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

Environmental Modelling and Software:

Accepted October 2016

Authors: Helen Mayfield[a], Carl Smith[b], Marcus Gallagher[c] , Marc Hockings[d]

a: helenmayfield@warpmail.net
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

c:
School of Information Technology and Electrical Engineering
University of Queensland

d: m.hockings@uq.edu.au
School of Geography, Planning and Environmental Management
University of Queensland

Corresponding Author:

Helen Mayfield
Phone number: + 44 7 946 924 158
Email: helenmayfield@warpmail.net
Present Address:
Ground Floor
87 Camborne Ave.
London, W13 9QZ
United Kingdom

Highlights

- Freely available datasets have proven valuable in predicating deforestation
- Machine learning techniques are a reliable alternative to statistics
- Gaussian processes are suggested as an alternative to artificial neural networks
- Bayesian networks were more stable across sample methods

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.

Keywords

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

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: https://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

List of Abreviations

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

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 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 images provide global data with a spatial resolution of 30 m x 30 m (Wang et al. 2009), with the most recent satellite, Landsat 8, being launched in 2013 (NASA 2015). Data from the Landsat satellite program (UT-Battelle 2013) are frequently used in deforestation studies for calculating slope and elevation variables (Mas et al. 2004, Gaveau et al. 2009, Wang et al. 2009). Satellite data is also available from the National Aeronautics and Space Administration's (NASA's) Geocover project which has been used by Conservation International (CI) to create land use change datasets for many deforestation hotspots. The dataset for Mexico is in raster format (28.5 m resolution) and maps forest lost between 1990 and 2000, and between 2000 and 2005, where forest is defined as old growth forest, secondary and degraded forests, and plantations (Vaca et al. 2012). An equivalent dataset covering Madagascar exists from the same source (Allnutt et al. 2013).

Other non-profit organisations also make data available without charge for scientific or other non-commercial purposes. Natural Earth (NE) has published a large number of datasets with global

coverage, including political boundaries and locations of populated places, ports and airports (NE 2013b). These datasets give access to a number of variables that have been linked to deforestation risk, such as distance to populations of different sizes (Mas et al. 2004) and political boundaries, such as states and countries that may have differing forest protection policies. Similarly, the World Wildlife Fund (WWF) has produced a global map of terrestrial ecoregions (Olson et al. 2001), which has been used to identify and control for differences in deforestation rates between the different ecoregions (Vaca et al. 2012). A global dataset of major roads is also available (mapAbility 2012). While a useful reference, this last dataset should be used with caution as the dates for when the roads were created are not given, meaning that it cannot be verified if the roads were in existence at the time when deforestation occurred.

As well as free information it is possible to purchase data. As an example, The Oak Ridge National Laboratory offers the Landscan population pressure raster dataset at 1 km (UT-Battelle 2013). While the cost may prove prohibitive for some studies, the data is considered high quality and the algorithms used to calculate population pressure make use of roads, populated areas (urban boundaries) and populated points (towns and villages). It has been used previously to estimate population pressure when analysing deforestation (Rogers et al. 2010). As with the road location data, values in the population pressure dataset represent the state of the world in a recent time period, rather than when deforestation was occurring. This may affect the relevance of the information, particularly if there have been significant population movements over the past few decades.

The datasets described have several advantages that make them widely applicable to deforestation studies. All provide extensive, in some cases global, coverage of deforestation hotspots while still having sufficient resolution for more local studies. Most also offer enough information to derive a selection of potentially useful predictor variables for estimating deforestation risk. The variety of the datasets provides an extensive range of potential predictor variables, including many of the most commonly studied predictors such as slope, elevation, population pressure and surrounding land use.

1.2 Modelling methods

GLMs (Hastie et al. 2009) are a family of statistical techniques commonly applied in deforestation prediction. In a GLM the output is modelled as a linear combination of the inputs, sometimes passed through a nonlinear function (e.g. a sigmoidal function for logistic regression). GLMMs are an extension of GLMs that can model random effects among groups of predictors and are becoming widely used in environmental and conservation analysis studies (Green et al. 2013,

Newman et al. 2014). Both GLMs and GLMMs are unable to account for interactions between predictors unless these are explicitly modelled and pre-specified.

ANNs are predictive models loosely based on the biological structure of the human brain (Rumelhart et al. 1986, or more recently Haykin 2009). An ANN is constructed by linking input nodes, with weighted connections, to output nodes via one or more layers of hidden nodes. Without these hidden layers an ANN is equivalent to either linear or logistic regression, depending on the output function used. During training, data are presented to the network via the input layer. These values are then processed by the first layer of hidden nodes, where they are multiplied by the weights for each node and processed according to a sigmoidal activation function. The output from each layer of hidden nodes is used as the input to the next layer of hidden nodes (Haykin 2009). The output from the final hidden layer is passed to the output nodes.

A BN is a graphical model that takes a probabilistic approach to representing relationships among variables (Fenton and Neil 2013). At the core of this approach is Bayes theorem, which uses conditional probabilities to estimate the probability of a hypothesis given the evidence. Key benefits of BNs are their ability to deal with uncertain or missing data and a clear, graphical representation of the relationships between variables (Uusitalo 2007). Another advantage of BNs is that both the network structure (cause and effect relationships among variables) and conditional probabilities can be either learnt from data or derived from expert knowledge.

GPs are a spatial model that can be viewed as a particular instance of the well-established geostatistics technique of kriging (Hastie et al. 2009) and allow for a very flexible range of response functions to be modelled. The model is defined by a mean and covariance function, which defines a prior distribution over the possible functions. Given a training set, this can be converted into a posterior distribution using Bayes Rule. The posterior distribution is then used to make predictions e.g. by taking the mean of the posterior distribution as the predicted value of the response function at a test point. A detailed explanation of the equations used in GPs is given in Section 1 of the online supplementary material.

ANNs have shown promising results when applied to deforestation risk modelling (Mas et al. 2004) and more generally to modelling changes in land use (Basse et al. 2014). BNs have been successfully applied to numerous environmental management studies including reforestation (Frayer et al. 2014) and forest dynamics (Liedloff and Smith 2010). While no research was found that specifically used GPs in deforestation risk assessment, the method is similar to Kriging (Rasmussen and Williams 2006) making it a suitable approach for modelling spatial patterns (Campos-Taberner et al. 2015, Yan et al. 2016).

2 Materials and methods

Two areas were selected for this study, one in Mexico (Figure 1) and the other in Madagascar (Figure 2). Mexico, is widely recognised as one of a handful of the remaining mega-diverse countries for biodiversity and has historically had high deforestation rates (Mas 2005). Changes in government policies and investment in infrastructure in the late 20th century played a major role in increasing forest loss (Ellis and Porter-Bolland 2008).

