International Journal of Applied Earth Observation and Geoinformation 54 (2017) 124–133

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Combining machine learning and ontological data handling for multi-source classification of nature conservation areas

Niklas Moran[a,*], Simon Nieland[a], Gregor Tintrup gen. Suntrup[b], Birgit Kleinschmit[a]

[a] Geoinformation in Environmental Planning Lab, Technische Universität Berlin, Straße des 17. Juni 145, 10623 Berlin, Germany
[b] Dept. Environmental Systems, RLP AgroScience – Institute for Agroecology, Breitenweg 71, 67435 Neustadt, Germany

ARTICLE INFO

Article history:
Received 20 July 2016
Received in revised form
11 September 2016
Accepted 20 September 2016

Keywords:
Remote sensing
Ontology
Biotope classification
Machine learning
Nature conservation
OWL
EUNIS
GEOBIA
Grasslands

ABSTRACT

Manual field surveys for nature conservation management are expensive and time-consuming and could be supplemented and streamlined by using Remote Sensing (RS). RS is critical to meet requirements of existing laws such as the EU Habitats Directive (HabDir) and more importantly to meet future challenges. The full potential of RS has yet to be harnessed as different nomenclatures and procedures hinder interoperability, comparison and provenance. Therefore, automated tools are needed to use RS data to produce comparable, empirical data outputs that lend themselves to data discovery and provenance. These issues are addressed by a novel, semi-automatic ontology-based classification method that uses machine learning algorithms and Web Ontology Language (OWL) ontologies that yields traceable, interoperable and observation-based classification outputs. The method was tested on European Union Nature Information System (EUNIS) grasslands in Rheinland-Palatinate, Germany. The developed methodology is a first step in developing observation-based ontologies in the field of nature conservation. The tests show promising results for the determination of the grassland indicators wetness and alkalinity with an overall accuracy of 85% for alkalinity and 76% for wetness.

1. Introduction

Recognizing the importance of functioning ecosystems to reduce biodiversity loss, the European Union has implemented an environmental conservation framework to protect and conserve vital habitats in accordance with the Convention on Biological Diversity. An integral part of this framework is the EU Habitats Directive (Council Directive) 92/43/EEC [1992], which established the Natura 2000 network of habitats. The directive requires member states to conserve and monitor designated habitats and submit a report every six years. Environmental data to determine biodiversity status must be collected to comply with reporting requirements. Comparing data used for these reports is difficult because of varying data collection methods and acquisition nomenclatures used by nature conservation authorities in each member state (Vanden Borre et al., 2011). The main issue is in the subjective nature of field surveys to identify habitats (Cherrill and McClean, 1999b,a; Hearn et al., 2011; Nieland et al., 2015a). Furthermore, habitat status is mainly generated in bottom-up approaches taking into account the national and regional interpretation guidelines (Vanden Borre et al., 2011). This subjective and time-consuming task of conducting field surveys could be partially replaced with an automated RS method that uses Geographic Object Based Image Analysis (GEOBIA) to reduce subjectivity, costs and time.

RS offers opportunities to collect and automatically interpret large amounts of computer-readable data useful for nature conservation and biodiversity monitoring (Corbane et al., 2015; Vanden Borre et al., 2011; Mayer and Lopez, 2011). RS image analysis implicitly incorporates the expertise of the person performing the analysis, reducing reproducibility as the analyst ultimately chooses class membership in non-crisp boundaries between classes. This can be divided into RS knowledge (spectral signature, remote sensing indices, etc.) and field knowledge (feature properties, spatial relations, etc.) (Andrés et al., 2013), which is often neither completely nor explicitly defined as it is based on trial and error but influences the classification (Arvor et al., 2013). To ensure accuracy and applicability of classification outputs for conservation, experts with detailed knowledge of the sites are needed to interpret the RS data. The distance between the high-level semantics used by experts to describe domain concepts and the low-level information quantified from data is referred to as the "semantic gap" (Smeulders et al., 2000).

* Corresponding author.
E-mail address: niklas@niklasmoran.com (N. Moran).

http://dx.doi.org/10.1016/j.jag.2016.09.009
0303-2434/

Ontologies can help bridge the “semantic gap” and allow for better data transferability, knowledge and workflow management (provenance) and logical consistency (Janowicz, 2012). The standards-compliant format designed and adopted to express rich semantics and enable the “Semantic Web” is called the OWL.[1] The format supports multiple syntaxes yet defines the Resource Description Framework (RDF) (subject, predicate, object triplets) saved as eXtensible Markup Language (XML) as a common exchange format. Moreover, through the use of reasoners (inference engines) that infer logical consequences over axioms and asserted facts and verify consistency, one can discover new knowledge (Arvor et al., 2013; Andrés et al., 2013). RS and field expert knowledge can be digitized in ontologies, thus allowing for a hierarchy of concepts for improved automatic image annotation and retrieval using concepts from both fields to produce more accurate results (Srikanth et al., 2005). Janowicz (2012) advocates for more observation-driven ontologies and for including machine learning and statistics to construct ontological primitives. While published research on using observation-based ontologies for biotope classifications is limited; the available research using ontologies in RS research is briefly summarized below.

