way the personal influence of individual interpreter was reduced. Interpreters assigned land change types related to agricultural land. Results show that land change process type had a significant impact on the consistency of visual interpretations, while the number of images was non-significant. Absence of change was most consistently identified. On the contrary, agricultural land abandonment and reforestation were the least agreed upon.

6.2 Reflection and outlook

There are many factors influencing visual image interpretation of agricultural land. Doubtlessly, the individual aspect of interpreters is highly important and leads to subjectivity in visual image interpretation. Interpretation consistency is also affected by land cover, land use and land change type. Furthermore, consistency of visual interpretation is influenced by the adopted method of enhancing interpretation agreement, such as reviews and feedbacks. In the sections below I reflect on possibility to replace visual image interpretation with automated image classification in the context of reference data acquisition and on the role of visual image interpretation for agricultural land monitoring.

6.2.1 Visual versus automated image interpretation

Visual interpretation of digital images is one of the most common forms in remote sensing analyses. It uses the human ability to qualitatively evaluate spatial patterns in an image. However, it is laborious and can be costly. Moreover, human ability to interpret spectral patterns is hard to replicate due to inconsistency in interpretations as stated in chapters 3 through 5. Visual interpretation is often aided or replaced by the automated image classifications (Foody and Mathur, 2004; Ma et al., 2017). The results of chapter 3 showed that automated methods can enhance or even replace the visual interpretation in case of cast shadow detection. The automated method could be considered as another interpreter, yet giving reproducible results after calibration.

Automated methods are reproducible and minimise human efforts. Recently, deep learning based methods have been exploited and researched in the remote sensing domain (Zhu et al., 2017). Together with advances in artificial intelligence, big data processing and hardware innovation, new standards in terms of processed data volumes are set. Nowadays land cover classification based on as much as 20 terabytes of high resolution satellite images from the National Agriculture Imagery Program can be performed in as little as ten minutes (Sirosh, 2018).

Automated methods also have drawbacks. Many of them rely on parameters to be tuned by the user or other human input, such as threshold values (Gamanya et al., 2007), interpretation keys or a training set of images (Barbieri et al., 2011; Postadjian et al., 2017; Qian et al., 2014). Accuracy assessment of automated methods require trustworthy data to compare with. In addition to the accuracy assessment, calibration of the methods require even larger number of training data. Visual interpretation is often used to collect reference data, but when performed by only one interpreter (Adeline et al., 2013; Tsai, 2006) it may be biased and actually far from the truth.

Following the numerous advances in machine learning, along with the increasing amount of earth observation data available, constantly increasing reference datasets (Smets et al., 2017; Zhao et al., 2014) and continuously growing computational capacities, it is likely that automated methods will largely replace visual interpretation. Automated methods, depicting better and better interpreter's cognitive reasoning will be merely aided by the human's input (Postadjian et al., 2017). Automated and visual image interpretation methods will likely coexist, because human input is indispensable and visual interpretation is foreseen to facilitate automated models (Robinson et al., 2019). Newly arising issues, such as appearance of new crops or new agricultural

spatial patterns (e.g. vertical agriculture, Goodman and Minner, 2019) will also need the input from the human interpreter. Automated and visual image interpretation shift from being perceived as separate, parallel data acquisition methods into complementary elements of a single image interpretation methodology.

6.2.2 Reference data acquisition

Data acquired by means of visual interpretation often serve as reference for map training and map validation. Such data also serve as a reference allowing assessment of an automated classification method (Adeline et al., 2013; Tsai, 2006). Methods for accuracy assessment of image classification products have been reviewed and described by, among others, Foody (2002). Following advances in accuracy assessment methods, evaluation of map accuracy is regarded as an important and fundamental issue (Foody, 2010). Good practices in accuracy assessment of land cover and land change have been summarised by Olofsson et al. (2014). Despite of evolution of accuracy assessment methods into more formal ways, the reference data themselves are error-prone. Imperfection of visually interpreted reference data (described in chapters 4 and 5) has to be taken into account.

Subjectivity of the visual image interpretation is a fact. There are two main approaches to tackle the imperfection of reference data caused by the subjectivity of visual interpretation. One way would be to acknowledge inconsistency, measure it and include in the assessment. Foody (2010) illustrated methods to reduce (or even remove) the impact of imperfect reference data on land cover change. Methods ranged from a simple algebraic means allowing estimation of actual values to a latent class analysis allowing assessment of classification accuracy and estimation of change extent on land without the use of ground reference data. The latter technique (latent class analysis) assigns a class using multiple classifications. To explore the suitability and limitations of latent class modelling in remote sensing and to develop methods taking into account the impact of interpretation inconsistency, vast amounts of data could be used, for example from crowdsourcing. Volunteered geographic information (VGI) is a form of crowdsourcing, where citizens collect georeferenced earth observation data through web sites (Goodchild and Li, 2012) and often use readily available images from applications such as Google Earth and Bing Maps. VGI is seen as a potentially powerful way of increasing the supply of data because acquisition cost is low (compared to the cost of data acquisition by the experienced and trained interpreters) and data is collected in vast amounts. On the other hand, there are a number of concerns over the subsequent use of VGI data, in particular over data quality (See et al.,

2013). VGI data collection needs to be as simple as possible, opting for one yes/no question being asked (Sturn et al., 2015).VGI is also increasingly being used to support land cover validation (See et al., 2013). The reliability of such data requires further research (Goodchild and Li, 2012; Zhao et al., 2017).

The other approach to tackle the imperfection of reference data caused by the subjectivity of visual interpretation would be to minimise the inconsistency of the visually interpreted reference data. There are various ways to improve data acquisition, such as involving multiple trained and experienced interpreters (Powell et al., 2004), training them (Pengra et al., 2019), including a review (Zhao et al., 2014) or including review and feedback procedures (Tsendbazar et al. (2018) and chapter 4 (this thesis)). However, collected data will never be perfect. It is known that some land types are more challenging to be correctly interpreted than other (Pengra et al., 2019) while some areas have poorer image availability to facilitate interpretation (Lesiv et al., 2018a). Lack of a standard international classification system for defining land use and land change additionally complicates the situation. Such standardised land use and land change classification system is particularly important for agricultural land monitoring. Standardised classifications facilitate data comparison and correlation allowing successful land monitoring (Jansen and Gregorio, 2002).

One option to increase the quality of visually interpreted reference data could be aiding visual interpretation with methods from machine learning and computer vision. Especially promising are recent advances in active learning, where a model exploits interpreter's knowledge by asking for visual interpretation whenever it encounters a particularly ambiguous data sample (Kellenberger et al., 2019; Tuia et al., 2011). Kellenberger et al. (2019) recently trained a convolutional neural network to detect animals and found that less than half a percent of image annotated by interpreters are enough to adapt to a new challenging set of UAV images and automatically find almost 80% of the objects in it. Similar models could automatically provide interpretation of reference datasets, target the most challenging sites and ask the human interpreter (or multiple interpreters) for an analysis of only those sites. Such approach would limit the workload involved in visually interpreted data acquisition and reduce subjectivity. In the future, reference data acquisition with the means of visual interpretation is likely to be indispensable and needed, but could be limited to the exceptional and unusual cases only. Examples could be newly arising issues described in next section.

6.2.3 Agricultural land monitoring

Agricultural activities on land remain to be the focus of multiple economic and environmental policies aiming to ensure food security, sustainable use of natural resources and balanced development of rural areas. Agricultural land monitoring is essential for those policies. Large-scale maps of land cover, land use and land change can be used towards those needs (Fritz et al., 2013). It can be expected that the importance of accurate large-scale maps of agricultural land will increase.