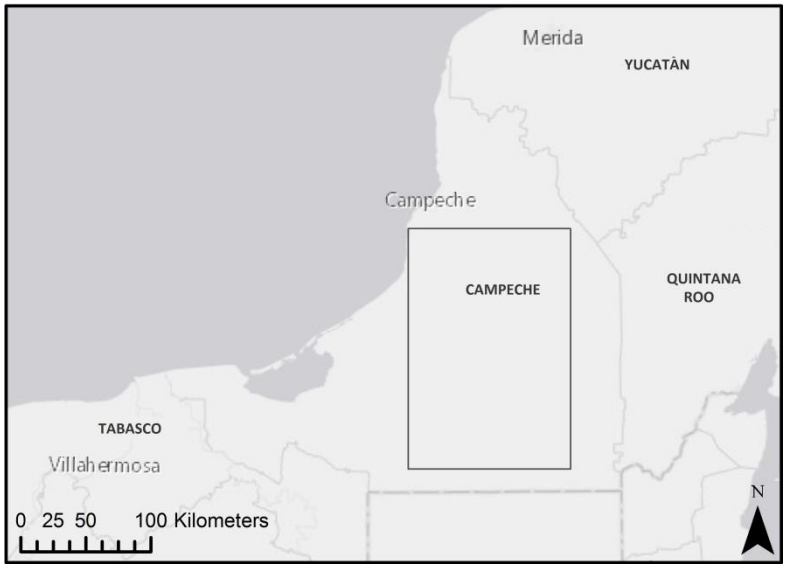

Figure 1: Location of study area (outlined in black) within the Mexican Yucatan peninsula.

Figure 2: Location of study area (outlined in black) within Madagascar

Within the Yucatan region there are variations in the types of agriculture, access to alternative incomes (such as tourism) and forest management policies. Therefore the underlying factors affecting deforestation (such as population pressure or land tenure) differ in influence from area to area (Ellis and Porter-Bolland 2008). There are several major protected areas in the region, including the Calakmul Biosphere Reserve, El Mirador National Park and Tikal National Park. For this study a 100 km x 200 km area was selected from the state of Campeche, overlapping the Calakmul Biosphere Reserve (WDPA 2010).

Like Mexico, Madagascar has high levels of deforestation and is a priority for conservation due to its substantial number of endemic species (Allnutt et al. 2013). There are several prominent causes of deforestation in the country including slash and burn farming carried out primarily for rain fed hill rice cultivation (McConnell et al. 2004), both legal and illegal logging of hardwoods, and mining (Allnutt et al. 2013). Exact deforestation rates are not known as estimates are generally considered inaccurate due to missing data caused by cloud cover, and also vary depending on the definition of forest (Agarwal et al. 2005). For this study, the north eastern province of Toamasina was chosen. The province has an area of approximately 75,000 km^2 and consists of two main ecoregions; Madagascan lowland forest along the east and a section of Madagascan sub humid

forests in the west (Olson et al. 2001). Land use change maps for both study regions are given in Section 2 of the online supplementary material.

2.1 Data and Variable selection

A selection of freely available or low cost datasets were used to model deforestation risk. The datasets used in this study were the WDPA (WDPA 2010), CI land use change data (Vaca et al. 2012), NE political boundaries and city locations (NE 2013b), WWF terrestrial ecoregion (Olson and Dinerstein 2002) the digital elevation model (DEM) from Landsat (Reuter et al. 2007), MapAbility road location (mapAbility 2012) and the Landscan population pressure (UT-Battelle 2013).

To create data samples for training and testing the models, random sample points were generated across the study areas and then overlaid on to the CI land use change datasets. A change in land cover from forest to non-forest across a time step represents deforestation. For this study, the deforestation response variable was defined as either 0 if deforestation was absent between 2000 and 2005 or 1 if deforestation was present during this timeframe. The value of the response variable for each sample point was initially defined as whether or not the underlying 30 m x 30 m cell in the CI dataset was deforested between 2000 and 2005. For the Mexican dataset, this resulted in a prevalence rate for deforested sample points of less than 1%. Data containing low prevalence rates presents an issue for both statistical and ML models because extremely large datasets are then required for models to have sufficient examples of deforestation to learn from. Two strategies were therefore used to deal with the low prevalence of deforested sample points within the sample.

The first strategy was to redefine the conditions that needed to be met for a sample point to be considered deforested. Instead of determining deforestation according to the land cover change only within the 30 m x 30 m cell underlying a sample point, a target region of 500 m x 500 m surrounding this point was considered. If any deforestation had occurred within this target region, then the sample point was classified as deforested (deforestation present). If not then the point was classified as not deforested (deforestation absent). This increased the prevalence rate of deforested sample points in the Mexican and Madagascan datasets to 4.8% and 18% respectively, providing more examples of deforestation for the models to learn from.

The second strategy was to create a second dataset for each study area using stratified random sampling (Haibo and Garcia 2009). For this sampling technique, half of the samples were randomly

selected from points where deforestation was present within the target region and half randomly selected from points where deforestation was absent within the target region. This created a sample with a 50% prevalence rate of deforested sample points. For both standard random sampled and stratified sampled datasets, sample points with no forest in the surrounding 500 m x 500 m target region in 2000 were removed. Sample points were spaced a minimum of 250 m apart to minimise spatial autocorrelation.

Mexican and Madagascan samples had 8000 and 7000 points respectively. 1000 points were removed from each sample and reserved as a final validation set for evaluating the best performing models of each method. To create the training and testing datasets, the remaining points in each sample were randomly split into two equal sized sets using repeated random subsampling. Both the training and testing set contained the same number of deforested samples points. This procedure was repeated 20 times, with replacement, creating the training and testing pairs used for all models. While greater confidence would be possible with a greater number of trials, the extensive time required to run each trial for the GPs meant that all models were restricted to 20 trials to allow for an equal assessment of all methods.

A range of predictors previously identified as relevant to deforestation (such as slope and surrounding deforestation) were selected as candidates for inclusion in the deforestation risk models developed in this study (Table 1). Predictors represented either measures of proximity to a sample point, the land use within the 500 m x 500 m target region surrounding a sample point, or the land use within the neighbourhood immediately surrounding the 500 m x 500 m target region. To select a land use neighbourhood size relevant to predicting deforestation risk, areas of 1 km x 1 km, 2 km x 2 km and 5 km x 5 km were tested over 20 trials using both the random and stratified random samples. The performance of each modelling method in predicting deforestation risk was assessed for each neighbourhood size using the True Skill Statistic (TSS) (this metric is explained in section 2.3). Based on the results, a 1 km x 1 km neighbourhood was selected for all modelling methods except for BNs, which performed better when a 5 km x 5 km neighbourhood was used.

All deforestation predictors were extracted from the georeferenced datasets using scripts written in Python 2.7 (Python Software Foundation 2012). Predictors representing land use (such as the distance to the forest edge) were derived from land cover values in 2000. Predictors measuring land use change (such as percentage of deforestation in the surrounding area) were derived from the change in land cover between 1990 and 2000. Fragmentation variables were calculated using Fragstats version 4.1 (McGarigal 2012) and were class-based metrics calculated on the forest class of land use in 2000. For road location, distance to cities and population pressure, the currently

available data was used, as no meta-data was available to clarify the dates represented in the datasets.

