Environmental Modelling & Software 87 (2017) 17–28

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Use of freely available datasets and machine learning methods in predicting deforestation

Helen Mayfield [a, *], Carl Smith [b], Marcus Gallagher [c], Marc Hockings [a]

[a] School of Geography, Planning and Environmental Management, University of Queensland, St Lucia, QLD 4072, Australia
[b] School of Agriculture and Food Science, University of Queensland, St Lucia, QLD 4072, Australia
[c] School of Information Technology and Electrical Engineering, University of Queensland, St Lucia, QLD 4072, Australia

ARTICLE INFO

Article history:
Received 17 March 2016
Received in revised form
4 October 2016
Accepted 21 October 2016

Keywords:
Artificial neural network
Bayesian network
Deforestation
Freely available data
Gaussian process
Logistic regression

ABSTRACT

The range and quality of freely available geo-referenced datasets is increasing. We evaluate the usefulness of free datasets for deforestation prediction by comparing generalised linear models and generalised linear mixed models (GLMMs) with a variety of machine learning models (Bayesian networks, artificial neural networks and Gaussian processes) across two study regions. Freely available datasets were able to generate plausible risk maps of deforestation using all techniques for study zones in both Mexico and Madagascar. Artificial neural networks outperformed GLMMs in the Madagascan (average AUC 0.83 vs 0.80), but not the Mexican study zone (average AUC 0.81 vs 0.89). In Mexico and Madagascar, Gaussian processes (average AUC 0.89, 0.85) and structured Bayesian networks (average AUC 0.88, 0.82) performed at least as well as GLMMs (average AUC 0.89, 0.80). Bayesian networks produced more stable results across different sampling methods. Gaussian processes performed well (average AUC 0.85) with fewer predictor variables.

1. Introduction

Forests around the world remain at risk from a range of threats including urban population growth (DeFries et al., 2010), agricultural and infrastructure expansion (Newman et al., 2014), illegal logging (Gaveau et al., 2009) and insecure property rights (Robinson et al., 2014). With the loss of the forests, we are also losing valuable ecosystem services (Rogers et al., 2010), critical habitats for maintaining biodiversity (Buchanan et al., 2008) and destroying an important carbon sink that could help mitigate increasing atmospheric concentrations of carbon dioxide (Wang et al., 2009). In order to better understand and ultimately reduce these risks, researchers frequently turn to data driven analyses (Mas et al., 2004; Vaca et al., 2012; Allnutt et al., 2013; Newman et al., 2014) for which access to relevant and quality information is crucial.

Despite their value, many datasets, especially at high resolution, still remain difficult or costly to obtain. Socio-economic data may rely on costly surveys and gaining access to data on dynamic variables, such as city or road locations, for the relevant time periods (i.e. when the deforestation was occurring) can be difficult and may require manual digitisation of maps. In contrast, other geo-referenced datasets, such as those describing land use change (Vaca et al., 2012; Allnutt et al., 2013), protected areas (WDPA, 2010), political boundaries (NE, 2013a) and ecoregions (Olson et al., 2001) are becoming freely available. To date however, there has been no rigorous assessment of the utility of using these freely available datasets for deforestation risk modelling.

To analyse these data, researchers have often relied on classical statistics such as generalised linear models – GLMs (Hastie et al., 2009), and more recently generalised linear mixed models – GLMMs (Green et al., 2013). While these techniques are well accepted and easily implemented, they assume explanatory variables are independent (unless dependencies are explicitly

List of Abbreviations: ANN, Artificial neural networks; BN, Bayesian networks; CI, Conservation International; DEM, Digital elevation model; FN, False negative; FP, False positive; GLM, Generalised linear model; GLMM, Generalised linear mixed model; GP, Gaussian process; IUCN, International Union for the Conservation of Nature; ML, Machine learning; NE, Natural Earth; PA, Protected area; TAN, Tree Augmented Naïve; TN, True negative; TP, True positive; TSS, True skill statistic; AUC, area under the (receiver operating) curve; WDPA, World database on protected areas; WWF, World Wildlife Fund.