Ontologies modeled on the Land Cover Classification System and the General Habitat Category were integrated into tools used to monitor and protect areas in the EU (Arvor et al., 2013). The authors note that using the taxonomy of the different classification systems makes it possible to include expert knowledge in the process. Lucas et al. (2015) used pixel-based analysis and GEOBIA for greater classification accuracy which relies on a rule-base created by an expert. Other research includes urban building classification using a three-layered architecture (di Sciascio et al., 2013) and another using semi-automated classification using the Random Forest classifier to determine variable importance of features from airborne laser scanner data (Belgiu et al., 2014). Ontologies have also been paired with different algorithms to automatically acquire classification rules: a genetic programming algorithm (Forestier et al., 2012) and the C4.5 machine learning algorithm (Sheeren et al., 2006). In biodiversity monitoring research, ontologies have been demonstrated to improve spatial data interoperability (Nieland et al., 2015a, 2015) and have been shown to aid in discovery of new relationships to consider for habitat management (Pérez-Luque et al., 2015). The addition of fuzzy data types to OWL and the development of a fuzzy spatial reasoner holds great promise for the future of GEOBIA ontology research using remote sensing (Mariana Belgiu and Hofer, 2013; Bobillo and Straccia, 2015). More recently a multi-scale fuzzy spatial reasoner was developed that could have significant impact on this research (Argyridis and Argialas, 2015).

Even though researchers recently developed a number of indicators using different sensors for habitat evaluation (Nagendra et al., 2013), classification procedures and rule-sets were not formalized to be computer-readable and therefore suffer from similar transferability and reproducibility problems as manual habitat mapping (Arvor et al., 2013; Nieland et al., 2015a, 2015). Therefore, a formalized computer-readable ontology could help solve these problems and allow scientists to see how the classification was performed and be aware of possible incompatibilities before combining data (Janowicz, 2012). Furthermore, there are no standardized transnational habitat evaluation RS indicators (Lucas et al., 2015; Vanden Borre et al., 2011). Therefore, technical solutions to increase interoperability by thematically harmonizing environmental data and systematize data collection methods from remote sensing inputs in an automated workflow are needed.

To help increase interoperability, an expert group that seeks to harmonize land cover (LC) and land use (LU) nomenclatures using an object-oriented data model that eases translations between nomenclatures was formed and is called EIONET Action Group on Land monitoring in Europe (EAGLE) (Arnold et al., 2013). The many different nomenclatures used in Europe each have their own specific thematic conceptualization suited towards a specific scale and data collection method – which reduces the ability tocompare thematic maps. Since LC and LU are interconnected and influence one another, nomenclatures often incorporate both definitions into one, class making separation difficult. To overcome this problem, the EAGLE data model describes landscapes in three main components: land cover (abiotic, vegetation, water) land use (agriculture, forestry, etc.) and characteristics (bio-physical, cultivation, etc.). The increased interoperability and transferability of RS data and the semantic layer on top helps decision-makers to better assess and compare outcomes.

In this paper we propose an automated system that can classify dry, mesic and wet grasslands according to the EUNIS biotope classification schema using earth observation data, existing thematic maps (biotope, forestry, etc.), and expert knowledge formalized in an ontology with rules generated by machine learning algorithms. This method contributes to the goal of empirically derived rule creation and enhances data interoperability and comparison as proposed by Janowicz (2012). While this paper focuses on nature conservation, the method is domain independent and could be used with other vocabularies that have more generalized terms such as EAGLE. The main goals of this paper are:

- to develop a RS classification methodology using a Decision Tree Classifier (DT) approach in combination with ontological formalism to generate highly interoperable, reproducible and exchangeable classification procedures and results,
- apply the methodology to indicators used to separate grassland habitats defined under EUNIS,
- and evaluate the developed approach by comparing it to an ensemble classification algorithm (Extra Tree Classifier (ET)).[2]

2. Method

This section proposes an ontology-based classification approach which uses a DT for the semantic annotation procedure. It furthermore describes the dataset used in the study, the nomenclatures used and the semantic conceptualization of classes in the nature conservation domain. To evaluate the method, the results were compared to a highly randomized tree classifier called ET (Geurts et al., 2006), which is a tree-based ensemble classifier known to be suitable for this kind of classification problem (Qian et al., 2015; Franke et al., 2012; Hladik et al., 2013; Zou et al., 2010; Barrett et al., 2014).

2.1. Study data bases and Nomenclatures

To evaluate the developed method, this work shows how it could be used to support the federal state of Rhineland Palatinate (RLP) in performing their regional biotope mapping and fulfilling Hab-Dir reporting obligations. Therefore we chose the nomenclature of the EUNIS to classify biotopes as it directly satisfies the HABDIR and has already been semantically transferred to the local mapping nomenclatures. A consistent classification process could be realized by describing EUNIS classes with biophysical and anthropogenic indicators that are supposed to be derivable with the help

[1] http://www.w3.org/TR/owl2-overview/

[2] http:// scikit- learn. org/ stable/ modules/ ensemble. html#extremely-randomized- trees