Table 1: Deforestation predictors considered in this study

Dataset	Feature	Predictor definition	Example of previous usage
Conservation International Land Use Change	Geographic Coordinates	Longitude, latitude of sample point	
	Surrounding deforestation	Distance to nearest deforestation in 2000, percentage of the area that was deforested between 1990 and 2000	Müller et al. (2011)
	Fragmentation of surrounding forest	Edge density, landscape division index, proximity index distribution, fractal index dimension, percentage of area forested (all measured for 2000)	Mas et al. (2004)
	Proximity to forest edge	Distance to edge of forest in 2000	McConnell et al. (2004)
World Database on Protected Areas	Surrounding protected areas	The percentage of the target region that falls within a PA, PA ID (if relevant) for PA containing the target region, PA IUCN category, PA size, distance to nearest PA (start (1990) middle (2000) and end of the study period (2004))	Gaveau et al. (2009)
Landsat Digital Elevation Data	Slope	Median slope of target region	Buchanan et al. (2008)
	Elevation	Median elevation of target region	Mas (2005)
	Ruggedness	Standard deviation of slope in target region	Müller et al. (2011)
MapAbility Road Location	Proximity to road	Distance to nearest road (up to 15 km)	Htun et al. (2013)
Landscan Population Data	Population pressure	The population density in either the 1 km or 3 km area surrounding the sample point	Laurance et al. (2002)
Natural Earth	Political boundary	The state political boundary. Used only as a control	-
	Proximity to river	Distance to nearest river (up to 15 km)	Laurance et al. (2002)
	Proximity to city	Distance to nearest city	(Mas et al. (2004)
WWF Ecoregions	Ecoregion	The ecoregion containing sample point. Used as a control in Mexico, but included as a variable in Madagascar due to variation within study region.	Vaca et al. (2012)

Multicollinearity among predictors may cause models, particularly GLMs, to become unstable and reduce the effects of any one variable (Aguilera et al. 2006). To reduce the possibility of this occurring, for any pair of predictors where the Pearson correlation coefficient was higher than 0.8 one variable was excluded. Both the Mexican and Madagascan datasets showed that landscape division index was heavily correlated with percentage of region forested. The former was therefore removed from the study.

For Mexico, median slope was excluded due to correlation with median elevation. The distance to nearest PA in the middle of the study period (2000) was excluded due to correlation with nearest PA at the start and end of the study period (1990 and 2005 respectively). For Madagascar, distance to the nearest PA at the start and end of the study period were removed due to correlation with distance to the nearest PA in the middle of the study period. In the Mexican region, only three PAs were present, resulting in insufficient examples to warrant the inclusion of the PA characteristics included for the Madagascan datasets. The distance to river was not available at the time the Mexican datasets were created and therefore was only included for Madagascar. Only one ecoregion was present for Mexico so this predictor was excluded from the Mexican datasets. Data for the GLMs, ANNs and GPs were normalised between 0 and 1.

2.2 Deforestation risk modelling

GLMs were implemented in the R programming language. Four models were tested; a GLM, a stepwise GLM and a GLMM (including X and Y coordinates of the sample points as random effects) using the MASS package (Venables 2002), and a GLM with interactions among the predictors using the glmulti package (Calcagno and de Mazancourt 2010). The glmulti package takes a maximum of 15 predictors and includes linear interactions between each combination of two or three predictors as separate predictors. Models testing for interactions were therefore implemented using the 15 most significant predictors from the GLM.

ANNs were implemented in Matlab 2014 (The MathWorks Inc 2014) using the ANN toolbox. ANNs used a resilient backpropagation learning algorithm and were allowed to run a maximum of 3000 epochs. During training, 350 sample points were selected as a validation set to help avoid overfitting. These were used to evaluate the network after each epoch and terminate training in any instances where no improvement was made after 300 epochs.

Ten separate sets of randomly selected starting weights were selected with those resulting in the lowest mean squared error being used for the final ANN. Initial trials were run using a single layer network and either 2, 5, 10, 20, 30, 60, 90, 120 or 150 hidden nodes, as well as a double layer network with 5, 10 or 30 nodes in each layer. From these trials a single layer network (with 10 or 60 hidden nodes) and a double hidden-layer ANN (with either 5, 10 or 30 nodes per layer) were tested for further trials. ANNs had two output nodes, representing 0 and 1. A prediction was interpreted as deforested when the predicted probability of a "1" was higher than for a "0".

Three different BN structures were implemented using the Netica software (Norsys Software Corp 2013); a naïve BN where predictor variables were considered independent, a Tree Augmented Naïve BN (TAN) that allowed for dependent relationships among predictor variables and an expert-designed BN where relationships among predictor variables were specified by experts. The expert-designed BN was developed with input from four experts with experience in deforestation. All experts were sent a description of the deforestation predictors that they could select from to construct the causal structure of the BN and an introduction to BNs. The expert-designed BN with the best TSS performance was selected for comparison with the other models.

For each BN, continuous variables were discretised using a custom R script developed for this study. The algorithm first split the range of a continuous variable into 150 equal interval buckets to ensure a fine initial discretisation. Any buckets with less than a specified minimum number of sample points, was merged with the adjoining bucket containing the least number of sample points. This was repeated until all buckets contained at least the specified minimum number of sample points. The minimum number of sample points was calculated as the number of points required to meet the 95% confidence level for that sample size (Moore 1996).

If, after satisfying the minimum number of sample points per bucket, the number of buckets exceeded 20, the buckets were further merged until a maximum of 20 buckets remained. This resulted in continuous nodes having a maximum of 20 states, with each state capturing the minimum number of sample points required for a 95% confidence level. A limit of 20 states was given to nodes to ensure that the BNs were computationally efficient (since the more states used in a BN, the larger the condition probability tables become and the less computationally efficient the model becomes) and all conditional probabilities within the models were learnt from a sufficient number of samples.

All GPs were run in Matlab, release 2014 (The MathWorks Inc 2014), using the GPML toolbox (Rasmussen and Nickisch 2010). Prior to the main training, a GP is presented with a subset of the

data from which it learns hyperparameters that define the covariance function. This stage of the process has significant runtime requirements (a single trial learning hyperparameters on 1500 points takes around ten hours to complete on a dedicated server). Runtime can be reduced by reducing the number of samples presented, however trials indicated this resulted in reduced performance. As GPs are a spatial model, they were also tested when trained using only the X and Y coordinates of sample points to test how well they would predict spatial patterns in deforestation with no predictors apart from the existing locations of deforestation.

2.3 Evaluation of datasets

The usefulness of the individual datasets was evaluated based on analysis of the predictor variables that were most influential in the GLMs and BNs. For the GLMs this was measured based on the p-values for each predictor. For the BNs, a sensitivity to finding analysis was run in Netica and the percent variance for each variable was used as a measure of its importance. Predictors with small p-vales or high sensitivities in the BNs were more important for predicting deforestation presence. Datasets providing these were considered as more useful in predicting deforestation risk in each study zone.