Current maps dedicated to agricultural land monitoring are often of coarse-resolution or derived from more general land cover maps serving other purposes (such as climate change) (Fritz et al., 2013). Moreover, the quality and reliability of data acquisition vary between different global maps (Fritz et al., 2013). In the advent of openly-accessible satellite images such as Sentinel, automatically derived global maps focused on real-time agricultural land monitoring can be expected in the near future.

Similarly to the idea of standard global land cover validation dataset proposed by Olofsson et al. (2012) and Stehman et al. (2012), the same type of standard global product and validation dataset, but dedicated to agricultural land monitoring, could be designed and created. While fundamental design principles for land cover validation dataset project were described (Olofsson et al., 2012) and stratified sampling designs have been recommended (Stehman et al., 2012), there is no standard idea combining classification, mapping, training and validation methodology for tackling standard product for agricultural land monitoring. As stated in chapter 1 of this thesis, a large-scale map facilitating agricultural land monitoring is essential for many policies aiming to ensure food security, the sustainable use of natural resources and the balanced development of rural areas.

With the arrival of changes in the forms of food production, legend descriptions of map units and spatial land representation (e.g. vertical agriculture) may encounter challenges. Definitions adopted for agricultural land identification need to be versatile and ready to accommodate future changes and innovations in food production. Vertical agriculture (Goodman and Minner, 2019) is likely to be one of those challenges. Current changes and trends in food production shift towards co-existence of commercial farming and city life (Goodman and Minner, 2019). Following the trend to reduce urban footprint, new forms of farming are emerging. Examples are all forms of food production related to urban buildings, including open rooftop farms and rooftop greenhouses (Thomaier et al., 2015). All these new forms of farming

will need to be interpreted, classified and mapped for the monitoring programmes of agricultural land. Issues with mapping vertical agriculture on orthogonal maps is likely to be one of the new challenges. Other issues will be unambiguous description and classification of commercial buildings with farming activities undertaken on the rooftops. Detection and distinction between buildings dedicated to indoor farming and those dedicated to non-farming purposes will likely be another challenge.

Observation of new trends in food production requires human input and as such it suggests that the role of human in interpreting agriculture is indispensable. Human interpreters will likely be needed at the moment of identification and description of newly emerged issues. Since the creation of large-scale maps require automation to achieve near-real time monitoring, the automated process will need to accommodate new issues after a while. Research part might require human input, while for an operational monitoring, automated methods can take over. For example, in a fast-changing domain such as modern agriculture, attempts to fully automate map production processes for agricultural land monitoring are likely to be unfeasible. A possible future scenario pictures automated methods that are complementary to visual image interpretation. For example, human interpreters could collect training data identifying new types of agricultural activities and feed the automated procedures with this data. This way automated image classification and visual image interpretation shift from being perceived as separate, parallel data acquisition methods to a methodology composed of complementary elements.

6.3 Future research

This thesis studied factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring. The influence of image quality, change process type and individual aspect on the interpretation consistency has been researched, but the topic is not fully explored and several directions for future research can be sketched.

The individual aspect of interpreters is highly important and leads to subjectivity in visual image interpretation. To remove subjectivity in reference data acquisition, future research could focus on replacing human interpreter with automated methods. Images could be classified not by a human interpreter but with automated method. Similar to results in chapter 3, the automated image classification could be treated as an objective interpreter. However, so far, every automated method at some stage requires training and calibration data which are collected (at least to some extent) by subjective human

interpreters. Future research could explore options to determine the minimal human input and optimise the automated image classification and map production chains.

Since the impact of subjectivity in image interpretation cannot be fully removed, it should be reduced as much as possible and accounted for. Ways to reduce visual image interpretation are discussed in chapter 4 and in section 6.2.2. Factors influencing consistency of image interpretation have been identified (chapters 2 through 5) and some recommendations have been suggested, such as review and feedback or usage of image in false colour composite. However, not all factors influencing consistency of visual image interpretation have been identified. Future interdisciplinary research combining human studies and remote sensing could help defining factors influencing consistency in visual image interpretation. Having identified the most challenging cases, comprehensive guidelines for consistent visual image interpretation could be proposed.

Possible ways to collect reference data range from involving trained and experienced interpreters, through crowdsourcing and VGI, to automated methods of land cover classification. Nowadays, there are two trends in remote sensing. The first trend is focusing on fully automatic ways to classify images. For these procedures already existing, openly-accessible reference data are used for training and calibration. Often, these reference data with unknown accuracy is treated as ground truth for assessment of automated methods for image classification. Neglecting the accuracy of input training data may impact the performance of automated methods. The second trend is the focus on the quality of visually interpreted reference data and correct accuracy assessment of map products. These two trends could be combined in a way that a standardised global reference dataset of known accuracy and quality could be obtained by visual interpretation aided by automated image classification. Active learning, a type of machine learning where the algorithm actively queries the interpreter could be a way forward to achieve the goal of a reference dataset with defined quality. Such a standard reference dataset could engage the remote sensing community to keep it updated and constantly improve it. Ignoring presence of subjectivity in reference datasets is not a solution. Since there is no reason to believe that we can leave the consequences of subjectivity in image interpretation behind us, we have to find optimal solutions to assess it.

References

Adeline, K.R.M., Chen, M., Briottet, X., Pang, S.K., Paparoditis, N., 2013. Shadow detection in very high spatial resolution aerial images: A comparative study. ISPRS J. Photogramm. Remote Sens. 80, 21–38. https://doi.org/10.1016/j.isprsjprs.2013.02.003

Aguilar, M.A., Saldaña, M. del M., 2013. Radiometric comparison between GeoEye-1 and WorldView-2 panchromatic and multispectral imagery, in: Congress INGEGRAF-ADM-AIP PRIMECA. Madrid, Spain.

Arévalo, V., González, J., Ambrosio, G., 2008. Shadow detection in colour high-resolution satellite images. Int. J. Remote Sens. 29, 1945–1963. https://doi.org/10.1080/01431160701395302

Astrand, P.J., Di Matteo, G., Wirnhardt, C., Burger, A., Vajsova, B., Walczynska, A., Hain, S., Kornhoff, A., Simon, E., 2014. VHR image acquisition specifications for the CAP checks (CwRS and LPIS QA): VHR profile-based specifications [WWW Document]. URL https://www.ng-lio.eu/Portals/0/17359.pdf (accessed 4.15.15).

Astrand, P.J., Wirnhardt, C., Biagini, B., Weber, M., Hellerman, R., 2004. Controls with remote sensing of Common Agricultural Policy (CAP) arable- and forage- area-based subsidies: a yearly more than 700-image and 3-M euro affair, in: Meynart, R., Neeck, S.P., Shimoda, H. (Eds.), SPIE 5570, Sensors, Systems, and Next-Generation Satellites VIII. p. 577. https://doi.org/10.1117/12.565580

Barbieri, A.L., de Arruda, G.F., Rodrigues, F.A., Bruno, O.M., Costa, L. da F., 2011. An entropy-based approach to automatic image segmentation of satellite images. Phys. A Stat. Mech. its Appl. 390, 512–518. https://doi.org/10.1016/j.physa.2010.10.015

Belgiu, M., Drăguţ, L., Strobl, J., 2014. Quantitative evaluation of variations in rule-based classifications of land cover in urban neighbourhoods using WorldView-2 imagery. ISPRS J. Photogramm. Remote Sens. 87, 205–215. https://doi.org/10.1016/j.isprsjprs.2013.11.007

Bianchetti, R., MacEachren, A., 2015. Cognitive themes emerging from air photo interpretation texts published to 1960. ISPRS Int. J. Geo-Information 4, 551–571. https://doi.org/10.3390/ijgi4020551

Bindings for the "Geospatial" Data Abstraction Library, 2017. The R Project for Statistical Computing [WWW Document]. URL https://cran.r-project.org/package=rgdal (accessed 11.12.17).