* Corresponding author.
E-mail addresses: helenmayfield@warpmail.net (H. Mayfield), c.smith2@uq.edu.au (C. Smith), marcusg@itee.uq.edu.au (M. Gallagher), m.hockings@uq.edu.au (M. Hockings).

http://dx.doi.org/10.1016/j.envsoft.2016.10.006
1364-8152/

Software and data availability

Software

Name Netica version 5.12
Developer Norsys Software Corporation
Address 3513 West 23rd Avenue, Vancouver, BC, Canada, V6S1k5
Email info@norsys.com
Availability www.norsys.com
Name ArcGIS 10.1
Developer ESRI
Address 380 New York Street, Redlands, CA 92373-8100
Email service@esri.com
Availability http://www.esri.com
Name Fragstats
Developer UMass Landscape Ecology Lab
Address 304 Holdsworth Natural Resources Center, Box 34210, Amherst, MA 01003
Email mcgarigalk@eco.umass.edu
Availability http://www.umass.edu/landeco/research/fragstats/fragstats.html
Name R Programming Language
Developer R Core Development Team
Availability http://www.r-project.org
Name MatLab 2014
Developer MathWorks
Address 1 Apple Hill Drive, Natick, MA 01760-2098, United States
Availability www.mathworks.com

Datasets

Name Land use change
Developer Conservation International
Availability Available on request. http://www.conservation.org
Name Terrestrial Ecoregions
Developer World Wildlife Fund
Availability http://worldwildlife.org/publications/terrestrial-ecoregions-of-the-world
Name VMAP0
Developer mapAbility
Availability http://www.mapability.com
Name World Database of Protected Areas
Developer United Nations World Conservation Monitoring Centre
Availability http://www.protectedplanet.net
Name Natural Earth large scale datasets
Developer Natural Earth
Availability http://www.naturalearthdata.com
Name Landscan global population distribution
Developer Oak Ridge National Laboratory
Availability http://web.ornl.gov/sci/landsca
Name U.S. Geological Survey's Landat data
Developer U.S. Geological Survey's Earth Resources Observation and Science
Availability http://landsat.usgs.gov/Landsat_Search_and_Download.php

modelled) and cannot exploit nonlinear relationships between dependent and independent variables (unless they are known to be nonlinear a priori and data can be transformed). Machine learning (ML) methods such as artificial neural networks - ANNs (Hastie et al., 2009), Bayesian networks – BNs (Fenton and Neil, 2013) and Gaussian processes – GPs (Rasmussen and Williams, 2006), do not make these assumptions. This may prove to be advantageous when it comes to modelling deforestation risk where predictor variables may not be independent or relationships linear.

While comparisons of multiple ML and statistical methods have been conducted in assessing landslide susceptibility (Pham et al., 2016), land use change (Tayyebi et al., 2014), and conservation biology (Kampichler et al., 2010), such broad comparisons have not been undertaken for deforestation risk assessment, with studies either offering no model comparison (Mas et al., 2004; Basse et al., 2014) or a limited comparison of only two methods (Pérez-Vega et al., 2012). This study aims to address this gap while at the same time evaluating a variety of relevant, freely available or low cost datasets to determine their usefulness in predicting deforestation risk, defined here as probability of the presence or absence of deforestation.

By using several statistical and machine learning techniques, we assess whether machine learning is able to improve on the more commonly used methods from classical statistics. In doing so we provide researchers with guidance on the comparative performance of these analytical methods in predicting deforestation risk. We first describe the datasets used in this study along with each deforestation risk modelling method compared. We then describe the design and implementation of each modelling method, the predictor variables included and the model evaluation metrics used in this study. Finally, we examine how the ML models compared against standard statistical models and the implications of these results.

1.1. Freely available datasets

Free or low cost datasets are becoming increasingly common and cover a range of factors relevant to analysing deforestation. While efforts are being made to look at methods for improving the quality of land use images in these datasets (Estes et al., 2016), many are already at a standard that is potentially useful for practical deforestation prediction. High levels of correlation amongst variables are common in land use change, with multiple factors sometimes resulting in the same result (van Vliet et al., 2016), and deforestation is no exception to this. While these correlations can create complications with model design and validation (van Vliet et al., 2016), it also suggests that the large range of available datasets (detailed further in this section) may provide an alternative source of variables in cases where more expensive or difficult to collect options are not available.

One major development in geo-referenced datasets is the World Database on Protected Areas (WDPA), which is maintained by the United Nations Environmental Program World Conservation Monitoring Centre (UNEP-WCMC). The positive influence of protected areas (PAs) on preventing deforestation within their boundaries has been shown (Mas, 2005; Gaveau et al., 2009), although there is some debate in the literature regarding the magnitude of this influence, with some evidence that the credit afforded to protected areas is due not to the protected status of the forest, but to other attributes, such as accessibility (Gaveau et al., 2009). The database is a global, geo-referenced dataset that details the location (as a polygon layer) and date of declaration for the world's PAs (WDPA, 2010). It also lists details such as the conservation category (if any) for each PA, as described by the International Union for the Conservation of Nature – IUCN (Dudley, 2008).

The Landsat Thematic Mapper and Enhanced Thematic Mapper