2.4 Model performance assessment

All models selected for comparison were evaluated on their ability to predict whether samples would be deforested in the 2000 – 2005 time period, including their ability to produce a reasonable map of deforestation risk for this time period. Models were evaluated over 20 trials using TSS, sensitivity and specificity, calculated using a confusion matrix (Table 2).

Table 2: *Confusion matrix used to assess the predictive performance of deforestation risk models*

		Actual Value	
		Recorded deforestation present	Recorded deforestation absent
Predicted Value	Predicted deforestation present	True positive (TP)	False positive (FP)
	Predicted deforestation absent	False negative (FN)	True negative (TN)

Sensitivity measures the proportion of observed presences (deforested sample points) that are predicted correctly, while specificity measures the proportion of observed absences (forested sample points) that are predicted correctly. TSS combines sensitivity and specificity (Table 3), with a value greater than 0 indicating better than random model performance (Allouche et al. 2006). The Kappa statistic and overall accuracy are commonly used performance metrics in deforestation modelling studies. We avoided their use to assess the performance of our deforestation risk models because these metrics become unreliable when prevalence rates are low (Allouche et al. 2006).

Table 3: *Metrics used to assess the predictive performance of deforestation risk models*

Metric	Description	Range	Formula
Sensitivity	Proportion of observed presences that are predicted correctly	0 - 1	$\frac{TP}{TP + FN}$
Specificity	Proportion of observed absences that are predicted correctly	0 - 1	$\frac{TN}{TN + FP}$
True Skill Statistic	Combined sensitivity and specificity (a value greater than 0 is better than random, a value less than 0 is worse than random)	-1 - 1	*Sensitivity + Specificity -1*

We used a 50% probability cut off to distinguish between predicted deforestation presence and absence (i.e. a sample point was categorised as deforested if it had more than a 50% predicted deforestation probability). To compensate for any bias caused by this, and evaluate the performance of models across a range of probability cut offs, the area under the receiver operating curve (AUC) was also used to assess model performance. AUC takes into account the sensitivity and specificity of predictions across all probability cut offs from 0 to 100% (Lobo et al. 2008).

The effect of selecting a 50% cut off was further examined by correcting for the known prevalence rate of deforested sample points within the datasets. To do this, sample points were ranked from highest to lowest predicted deforestation probability. Then the known deforestation prevalence in the datasets was used to categorise sample points as deforested or not. For example, if the known prevalence rate of deforested samples in the dataset is 10% then the top 10% of sample points (ranked from highest to lowest predicted deforestation probability) were categorised as deforested.

3 Results

Results presented in this section cover both the usefulness of the datasets and a comparison of the statistical and machine learning methods tested for predicting the probability of deforestation.

3.1 Model selection

Neither the GLM using stepwise selection or the GLM including interactions performed better than the standard GLM, however performance was improved when the X and Y coordinates of the sample points were modelled as random effects. The structure of the ANNs was found to have little influence on model performance (compared to other factors such as sample method) and a single layer ANN with 60 hidden nodes was selected for comparison with the other modelling techniques. The naïve and TAN BNs both outperformed the expert designed BNs, with the naïve BN scoring higher on sensitivity. GPs were run with 1500 sample points used for learning the hyperparameters, with the exception of the GPs trained only on geographic coordinates of the sample points for Madagascar, which required 2500 points to successfully learn the hyperparameters. The model designs selected for method comparison are summarised in Table 4.

Table 4: Final designs used in method comparison

Technique	Selected designs
GLM	GLM; GLMM with X and Y as random effects
ANN	Single hidden layer with 60 hidden nodes
BN	Naïve BN; TAN BN
GP	GP (all variables); GP (X and Y coordinates only)

Selected results for preliminary model selection for each technique are provided in Section 3 of the online supplementary material. Both naïve and TAN BNs are shown due to differences in their comparative performance across the different metrics. GLM and GLMM performance is also shown for comparison.

3.2 Dataset usefulness

Both GLMMs and BNs are able to measure the order of influence of predictors on model predictions. In both study areas, CI land use change proved to be the most influential dataset, with surrounding deforestation and distance to the nearest deforestation being amongst the most influential predictors in Mexico and edge density of the target region being the most influential predictor in Madagascar. Distance to the nearest city (from the NE dataset) and median elevation (from the Landsat DEM) were also important predictors. Several of the protected area variables from the WDPA dataset (percentage of target region protected and percentage of neighbourhood protected) were also important predictors in Mexico, but not in Madagascar (possibly due to the small amount of protected area within the Madagascan study area). The one dataset which was not free, Landscan population pressure, did not provide any significant predictors in either area. Detailed results of individual variable importance for the GLMs and BNs are provided in Section 4 of the supplementary online material.

3.3 Method Comparison

The performance results for the selected models of each machine learning method are presented in Figure 3 to Figure 6.

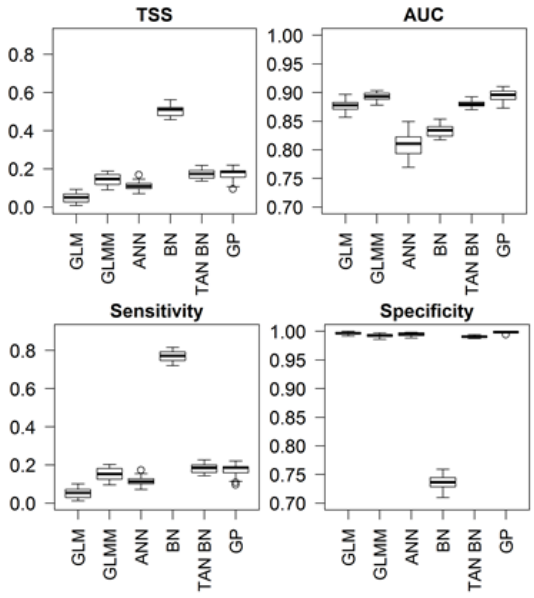

Figure 3: Results over 20 trials for the Mexican study region for models trained on randomly sampled points (note change in scale between plots). Naïve BNs had higher overall TSS scores as a result of high sensitivity. GLMMs outperformed both GLMs and ANNs. Boxplots show the minimum, maximum and median values as well as the 25th and 75th quartiles. Outliers are shown as circles.