Bonferroni, C., 1936. Teoria statistica delle classi e calcolo delle probabilita. Pubbl. del R Ist. Super. di Sci. Econ. e Commericiali di Firenze 8, 3–62.

Boud, D., Molloy, E., 2013. Rethinking models of feedback for learning: the challenge of

design. Assess. Eval. High. Educ. 38, 698–712. https://doi.org/10.1080/02602938.2012.691462

Bregt, A.K., Denneboom, J., Gesink, H.J., Van Randen, Y., 1991. Determination of rasterizing error a case study with the soil map of The Netherlands. Int. J. Geogr. Inf. Syst. 5, 361–367. https://doi.org/10.1080/02693799108927861

Breiman, L., 2002. Manual on setting up, using, and understanding Random Forests V3.1 [WWW Document]. Stat. Dep. Univ. Calif. Berkeley. URL https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf (accessed 5.15.19).

Breiman, L., 2001. Random Forest. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Bruinsma, J., Fischer, G., Nachtergaele, F., Poulisse, J., Tran, D., Griffee, P., Nguyen, N., Bishop, C., Clarke, L., 2003. Crop production and natural resource use, in: Bruinsma, J. (Ed.), World Agriculture: Towards 2015 / 2030. An FAO Perspective. p. 97.

Brus, J., Pechanec, V., Machar, I., 2018. Depiction of uncertainty in the visually interpreted land cover data. Ecol. Inform. 47, 10–13. https://doi.org/10.1016/j.ecoinf.2017.10.015

CEOS, 2019. CEOS Working Group on Calibration and Validation: The Land Product Validation Subgroup [WWW Document]. URL https://lpvs.gsfc.nasa.gov/ (accessed 2.6.19).

CGLS, 2019. Copernicus Global Land Service [WWW Document]. URL https://land.copernicus.eu/global/index.html (accessed 2.6.19).

Chen, J.J., Chen, J.J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30m resolution: A POK-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27. https://doi.org/10.1016/j.isprsjprs.2014.09.002

Choi, M.G., Jung, J.H., Jeon, J.W., 2009. No-reference image quality assessment using blur and noise. Proc. World 38, 153–157.

Classes and Methods for Spatial Data, 2017. The R Project for Statistical Computing [WWW Document]. URL https://cran.r-project.org/package=sp (accessed 11.12.17).

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., Foody, G., 2013. Using control data to determine the reliability of volunteered geographic information about land cover. Int. J. Appl. Earth Obs. Geoinf. 23, 37–48. https://doi.org/10.1016/j.jag.2012.11.002

Comber, A., Wulder, M., 2019. Considering spatiotemporal processes in big data analysis : Insights from remote sensing of land cover and land use. Trans. GIS 1–13. https://doi.org/10.1111/tgis.12559

Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., Lambin, E., 2004. Digital change detection methods in ecosystem monitoring: a review. Int. J. Remote Sens. 25, 1565–1596. https://doi.org/10.1080/0143116031000101675

CORINE, 2019. CORINE Land Cover data, European Environment Agency (EEA)

[WWW Document]. URL https://land.copernicus.eu/pan-european/corine-land-cover (accessed 6.26.19).

da Silva, S.M., Rodrigues, S.C.M., Bissaco, M.A.S., Scardovelli, T., Boschi, S.R.M.S., Marques, M.A., Santos, M.F., Silva, A.P., 2019. A Novel Online Training Platform for Medical Image Interpretation, in: Lhotska, L., Sukupova, L., Lacković, I., Ibbott, G. (Eds.), World Congress on Medical Physics and Biomedical Engineering 2018. IFMBE Proceedings. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-10-9035-6_153

Dare, P.M., 2005. Shadow analysis in high-resolution satellite imagery of urban areas. Photogramm. Eng. Remote Sens. 71, 169–177. https://doi.org/10.14358/PERS.71.2.169

Deeb, S.S., 2005. The molecular basis of variation in human color vision. Clin. Genet. 67, 369–377. https://doi.org/10.1111/j.1399-0004.2004.00343.x

Deilami, B.R., Ahmad, B. Bin, Saffar, M.R.A., Umar, H.Z., Bahru, J., 2015. Review of change detection techniques from remotely sensed images. Res. J. Appl. Sci. Eng. Technol. 10, 221–229. https://doi.org/10.19026/rjaset.10.2575

Devi, R.N., Jiji, G.W., 2015. Change detection techniques - A survey. Int. J. Comput. Sci. Appl. 5, 45–57. https://doi.org/10.5121/ijcsa.2015.5205

Devos, W., Milenov, P., 2013. Introducing the TEGON as the elementary physical land cover feature, in: 2013 Second International Conference on Agro-Geoinformatics (Agro-Geoinformatics). IEEE, Fairfax, VA, USA, pp. 562–567. https://doi.org/10.1109/Argo-Geoinformatics.2013.6621939

Devos, W., Milenov, P., Wojda, P., Tarko, A., Franielczyk, R., 2012. The first year of implementation of the LPIS Quality Assessment in the frame of COMM. Reg. No 1122/2009 Art. 6.2. Publ. Off. Eur. Union. https://doi.org/10.2788/91513

DG AGRI, 2018. CAP explained. Direct payments for farmers 2015-2020. https://doi.org/10.2762/149509

Di Gregorio, A., Jansen, L.J.., 2000. Land Cover Classification System: Classification concepts and user manual.

Di Gregorio, A., Jansen, L.J.M., 1998. A new concept for a land-cover classification system. Land. 2, 55–65.

Digital Globe, 2017. Information Products, Standard Imagery, Specifications [WWW Document]. URL https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/21/Standard_Imagery_DS_10-7-16.pdf (accessed 12.11.17).

Eliazar, I., Sokolov, I.M., 2010. Maximization of statistical heterogeneity: From Shannon's entropy to Gini's index. Phys. A Stat. Mech. its Appl. 389, 3023–3038. https://doi.org/10.1016/j.physa.2010.03.045

ENVI, 2014. ENVI [WWW Document]. URL http://www.exelisvis.com/ProductsServices/ENVIProducts/ENVI.aspx (accessed 4.1.14).

ESA, 2019. Sentinel Online [WWW Document]. URL

https://earth.esa.int/web/sentinel/ (accessed 7.24.19).

European Commission, 2014a. Commission delegated regulation (EU) no 640/2014 [WWW Document]. Off. J. Eur. Union L 181/48. URL http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0640&rid=1 (accessed 12.11.17).

European Commission, 2014b. Budget Explained [WWW Document]. URL http://ec.europa.eu/budget/explained/myths/myths_en.cfm#9of15 (accessed 4.1.14).

European Commission, 2014c. Budget 2013 in figures [WWW Document]. URL http://ec.europa.eu/budget/figures/2013/2013_en.cfm (accessed 4.1.14).

European Commission, 2014d. JRC IES DERD Unit - Community Image Data Portal [WWW Document]. URL http://cidportal.jrc.ec.europa.eu/home/ (accessed 11.20.14).