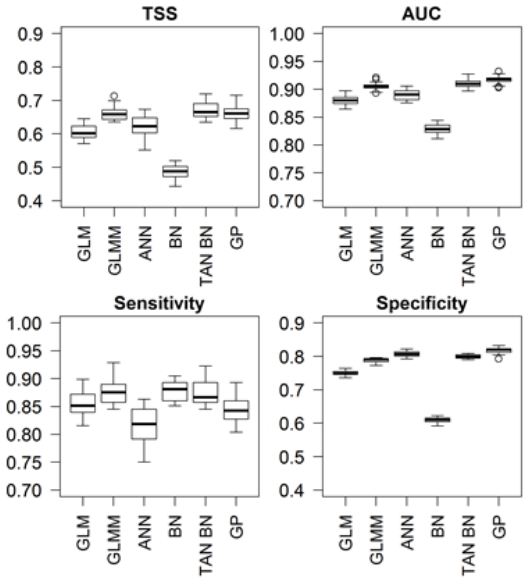

Figure 4: Results over 20 trials for the Mexican study region for models trained on stratified randomly sampled points (note change in scale between plots). Naïve BNs did not show the same improvement in performance when trained on stratified randomly sampled points as other models (the performance of naïve BNs remained stable when trained on randomly versus stratified randomly sampled points). Boxplots show the minimum, maximum and median values as well as the 25th and 75th quartiles. Outliers are shown as circles.

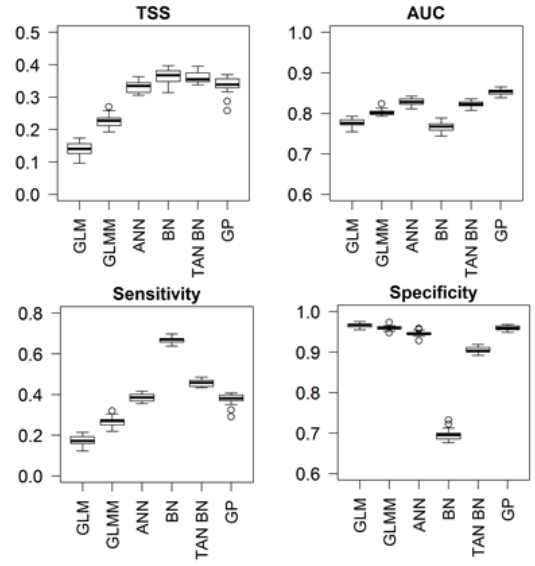

Figure 5: Results over 20 trials for the Madagascan study region for models trained on randomly sampled points (note change in scale between plots). In contrast to Figure 3, ANNs outperformed GLMMs. GPs were again amongst the best performing models. Boxplots show the minimum, maximum and median values as well as the 25th and 75th quartiles. Outliers are shown as circles.

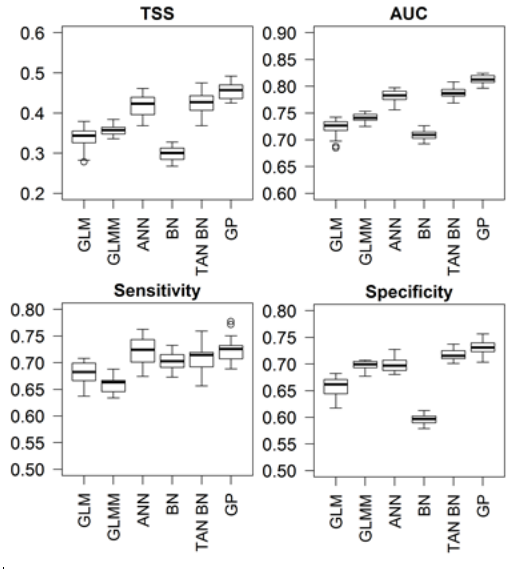

Figure 6: Results over 20 trials for the Madagascan study region for models trained on stratified randomly sampled points (note change in scale between plots). All models except the naïve BN again show an increase in sensitivity. Boxplots show the minimum, maximum and median values as well as the 25th and 75th quartiles. Outliers are shown as circles.

In most cases the GLMM had a higher median TSS and AUC score than the standard GLM, with the exception being models trained on stratified sampled data in the Madagascan study area (Figure 6). In this instance the TSS scores were similar for both. In both study areas, when trained on the randomly sampled points (Figure 3 and Figure 5), the naïve BN had higher sensitivity than the other models and a higher TSS score. This was a result of the naïve BN over predicting the amount of deforestation. The ANN outperformed the GLMM in Madagascar (Figure 5 and Figure 6), but the reverse was true for Mexico (Figure 3 and Figure 4).

3.4 Spatial Analysis

Mapped results for models that were retrained using all points from the training sample (n = 7000 for Mexico, n = 6000 for Madagascar) and tested on the unseen set of validation samples (n = 1000), showed that all models predicted deforestation in roughly the same locations. Predicted and actual deforestation maps for the Mexican study area (Figure 7) between 2000 and 2005 show how locations predicted to have a high probability of deforestation corresponded to areas where deforestation had actually occurred. The exception was the naïve BN, which over predicted deforestation occurrence. This is reflected in the high sensitivity scores obtained for the naïve BN (Figure 3 and Figure 5). Mapped results for the Madagascan study area are presented in Section 5

of the supplementary material, available online, as well as those for models trained on the stratified datasets (given in Section 6).