European Commission, 2014e. Communication from the Commission to the Council and the European Parliament, Technical adjustment of the financial framework for 2015 in line with movements in GNI. COM(2014) 307 Final [WWW Document]. Off. J. Eur. Union L 181/48. URL http://eur-lex.europa.eu/resource.html?uri=cellar:1464a97d-eb02-11e3-8cd4-01aa75ed71a1.0023.03/DOC_1&format=PDF (accessed 12.11.17).

European Commission, 2013. Regulation (EU) of the European Parliament and of the Council no 1306/2013 [WWW Document]. Off. J. Eur. Union L 347/549. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R1306&from=EN (accessed 6.7.18).

FAO, 2019. FAO Term Portal [WWW Document]. URL http://www.fao.org/faoterm (accessed 6.26.19).

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., Snyder, P.K., 2005. Global consequences of land use. Science (80-.). 309, 570–574. https://doi.org/10.1126/science.1111772

Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. Remote Sens. Environ. 114, 2271–2285. https://doi.org/10.1016/j.rse.2010.05.003

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4

Foody, G.M., Mathur, A., 2004. Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sens. Environ. 93, 107–117. https://doi.org/10.1016/j.rse.2004.06.017

Fritz, S., See, L., You, L., Justice, C., Becker-Reshef, I., Bydekerke, L., Cumani, R., Defourny, P., Erb, K., Foley, J., Gilliams, S., Gong, P., Hansen, M., Hertel, T., Herold, M., Herrero, M., Kayitakire, F., Latham, J., Leo, O., McCallum, I., Obersteiner, M., Ramankutty, N., Rocha, J., Tang, H., Thornton, P., Vancutsem, C.,

van der Velde, M., Wood, S., Woodcock, C., 2013. The need for improved maps of global cropland. Eos, Trans. Am. Geophys. Union 94, 31–32. https://doi.org/10.1002/2013EO030006

FSA, 2017. Farm Service Agency, United States Department of Agriculture [WWW Document]. URL https://www.fsa.usda.gov/ (accessed 12.11.17).

Gamanya, R., De Maeyer, P., De Dapper, M., 2007. An automated satellite image classification design using object-oriented segmentation algorithms: A move towards standardization. Expert Syst. Appl. 32, 616–624. https://doi.org/10.1016/j.eswa.2006.01.055

Gardin, S., van Laere, S.M.J., van Coillie, F.M.B., Anseel, F., Duyck, W., de Wulf, R.R., Verbeke, L.P.C., 2011. Remote sensing meets psychology: a concept for operator performance assessment. Remote Sens. Lett. 2, 251–257. https://doi.org/10.1080/01431161.2010.516280

Geographic Data Analysis and Modeling, 2017. The R Project for Statistical Computing [WWW Document]. URL https://cran.r-project.org/package=raster (accessed 11.12.17).

Global Forest Change, 2019. University of Maryland [WWW Document]. URL https://earthenginepartners.appspot.com (accessed 6.26.19).

GLOBELAND30, 2019. Earth land-cover map [WWW Document]. URL http://www.globallandcover.com (accessed 6.26.19).

GNU, 2014. Image Manipulation Program [WWW Document]. URL http://www.gimp.org/ (accessed 4.1.14).

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: The challenge of feeding 9 billion people. Science (80-.). 327, 812–818. https://doi.org/10.1126/science.1185383

Gong, X., Marklund, L.G., Tsuji, S., 2009. Land Use Classification, in: 14th Meeting of the London Group on Environmental Accounting. Canberra, pp. 27–30.

Goodchild, M.F., Li, L., 2012. Assuring the quality of volunteered geographic information. Spat. Stat. 1, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Goodman, W., Minner, J., 2019. Will the urban agricultural revolution be vertical and soilless? A case study of controlled environment agriculture in New York City. Land use policy 83, 160–173. https://doi.org/10.1016/j.landusepol.2018.12.038

Hansen, M.C., Potapov, P. V, Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science (80-.). 342, 850–853. https://doi.org/10.1126/science.1244693

Harris, B., 1975. The Statistical Estimation of Entropy in the Non-Parametric Case. University of Wisconsin-Madison Mathematics Research Center, Madison, WI, USA.

Honkavaara, E., Arbiol, R., Markelin, L., Martinez, L., Cramer, M., Bovet, S., Chandelier, L., Ilves, R., Klonus, S., Marshal, P., Schläpfer, D., Tabor, M., Thom, C., Veje, N.,

2009. Digital airborne photogrammetry—A new tool for quantitative remote sensing?—A state-of-the-art review on radiometric aspects of digital photogrammetric images. Remote Sens. 1, 577–605. https://doi.org/10.3390/rs1030577

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., Van Der Laan, M., 2006. Survival Ensembles. Biostatistics 7, 355–373.

IACS, 2017. Agriculture and Rural Development [WWW Document]. Integr. Adm. Control Syst. URL https://ec.europa.eu/agriculture/direct-support/iacs_en (accessed 12.11.17).

Interface to Geometry Engine - Open Source ('GEOS'), 2017. The R Project for Statistical Computing [WWW Document]. URL https://cran.r-project.org/package=rgeos (accessed 11.12.17).

Ishihara, S., 1972. The series of plates designed as a tests for colour-blindness. Kanehara Shuppan Co., Ltd., Tokyo.

Jansen, L.J.M., Gregorio, A. Di, 2002. Parametric land cover and land-use classifications as tools for environmental change detection. Agric. Ecosyst. Environ. 91, 89–100. https://doi.org/10.1016/S0167-8809(01)00243-2

Jensen, J.R., 2000. Remote sensing of the environment: An Earth resource perspective. Prentice Hall Series in Geographic Information Science, NJ, USA.

Jia, X., Khandelwal, A., Gerber, J., Carlson, K., West, P., Kumar, V., 2016. Learning large-scale plantation mapping from imperfect annotators, in: 2016 IEEE International Conference on Big Data (Big Data). IEEE, pp. 1192–1201. https://doi.org/10.1109/BigData.2016.7840723

Jun, C., Ban, Y., Li, S., 2014. Open access to Earth land-cover map. Nature 514, 434–434. https://doi.org/10.1038/514434c

Kellenberger, B., Marcos, D., Lobry, S., Tuia, D., 2019. Half a percent of labels is enough: Efficient animal detection in UAV imagery using deep CNNs and Active Learning. IEEE Trans. Geosci. Remote Sens. 1–12. https://doi.org/10.1109/TGRS.2019.2927393

Krupinski, E.A., Williams, M.B., Andriole, K., Strauss, K.J., Applegate, K., Wyatt, M., Bjork, S., Seibert, J.A., 2007. Digital radiography image quality: image processing and display. J. Am. Coll. Radiol. 4, 389–400. https://doi.org/10.1016/j.jacr.2007.02.001

Lallé, S., Conati, C., Carenini, G., 2016. Prediction of individual learning curves across information visualizations. User Model. User-adapt. Interact. 26, 307–345. https://doi.org/10.1007/s11257-016-9179-5

Lesiv, M., See, L., Laso Bayas, J., Sturn, T., Schepaschenko, D., Karner, M., Moorthy, I., McCallum, I., Fritz, S., 2018a. Characterizing the spatial and temporal availability of very high resolution satellite imagery in Google Earth and Microsoft Bing Maps as a source of reference data. Land 7, 118. https://doi.org/10.3390/land7040118

Lesiv, M., Tsendbazar, N., Herold, M., Smets, B., Kerchove, R. Van De, 2018b. Copernicus Global Land Operations "Vegetation and Energy", "CGLOPS-1", Product Specifications, Dynamic Land Cover, Draft 1.0.