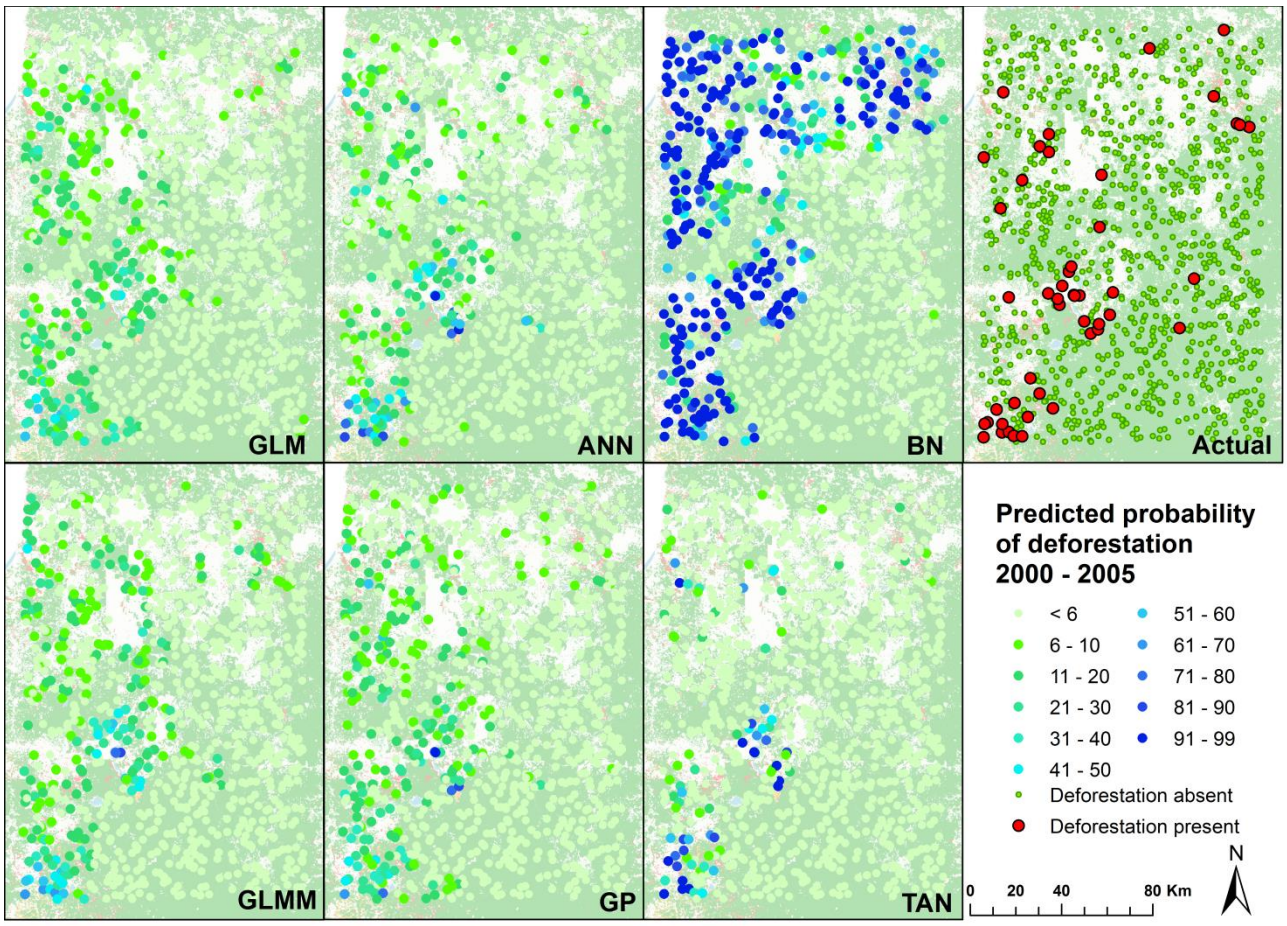

Figure 7*:* Predicted probability of deforestation (2000-2005) for models trained on randomly sampled points in the Mexico study area and tested on 1000 validation sample points previously unused in model training or testing.

The TSS scores for the models tested on the validation samples (Figure 8) show that adjusting deforestation predictions to reflect the known prevalence rate of deforested samples within the datasets improved the performance of models trained on randomly sampled points, but reduced the performance of the models trained on stratified randomly sampled data. This reduced the differences in TSS scores between models trained on standard randomly sampled and stratified randomly sampled points.

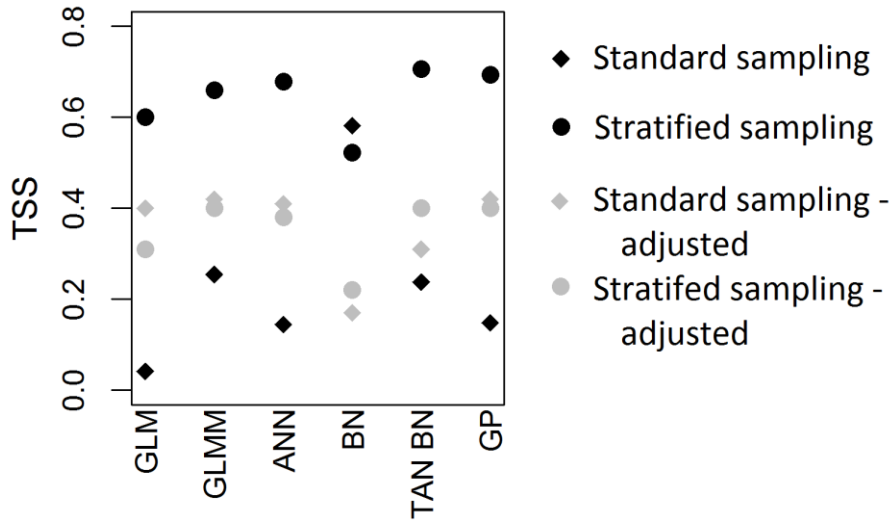

Figure 8: TSS results for models with the cut off for a deforested sample set as either a 50% probability of deforestation, or adjusted to replicate the known prevalence of deforested samples in the datasets).

3.5 GPs and deforestation location

The mapped predictions of the GP tested on the validation sample, when trained using only the X and Y coordinates of sample points as predictor variables (Figure 9), corresponded reasonably well with locations where deforestation had actually occurred. Mapped results for the Madagascan study area are presented in Section 5 of the supplementary material, available online.

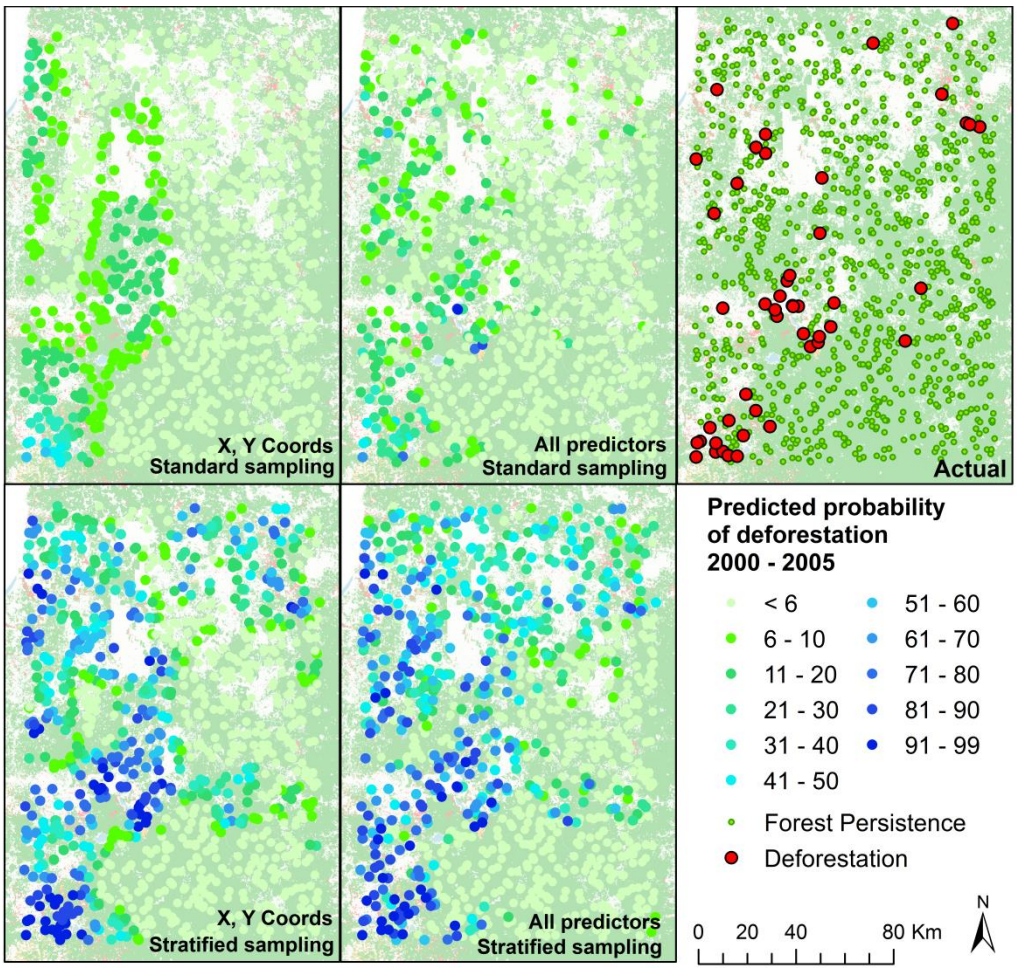

Figure 9: Probability of deforestation (2000-2005) of GPs for the Mexican study area trained using either only the geographic coordinates as predictors, or the entire set of predictors, and tested on the validation sample previously unused in model training or testing.