Li, Y., Gong, P., Sasagawa, T., 2005. Integrated shadow removal based on photogrammetry and image analysis. Int. J. Remote Sens. 26, 3911–3929. https://doi.org/10.1080/01431160500159347

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Lillesand, T., Kiefer, R.W., Chipman, J., 2008. Remote sensing and image interpretation, 6th ed. John Wiley &Sons, Inc., USA.

Liu, W., Yamazaki, F., 2012. Object-based shadow extraction and correction of high-resolution optical satellite images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5, 1296–1302. https://doi.org/10.1109/JSTARS.2012.2189558

LPIS TG ETS, 2017a. ETS Sampling zones - WikiCAP [WWW Document]. URL https://marswiki.jrc.ec.europa.eu/wikicap/index.php/ETS_Sampling_zones (accessed 11.12.17).

LPIS TG ETS, 2017b. Executive Summary - WikiCAP [WWW Document]. URL https://marswiki.jrc.ec.europa.eu/wikicap/index.php/ETS_Documentation (accessed 11.12.17).

LPIS TG ETS, 2014. Technical Documentation of the 2014 Implementation of Quality Assessment ETS v.5.3 [WWW Document]. URL https://marswiki.jrc.ec.europa.eu/wikicap/images/2/26/LPISQA2014_v5_3.pdf (accessed 12.11.17).

LPISQA, 2014a. Rationale - WikiCAP [WWW Document]. URL http://marswiki.jrc.ec.europa.eu/wikicap/index.php/LPISQA_1_Rationale (accessed 4.1.14).

LPISQA, 2014b. LPIS Control Zones - WikiCAP [WWW Document]. URL http://marswiki.jrc.ec.europa.eu/wikicap/index.php/LPISQA_2.b.ii_LPIS_control_zones (accessed 4.1.14).

LPISQA, 2014c. Orthoimage Technical Specifications for the Purpose of LPIS - WikiCAP [WWW Document]. URL http://marswiki.jrc.ec.europa.eu/wikicap/index.php/Orthoimage_technical_specifications_for_the_purpose_of_LPIS (accessed 4.1.14).

LPISQA, 2014d. Image Radiometric Quality Assurance - WikiCAP [WWW Document]. URL http://marswiki.jrc.ec.europa.eu/wikicap/index.php/Image_radiometric_quality_assurance (accessed 4.1.14).

LPISQA, 2014e. Annual Report - WikiCAP [WWW Document]. URL http://marswiki.jrc.ec.europa.eu/wikicap/index.php/LPISQA_2.c_Annual_report (accessed 4.1.14).

LPISQA, 2014f. Non-CwRS Image Upload - WikiCAP [WWW Document]. URL http://marswiki.jrc.ec.europa.eu/wikicap/index.php/LPISQA_2.c.i_Non-CwRS_image_upload (accessed 4.1.14).

LPISQA, 2012. Annex II: Executable Test Suite (ETS). Flow of Events , Related to the Inspection of the Reference Parcel, version 5.2 [WWW Document]. URL

ftp://mars.jrc.ec.europa.eu/lpis/Documents/v52_June2012/Annex_II_Flow_of_events_ver5_2.pdf (accessed 4.14.15).

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y., 2017. A review of supervised object-based land-cover image classification. ISPRS J. Photogramm. Remote Sens. 130, 277–293. https://doi.org/10.1016/j.isprsjprs.2017.06.001

Markelin, L., Honkavaara, E., 2008. Radiometric calibration and characterization of large-format digital photogrammetric sensors in a test field. Photogramm. Eng. Remote Sensing 74, 1487–1500.

McFeeters, S., 2013. Using the Normalized Difference Water Index (NDWI) within a Geographic Information System to detect swimming pools for Mosquito Abatement: A practical approach. Remote Sens. 5, 3544–3561. https://doi.org/10.3390/rs5073544

McGill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of Box Plots. Am. Stat. 32, 12. https://doi.org/10.2307/2683468

McRoberts, R.E., Stehman, S. V., Liknes, G.C., Næsset, E., Sannier, C., Walters, B.F., 2018. The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions. ISPRS J. Photogramm. Remote Sens. 142, 292–300. https://doi.org/10.1016/j.isprsjprs.2018.06.002

Mincer, J., 1974. Schooling, Experience, and Earnings. Hum. Behav. Soc. Institutions.

Mittal, A., Soundararajan, R., Bovik, A.C., 2013. Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. 20, 209–212. https://doi.org/10.1109/LSP.2012.2227726

Movia, A., Beinat, A., Crosilla, F., 2015. Comparison of unsupervised vegetation classification methods from VHR images after shadows removal by innovative algorithms. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 40, 1269–1276. https://doi.org/10.5194/isprsarchives-XL-7-W3-1269-2015

Nagao, M., Matsuyama, T., Ikeda, Y., 1979. Region extraction and shape analysis in aerial photographs. Comput. Graph. Image Process. 10, 195–223. https://doi.org/10.1016/0146-664X(79)90001-7

Narciss, S., 2008. Feedback strategies for interactive learning tasks, in: Handbook of Research on Educational Communications and Technology. pp. 125–144.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S. V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Olofsson, P., Stehman, S. V., Woodcock, C.E., Sulla-Menashe, D., Sibley, A.M., Newell, J.D., Friedl, M.A., Herold, M., 2012. A global land-cover validation data set, part I: fundamental design principles. Int. J. Remote Sens. 33, 5768–5788. https://doi.org/10.1080/01431161.2012.674230

Olsen, D., Dou, C., Zhang, X., Hu, L., Kim, H., Hildum, E., 2010. Radiometric Calibration for AgCam. Remote Sens. 2, 464–477. https://doi.org/10.3390/rs2020464

Otsu, N., 1979. A threshold selection method from gray level histograms. IEEE Trans.

Syst. Man, Cybern. 9, 62–66.

Paine, D.P., Kiser, J.D., 2012. Aerial photography and image interpretation, Third edit. ed. John Wiley & Sons, Inc., NJ, USA.

Patrício, D.I., Rieder, R., 2018. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. Comput. Electron. Agric. 153, 69–81. https://doi.org/10.1016/j.compag.2018.08.001

Peel, M.C., Finlayson, B.L., McMahon., T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. earth Syst. Sci. Discuss. 4, 439–473.

Pengra, B.W., Stehman, S. V., Horton, J.A., Dockter, D.J., Schroeder, T.A., Yang, Z., Cohen, W.B., Healey, S.P., Loveland, T.R., 2019. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2019.111261

Phipson, B., Smyth, G.K., 2010. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. Stat. Appl. Genet. Mol. Biol. 9. https://doi.org/10.2202/1544-6115.1585

Phiri, D., Morgenroth, J., 2017. Developments in Landsat land cover classification methods: A review. Remote Sens. 9, 967. https://doi.org/10.3390/rs9090967

Pinson, M.H., Wolf, S., 2003. Comparing subjective video quality testing methodologies, in: Ebrahimi, T., Sikora, T. (Eds.), SPIE Video Communications and Image Processing Conference. Lugano, Switzerland, pp. 573–582. https://doi.org/10.1117/12.509908

Poli, D., 2014. Mapping using high-resolution satellite imagery. EuroSDR, EDUSERV 12, Trento, Italy.