Including predictors in addition to the X and Y coordinates of the sample points resulted in slightly better model performance for both AUC and TSS (full results available in Section 7 of the supplementary online material). It should also be noted that, while a good result was generally obtained, GPs trained only on the geographic coordinates failed to produce reasonable predictions on any metric for several trials unless presented with additional sample points for learning the hyperparameters.

4 Discussion and Conclusion

Although obtaining datasets for deforestation studies can be challenging, the range and quality of freely available data sources is increasing. The datasets obtained for this study were sufficient to produce reasonable predictions of deforestation risk using even basic statistical models (GLMs and GLMMs), with AUC values generally above 0.8, which is considered good (Platts et al. 2008). The

freely available data allowed for a variety of different predictors to be used in modelling deforestation risk, which is crucial considering that the deforestation predictors can differ between regions. They also allowed different deforestation predictors to be used to represent the same deforestation pressure (for instance using population pressure from the Landscan dataset or distance to cities from the NE data to represent population pressure).

The CI data proved to be valuable for both the Mexican and Madagascan study areas, providing information on land use change from which the most influential deforestation predictors were derived (such as surrounding deforestation, distance to the nearest deforestation and edge density of the target region). The usefulness of the WDPA was naturally affected by the characteristics of the protected areas within each study area. The Mexican study area contains large, adjacent protected areas, creating an uninterrupted and sizable protected area categorised in one IUCN category. In contrast, the Madagascan study area has multiple small protected areas categorised in several IUCN categories. This meant that for each study area, different deforestation predictors were relevant for comparing the effect of PAs and illustrating the importance of having access to a variety of datasets.

In the Madagascan study area, distance to roads (from the MapAbilty dataset) was not a significant deforestation predictor for the GLM or BN models, and was only significant in the Mexican study area when a GLMM was used. Based on previous studies (Htun et al. 2013, Robinson et al. 2014) it is unlikely that an area's proximity to roads is entirely irrelevant to its deforestation risk. Our results are most likely a reflection of inaccuracies in the road dataset resulting from new roads continually being built with no road construction date being recorded. It is therefore difficult to match the timing of road construction with the timing of deforestation. When approximating population pressure, the distance to the nearest city (calculated from the NE data) proved to be more useful than population pressure derived from the Landscan data (which had to be paid for). This shows that proxy deforestation predictors obtained from freely available data may be just as useful as those obtained from purchased data.

The freely available data used in this study are not an exhaustive list, and all indications are that freely available geo-referenced datasets, such as those provided by Landsat and the WDPA, will continue to improve in quality and availability. The factors that affect deforestation differ from region to region and therefore the success of models trained on freely available datasets will naturally depend on whether these datasets contain data relevant to that area. The difference in the importance of variables between the Mexican and Madagascan study zones is possibly a reflection of the differences in the context of deforestation between the two regions (such as differences in underlying cause of deforestation or climatic factors). Efforts should be made to seek out data for

predictor variables that are considered to be relevant to the context. For example, datasets such as national census data could also prove valuable for approximating the causes of deforestation, such as fuel wood use (Davidar et al. 2010).

Despite theoretical advantages, ANNs only outperformed the more basic GLMMs in the Madagascan study area. This is consistent with current thinking that neither method consistently produces better results and supports a previously suggested approach to first try the basic GLMs and then move to ANNs (Tan et al. 2006). It should be noted that, although this study applied the same normalisation technique to the data for the GLMs, ANNs and GPs, the GLMs performance over the ANNs may have been improved had different normalisation or transformation techniques been applied to the data. In contrast to other studies (Bradshaw et al. 2007) modelling the geographic coordinates as random effects improved GLM performance in both regions, highlighting the importance of taking into account the spatial nature of deforestation when relying on statistical models. This was also evident in the performance of the GPs. Although performance was improved when predictors in addition to the geographic coordinates for sample points were used, the spatial nature of GPs meant they were able to make better than random predictions of deforestation risk when presented with only the X and Y coordinates of the sample points as predictors.

Although reasonable results for GPs were obtained when trained only the geographic coordinates of sample points, it is acknowledged that there are a number of limitations to this approach. First, models trained only on the X and Y coordinates of sample points are unable to account for sudden changes in terrain that could affect deforestation risk. Secondly, they often require more training samples to produce good prediction performance compared to models trained on a range of deforestation predictors. Nevertheless it does provide evidence that for a spatial process like deforestation, the ability of spatial models such as GPs to extrapolate patterns from surrounding areas may offer useful results when the geographic coordinates of existing deforestation locations is the only predictor available.

In the low prevalence rate situation (random sampling) the naïve BNs had higher TSS scores than the GLMs. This demonstrates the importance of using a range of metrics to evaluate model performance and supports the work of Lobo et al. (2008) and Platts et al. (2008), who propose that TSS, sensitivity and specificity are required to obtain a true picture of model performance. Furthermore, our results showed that models with poor TSS scores (which are based on a 50% probability cut off) may still produce high AUC scores, further supporting the notion that models should not be evaluated using a single metric.

The poor sensitivity and TSS results for all models (except the naïve BNs) trained using data derived from standard random sampling is at least in part a reflection of the low number of deforestation samples in the training data (i.e an imbalance between the number of deforestation presence and absence samples). This meant that the models had fewer examples of deforestation to learn from. Sensitivity and TSS were improved when models were trained using data derived from stratified sampling because the models had far more examples of deforestation from which to learn.

When assessed using AUC, model performance was less affected than the other metrics by how the training data were sampled. Training models using data derived from stratified sampling resulted in smaller improvements in AUC compared to the improvement seen in sensitivity and TSS. This is because AUC evaluates model performance across a range of probability cut-offs for deforestation presence rather than a fixed 50% probability cut off used by sensitivity and TSS. The fixed cut off used by sensitivity and TSS means that the prediction of deforestation presence only has to vary slightly above or below 50% for it to be classified as correct or incorrect, making the model performance results more sensitive to the training data.

The performance of BNs was more stable than the other modelling methods across the two different sampling strategies (standard random sampling and stratified random sampling). While GLMs, ANNs and GPs in this study performed relatively poorly when training on data derived from standard sampling (i.e. imbalanced data), the number of design options available for dealing with imbalanced data (Haibo and Garcia 2009) means that these methods should not be overlooked when one data class (in this case deforestation) is uncommon. We used stratified random sampling to correct for data imbalance because our data contained only a relative imbalance, meaning that enough deforestation samples existed in the data without having to generating synthetic or duplicated points.