Ponomarenko, N., Lukin, V., 2009. TID2008-A database for evaluation of full-reference visual quality assessment metrics. Adv. Mod. 10, 30–45.

Postadjian, T., Le Bris, A., Sahbi, H., Mallet, C., 2017. Investigating the potential of deep neural networks for large-scale classification of very high resolution satellite images. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-1/W1, 183–190. https://doi.org/10.5194/isprs-annals-IV-1-W1-183-2017

Powell, R.L., Matzke, N., de Souza, C., Clark, M., Numata, I., Hess, L.L., Roberts, D.A., 2004. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sens. Environ. 90, 221–234. https://doi.org/10.1016/j.rse.2003.12.007

Pyka, K., 2009. Jak ocenić jakość fotometryczną ortofotomapy? Arch. Fotogram. Kartogr. i Teledetekcji 19, 363–372.

Qian, Y., Zhou, W., Yan, J., Li, W., Han, L., 2014. Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. Remote Sens. 7, 153–168. https://doi.org/10.3390/rs70100153

R Core Team, 2017. R: A Language and Environment for Statistical Computing.

Rabley, P., Yuen, E., 2009. In China, GIS-based Land Registry aims to protect farming

rights and enhance food security [WWW Document]. ArcNews. URL http://www.esri.com/news/arcnews/spring09articles/in-china.html (accessed 12.11.17).

Radhika, V., Kost, C., Bartram, S., Heil, M., Boland, W., 2008. Testing the optimal defence hypothesis for two indirect defences: extrafloral nectar and volatile organic compounds. Planta 228, 449–457. https://doi.org/10.1007/s00425-008-0749-6

Robinson, C., Ortiz, A., Malkin, K., Elias, B., Peng, A., Morris, D., Dilkina, B., Jojic, N., 2019. Human-machine collaboration for fast land cover mapping. arXiv Prepr.

Sarabandi, P., Yamazaki, F., Matsuoka, M., Kiremidjian, A., 2004. Shadow detection and radiometric restoration in satellite high resolution images, in: IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. IEEE, pp. 3744–3747. https://doi.org/10.1109/IGARSS.2004.1369936

Schank, R.C., Berman, T.R., Macpherson, K.A., 1999. Learning by doing, in: Instructional-Design Theories and Models: A New Paradigm of Instructional Theory. pp. 161–181.

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., Obersteiner, M., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. PLoS One 8, e69958. https://doi.org/10.1371/journal.pone.0069958

See, L., Fritz, S., Perger, C., Schill, C., McCallum, I., Schepaschenko, D., Duerauer, M., Sturn, T., Karner, M., Kraxner, F., Obersteiner, M., 2015. Harnessing the power of volunteers, the internet and Google Earth to collect and validate global spatial information using Geo-Wiki. Technol. Forecast. Soc. Change 98, 324–335. https://doi.org/10.1016/j.techfore.2015.03.002

Shahtahmassebi, A., Yang, N., Wang, K., Moore, N., Shen, Z., 2013. Review of shadow detection and de-shadowing methods in remote sensing. Chinese Geogr. Sci. 23, 403–420. https://doi.org/10.1007/s11769-013-0613-x

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423, 623–656.

Sheikh, H.R., Bovik, A.C., Cormack, L., 2005. No-reference quality assessment using natural scene statistics: JPEG2000. IEEE Trans. Image Process. 14, 1918–1927.

Sievert, C., 2018. plotly for R.

Singh, V.P., 2013. Entropy theory and its application in environmental and water engineering. John Wiley & Sons, Inc., NJ, USA.

Sirosh, J., 2018. Planet-scale land cover classification with FPGAs, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18. ACM Press, New York, New York, USA, pp. 2877–2877. https://doi.org/10.1145/3219819.3226068

Smets, B., Buchhorn, M., Lesiv, M., Tsendbazar, N., 2017. Copernicus Global Land Operations "Vegetation and Energy", "CGLOPS-1", Product Usser Manual, Moderate Dynamic Land Cover 100m, Version 1.

Smits, P.C., Dellepiane, S.G., Schowengerdt, R. a., 1999. Quality assessment of image

classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach. Int. J. Remote Sens. 20, 1461–1486. https://doi.org/10.1080/014311699212560

Song, X.-P., Hansen, M.C., Stehman, S. V., Potapov, P. V., Tyukavina, A., Vermote, E.F., Townshend, J.R., 2018. Global land change from 1982 to 2016. Nature 560, 639–643. https://doi.org/10.1038/s41586-018-0411-9

Speelman, C.P., Kirsner, K., 2005. Beyond the learning curve: The construction of mind. Oxford University Press, Oxford.

Stehman, S. V., Olofsson, P., Woodcock, C.E., Herold, M., Friedl, M.A., 2012. A global land-cover validation data set, II: Augmenting a stratified sampling design to estimate accuracy by region and land-cover class. Int. J. Remote Sens. 33, 6975–6993. https://doi.org/10.1080/01431161.2012.695092

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional Variable Importance for Random Forests. BMC Bioinformatics 9.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25. https://doi.org/10.1186/1471-2105-8-25

Sturn, T., Wimmer, M., Salk, C., Perger, C., See, L., Fritz, S., 2015. Cropland Capture—A game for improving global cropland maps, in: In Proceedings of the Foundation of Digital Games (FDG 2015).

Tarko, A., de Bruin, S., Bregt, A.K., 2018. Comparison of manual and automated shadow detection on satellite imagery for agricultural land delineation. Int. J. Appl. Earth Obs. Geoinf. 73, 493–502. https://doi.org/10.1016/j.jag.2018.07.020

Tarko, A., de Bruin, S., Fasbender, D., Devos, W., Bregt, A., 2015. Users' assessment of orthoimage photometric quality for visual interpretation of agricultural fields. Remote Sens. 7, 4919–4936. https://doi.org/10.3390/rs70404919

Tarko, A., Tsendbazar, N.E., Bruin, S. de, Bregt, A.K., 2019. Producing consistent visually interpreted land cover reference data: Learning from feedback. Submitted for publication.

Tewkesbury, A.P., Comber, A.J., Tate, N.J., Lamb, A., Fisher, P.F., 2015. A critical synthesis of remotely sensed optical image change detection techniques. Remote Sens. Environ. 160, 1–14. https://doi.org/10.1016/j.rse.2015.01.006

Thomaier, S., Specht, K., Henckel, D., Dierich, A., Siebert, R., Freisinger, U.B., Sawicka, M., 2015. Farming in and on urban buildings: Present practice and specific novelties of Zero-Acreage Farming (ZFarming). Renew. Agric. Food Syst. 30, 43–54. https://doi.org/10.1017/S1742170514000143

Tóth, K., Kučas, A., 2016. Spatial information in European agricultural data management. Requirements and interoperability supported by a domain model. Land use policy 57, 64–79. https://doi.org/10.1016/j.landusepol.2016.05.023

Tsai, V.J.D., 2006. A comparative study on shadow compensation of color aerial images in invariant color models. IEEE Trans. Geosci. Remote Sens. 44, 1661–1671. https://doi.org/10.1109/TGRS.2006.869980

Tsendbazar, N.-E., Herold, M., de Bruin, S., Lesiv, M., Fritz, S., Van De Kerchove, R., Buchhorn, M., Duerauer, M., Szantoi, Z., Pekel, J.-F., 2018. Developing and applying a multi-purpose land cover validation dataset for Africa. Remote Sens. Environ. 219, 298–309. https://doi.org/10.1016/j.rse.2018.10.025

Tuia, D., Munoz-Mari, J., 2012. Putting the user into the active learning loop: Towards realistic but efficient photointerpretation, in: 2012 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 75–78. https://doi.org/10.1109/IGARSS.2012.6351633

Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J., 2011. A Survey of active learning algorithms for supervised remote sensing image classification. IEEE J. Sel. Top. Signal Process. 5, 606–617. https://doi.org/10.1109/JSTSP.2011.2139193

Tuomisto, H., 2010. A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography (Cop.). 33, 2–22. https://doi.org/10.1111/j.1600-0587.2009.05880.x

USGS, 2019. Earth Explorer [WWW Document]. URL https://earthexplorer.usgs.gov/ (accessed 6.26.19).

Van Coillie, F.M.B., Gardin, S., Anseel, F., Duyck, W., Verbeke, L.P.C., De Wulf, R.R., 2014. Variability of operator performance in remote-sensing image interpretation: the importance of human and external factors. Int. J. Remote Sens. 35, 754–778. https://doi.org/10.1080/01431161.2013.873152

Van Westen, C.J., 2000. The modelling of landslide hazards using GIS. Surv. Geophys. 21, 241–255. https://doi.org/10.1023/A:1006794127521

Wang, Z., Bovik, A.C., Lu, L., 2002. Why is image quality assessment so difficult?, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Orlando, FL, USA, pp. 3313–3316.

Wei, T., Simko, V., 2017. R package "corrplot": Visualization of a Correlation Matrix.

White, A.R., 2019. Human expertise in the interpretation of remote sensing data: A cognitive task analysis of forest disturbance attribution. Int. J. Appl. Earth Obs. Geoinf. 74, 37–44. https://doi.org/10.1016/j.jag.2018.08.026

Wu, S.T., Hsieh, Y.T., Chen, C.T., Chen, J.C., 2014. A comparison of 4 shadow compensation techniques for land cover classification of shaded areas from high radiometric resolution aerial images. Can. J. Remote Sens. 40, 315–326. https://doi.org/10.1080/07038992.2014.979488

Yamazaki, F., Liu, W., Takasaki, M., 2009. Characteristics of shadow and removal of its effects for remote sensing imagery, in: International Geoscience and Remote Sensing Symposium. IEEE, pp. IV-426-IV–429. https://doi.org/10.1109/IGARSS.2009.5417404

Young, S., 2017. Land change monitoring, assessment, and projection (LCMAP) revolutionizes land cover and land change research. United States Geol. Surv. Inf. Prod. 172 2008, 4. https://doi.org/10.3133/gip172

Zhang, Y., Chandler, D.M., 2013. An algorithm for no-reference image quality assessment based on log-derivative statistics of natural scenes. SPIE 8653, 86530J.

https://doi.org/10.1117/12.2001342

Zhao, Y., Feng, D., Yu, L., See, L., Fritz, S., Perger, C., Gong, P., 2017. Assessing and improving the reliability of volunteered land cover reference data. Remote Sens. 9, 1034. https://doi.org/10.3390/rs9101034

Zhao, Yuanyuan, Gong, P., Yu, L., Hu, L., Li, X., Li, C., Zhang, H., Zheng, Y., Wang, J., Zhao, Yongchao, Cheng, Q., Liu, C., Liu, S., Wang, X., 2014. Towards a common validation sample set for global land-cover mapping. Int. J. Remote Sens. 35, 4795–4814. https://doi.org/10.1080/01431161.2014.930202

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Summary

Agricultural activities have played a key role in shaping the landscape for millennia and will continue to do so in the future. As such, it remains to be the focus of multiple economic and environmental policies aiming to ensure food security, sustainable use of natural resources and balanced development of rural areas. Agricultural land monitoring is essential for those policies.

For agricultural land monitoring, field observations of large-scale (national, multi-national to global scale) areas can be costly and often not feasible. On the other hand, remote sensing offers more efficient way to monitor agricultural land in large areas. Through remote sensing, images of the interest area can be remotely acquired by cameras on air-borne or space-borne platforms, which measure the electromagnetic energy reflected by the earth's surface. Visual interpretation of remotely sensed images is commonly applied to acquire reference data used for development and validation of large-scale land maps. Visual image interpretation is a process of identifying objects or phenomena on images. Information resulting from visual image interpretation can be presented as a map or as a tabular reference dataset.

There are many factors affecting the quality of visual image interpretation. It depends on the interpreter's training and experience, on the nature and complexity of interpreted objects or phenomena and on quality of images used for the interpretation. While the geometric quality of images used for the visual interpretation is well understood, little is known about the interpreters' preferred image characteristics such as brightness and contrast, image format, colour composite or about interpreters' assessments of images with shadowed areas. Cast shadows partially covering objects or phenomena can hinder effective interpretation. Since visual interpretation is subjective and is likely to produce inconsistencies between interpreters, one way to assure reliability is interpretation of sites by several interpreters. Still there are inconsistencies in visual image interpretations. Little is known about factors influencing consistency of interpretations done by several individuals.

This thesis aims to identify factors affecting visual image interpretation for acquiring reference data for agricultural land monitoring. The thesis consists of

six chapters. In chapter 1, I introduce the topic and formulate the research questions.

In chapter 2, I report on image characteristics that are preferred to facilitate visual interpretation. For this purpose, I designed and executed a survey among interpreters of agricultural land. Survey questions tested appreciation of image brightness and contrast ranges and also the influence of a background colour on the preferred image brightness and contrast, the preferred image format, colour composite, assessments of images with shadowed areas, appreciation of image enhancements and, finally, consistency of individuals' preferred brightness and contrast settings across multiple sample images. I found that appreciation of image photometric settings is stable at the individual level, but preferences vary across respondents. It was therefore recommend that interpreters are enabled to personalise image settings, such as brightness and contrast values, and to choose displayed band combination.

In chapter 3, I compared visual interpretation and a proposed automated detection method for identifying cast shadows. Cast shadows were identified by multiple interpreters and as well as by an automated method. The automated method determines a threshold from automatically generated training data using a risk-based approach. In general, more area was labelled as cast shadow by the automatic procedure than by visual interpretation. However, the cast shadow area labelled by the automated method was in between the intersection and the union of the areas labelled by interpreters. The cast shadow areas identified by the individual interpreters had limited intersection demonstrating that the interpreters strongly disagreed. Furthermore, the automated shadow detection method is reproducible and reduces interpretation effort and required skills. It seems promising for the screening of large image scenes and datasets and as a first assessment of image usability for agriculture monitoring programs.

In chapter 4, I investigated how the consistency of visual image interpretations of land cover can be improved. I evaluated relationships between the number of interpretation updates following feedback on submitted interpretation. It was found that feedback loops reduced the amount of required updates and hence improved consistency of the interpretations. Implementing feedback loops into the visually interpreted data collection process increases the consistency of acquired land cover reference data. Therefore review and feedback should be planned and customised for each interpreter to optimise the quality of visual image interpretation.

Chapter 5 evaluates whether the number of available images and land change process type influence agreement between visual image interpretations of land

changes based on results from an experiment conducted with 18 interpreters. I found that land change process type had a significant impact on interpretation consistency, while number of images had no influence. Absence of change was most consistently identified. On the contrary, agricultural land abandonment and reforestation were the least agreed upon. The results, further highlight subjectivity in visual interpretation when land undergo changes. Efforts allocated to acquire reference data through means of visual interpretation can be adjusted based on the expected type of change process and land cover/land use type. For example, a dedicated training addressing agricultural land abandonment and reforestation can be provided to the interpreters.

The final chapter presents and discusses the main conclusions of my work. Interpretation consistency is affected by individual aspects of interpreters, by analysed land cover, land use and land change type. Furthermore, consistency of visual interpretation is influenced by the adopted method of enhancing interpretation agreement, such as reviews and feedbacks.

Future challenges lay in identifying ways and extent to replace visual image interpretation with automated image classification in the context of reference data acquisition. Future research should also focus on the role of visual image interpretation for agricultural land monitoring. Presence of subjectivity in reference dataset cannot be ignored. Since there is no way to fully reduce inconsistency in visual image interpretations, further studies on inconsistency assessment are needed.

Acknowledgements

This thesis is a result of a long and challenging journey which started back in 2012. It could never be successful if not for my promotor Arnold K. Bregt and my co-promotor and supervisor Sytze de Bruin. I would like to thank them for providing excellent inputs during different research phases, for their countless advises, help and patience throughout all the years of my PhD.

I would like to thank my co-promotor Nandika Tsendbazar for her precious help during the latest stage of my PhD, for giving me the opportunity to be part of the Global Land Cover project and to actually make my moving to Wageningen happen.

I would like to thank Wim Devos, my supervisor in JRC, Ispra, Itlay, whose expertise in and broad knowledge of LPIS and its Quality Assessment were great support during the first years of my PhD.

I would like to thank my opponents for reading and constructively commenting my thesis and for taking the effort to travel to Wageningen for my defence.

I would like to thank my friends, colleagues and co-workers from JRC in Italy, who actively helped me with the experiments presented in second chapter of this book. I would like to thank the Quality Control team from the project in Ankara, Turkey, whose contribution is reflected in third chapter. I would like to thank students from Remote Sensing and GIS Integration 2019 course who participated in the survey mentioned in fifth chapter. My thanks go to all friends and colleagues from Wageningen University and Research with whom I have pleasantly spent the last years of the PhD. Many thanks to my paranymphs Sabina Roşca and Johannes Balling for their great help and support during the last stretch of the PhD.

Special thanks to my family, the loved ones and friends, whose support through all those years meant a lot to me.

About the author

Agnieszka was born and grew up in South of Poland. She obtained a Master degree in Geography from Jagiellonian University in Cracow (2009) and a degree of Master of Science in Engineering in the field of Geodesy and Cartography from AGH University of Science and Technology (2012).

In 2011 she moved to Italy for an internship as a satellite and aerial imagery interpreter in Joint Research Centre of European Commission, where she continued as a grant holder till 2015. The grant was aiming on image assessment for Land Parcel Identification System applications and was set in collaboration with Wageningen University & Research. Together with the grant, Agnieszka started the external PhD programme in Wageningen University & Research in 2012.

After her contract in Italy was over, Agnieszka worked in Ankara as a consultant in the Europe Aid project "External Quality Control under digitalisation of Land Parcel Identification System" and in her free time she continued to work on her PhD.

When the Turkish experience was over, she got an opportunity to work at the Wageningen University & Research. She moved to The Netherlands at the beginning of 2018. In 2019 she successfully completed her PhD thesis.

Journal publications

Tarko, A., de Bruin, S., Bregt, A.K., 2018. Comparison of manual and automated shadow detection on satellite imagery for agricultural land delineation. Int. J. Appl. Earth Obs. Geoinf. 73, 493–502. https://doi.org/10.1016/j.jag.2018.07.020

Tarko, A., de Bruin, S., Fasbender, D., Devos, W., Bregt, A., 2015. Users' Assessment of Orthoimage Photometric Quality for Visual Interpretation of Agricultural Fields. Remote Sens. 7, 4919–4936. https://doi.org/10.3390/rs70404919

Project reports

Devos, W., Tarko, A., Milenov, P., Franielczyk, R., 2013. Findings from the 2010 and 2011 LPIS QA records produced by the MS in the Frame of Comm. Reg. No 1122/2009 art. 6.2. Publications Office of the European Union, Luxembourg. ISBN 978-92-79-29473-0

Devos, W., Milenov, P., Wojda, P., Tarko, A., Franielczyk R., 2012. The first year of implementation of the LPIS quality assessment in the frame of Comm. Reg. No 1122/2009 art. 6.2. Publications Office of the European Union, Luxembourg. ISBN 978-92-79-23067-7

Milenov, P., Devos, W., Wojda, P., Tarko, A., 2012. LPIS Quality Assurance Framework Annex I - Executable Test Suite (ETS). Publications Office of the European Union, Luxembourg. ISBN 978-92-79-22803-2

Milenov, P., Devos, W., Wojda, P., Tarko, A., 2011. LPIS Quality Assurance Framework Annex II - Executable Test Suite (ETS). Publications Office of the European, Union, Luxembourg. JRC 68017

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (4.5 ECTS)

- International Archives of Photogrammetry and Remote Sensing, International Journal of Geographical Information Systems, International Journal of Remote Sensing, ISPRS Journal of Photogrammetry and Remote Sensing, Photogrammetric Engineering and Remote Sensing, Remote Sensing of Environment, Remote Sensing Reviews, Remote Sensing Series, The Photogrammetric Record, ISO Standards, CwRS Specifications, LPIS Specifications, WikiCAP website, Orthoimagery guidelines, OGC Standards

Writing of project proposal (4.5 ECTS)

- Image spectral/radiometry assessment for Land Parcel Identification System (LPIS) applications (2012)

Post-graduate courses (11.6 ECTS)

- Introduction to geoprocessing using Python; JRC Ispra (2013)
- PhD course: geostatistics; PE&RC (2013)
- Course radiometric performance of digital photogrammetric cameras and airborne laser scanners; EuroSDR (2013)
- Course mapping using high-resolution satellite imagery; EuroSDR (2014)
- Course change detection in high-resolution land use/cover geodatabases (at object level); EuroSDR (2014)

Deficiency, refresh, brush-up courses (0.6 ECTS)

- Exploring Envi; JRC Ispra (2012)

Competence strengthening / skills courses (2.4 ECTS)

- Scientific writing; JRC Ispra (2012)
- Active reading; JRC Ispra (2012)
- Public speaking; JRC Ispra (2013)
- Adobe photoshop Cs5; JRC Ispra (2013)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.2 ECTS)

- PE&RC Mid-term weekend (2014)
- PE&RC Last years weekend (2019)

Discussion groups / local seminars / other scientific meetings (4.5 ECTS)

- MARS Unit scientific seminars; JRC Ispra (2012-2015)

International symposia, workshops and conferences (4.2 ECTS)

- LPIS QA Training; oral presentation; Tallinn, Estonia (2011)
- LPIS Workshop; oral presentation; Baveno, Italy (2013)
- Inspector training; oral presentation; Baveno, Italy (2013)

Supervision of MSc students

- Comparison of radiometric image characteristics for the measurement and identification of selected Landscape Features relevant for the Common Agriculture Policy (CAP) (2014)

This research was financed by European Commission, Joint Research Centre, Ispra, Italy through Grantholder contract Category 20 in years 2012-2015 and received funding from Chair group of Laboratory of Geo-information Science and Remote Sensing, Wageningen University for the last seven months stretch of the PhD.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover images: Wageningen University & Research campus and surroundings (Google Earth historical images, Sentinel-2/Copernicus images, OpenStreetMap)

Printed by ProefschriftMaken, www.proefschriftmaken.nl