The stratified sampling provided a balanced dataset (with equal numbers of forested and deforested points), which improved model performance on most metrics used in this study. However it also caused models to over-predict deforestation and resulted in a corresponding drop in specificity. It is also unsuitable in instances where a genuinely rare class is present. In these cases, data imbalance can still be corrected for using methods such as synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002).

The deforestation risk models tested in this study are presence/absence models and therefore their use is restricted to predicting the presence/absence of deforestation. While this was sufficient for the purpose of this study, and allowed the use of common performance metrics to evaluate the

models (such as TSS and AUC), it has the disadvantage that the models cannot be used to predict the amount or total area of deforestation, which would be useful for planning or zoning studies. A further limitation of the models we tested is that they assume deforestation predictors do not change over time. This is a common assumption amongst many deforestation models based on machine learning (Mas et al. 2004), but is particularly relevant for our models where the deforestation predictors were selected based on previously known predisposing deforestation risk factors (Geist and Lambin 2001). This means that our models are not able to account for changes in the influence of predisposing deforestation risk factors over time, and if they change in future, our models may become less reliable.

Despite these limitations, our results show that freely available datasets can be used to predict the probability of deforestation within the study zones. It is hoped that this will encourage those studying deforestation to consider what information may already be available to either compensate for missing datasets or complement existing ones. Furthermore we have shown that machine learning methods can be used to analyse these data and provide a reliable alternative to traditional statistical methods when modelling deforestation risk.

Acknowledgements

We would like to thank Conservation International, for making their dataset available to us for this study as well as the deforestation experts who assisted with the expert derived BNs. We would also like to thank Dr Toby Marthews for his assistance with the statistical models and Dr David Pullar for his early assistance in extracting the predictor variables from the spatial datasets. The input of Dr Lauren Coad in the design of the study is also gratefully acknowledged. Finally we are grateful for funding provided by an Australian Postgraduate Award and feedback from four anonymous reviewers.

References

Agarwal, D. K., Silander, J. J. A., Gelfand, A. E., Dewar, R. E. and Mickelson, J. J. G. (2005). "Tropical deforestation in Madagascar: analysis using hierarchical, spatially explicit, Bayesian regression models." Ecological Modelling **185**(1): 105-131.

Aguilera, A. M., Escabias, M. and Valderrama, M. J. (2006). "Using principal components for estimating logistic regression with high-dimensional multicollinear data." Computational Statistics & Data Analysis **50**(8): 1905-1924.

Allnutt, T. F., Asner, G. P., Golden, C. D. and Powell, G. V. (2013). "Mapping recent deforestation and forest disturbance in northeastern Madagascar." Tropical Conservation Science **6**(1): 1-15.
Allouche, O., Tsoar, A. and Kadmon, R. (2006). "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)." Journal of Applied Ecology **43**: 1223–1232.
Basse, R. M., Omrani, H., Charif, O., Gerber, P. and Bódis, K. (2014). "Land use changes modelling using advanced methods: Cellular automata and artificial neural networks. The spatial and explicit representation of land cover dynamics at the cross-border region scale." Applied Geography **53**(0): 160-171.
Bradshaw, C. J. A., Sodhi, N. S., Peh, K. S. H. and Brook, B. W. (2007). "Global evidence that deforestation amplifies flood risk and severity in the developing world." Global Change Biology **13**(11): 2379-2395.
Buchanan, G. M., Butchart, S. H. M., Dutson, G., Pilgrim, J. D., Steininger, M. K., Bishop, K. D. and Mayaux, P. (2008). "Using remote sensing to inform conservation status assessment: Estimates of recent deforestation rates on New Britain and the impacts upon endemic birds." Biological Conservation **141**(1): 56-66.
Calcagno, V. and de Mazancourt, C. (2010). "glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models." 2010 **34**(12): 29.
Campos-Taberner, M., Garcia-Haro, F. J., Moreno, A., Amparo Gilabert, M., Sanchez-Ruiz, S., Martinez, B. and Camps-Valls, G. (2015). "Mapping Leaf Area Index With a Smartphone and Gaussian Processes." Geoscience and Remote Sensing Letters, IEEE **12**(12): 2501-2505.
Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research **16**: 321-357.
DeFries, R. S., Rudel, T., Uriarte, M. and Hansen, M. (2010). "Deforestation driven by urban population growth and agricultural trade in the twenty-first century." Nature Geoscience **3**(3): 178-181.
Dudley, N. (2008). "Guidelines for Applying Protected Area Management Categories." IUCN: Gland, Switzerland.
Ellis, E. A. and Porter-Bolland, L. (2008). "Is community-based forest management more effective than protected areas?: A comparison of land use/land cover change in two neighboring study areas of the Central Yucatan Peninsula, Mexico." Forest Ecology and Management **256**(11): 1971-1983.
Estes, L. D., McRitchie, D., Choi, J., Debats, S., Evans, T., Guthe, W., Luo, D., Ragazzo, G., Zempleni, R. and Caylor, K. K. (2016). "A platform for crowdsourcing the creation of representative, accurate landcover maps." Environmental Modelling & Software **80**: 41-53.
Fenton, N. and Neil, M. (2013). Risk Assessment and Decision analysis with Bayesian Networks. New York, CRC Press.
Frayer, J., Sun, Z., Müller, D., Munroe, D. K. and Xu, J. (2014). "Analyzing the drivers of tree planting in Yunnan, China, with Bayesian networks." Land Use Policy **36**(0): 248-258.
Gaveau, D. L. A., Epting, J., Lyne, O., Linkie, M., Kumara, I., Kanninen, M. and Leader-Williams, N. (2009). "Evaluating whether protected areas reduce tropical deforestation in Sumatra." Journal of Biogeography **36**(11): 2165-2175.
Geist, H. J. and Lambin, E. F. (2001). "What Drives Tropical Deforestation? A meta-analysis of proximate and underlying causes of deforestation based on subnational case study evidence." LUCC Report Series No. 4.
Green, J. M. H., Larrosa, C., Burgess, N. D. and Balmford, A. (2013). "Deforestation in an African biodiversity hotspot: Extent, variation and the effectiveness of protected areas." Biological Conservation **164**: 62-72.
Haibo, H. and Garcia, E. A. (2009). "Learning from Imbalanced Data." Knowledge and Data Engineering, IEEE Transactions on **21**(9): 1263-1284.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, Springer.
Haykin, S. (2009). Neural Networks and Learning Machines. Ontario, Pearson.
Htun, N., Mizoue, N. and Yoshida, S. (2013). "Changes in Determinants of Deforestation and Forest Degradation in Popa Mountain Park, Central Myanmar." Environmental Management **51**(2): 423-434.
Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H. and Arriaga-Weiss, S. (2010). "Classification in conservation biology: A comparison of five machine-learning methods." Ecological Informatics **5**(6): 441-450.
Laurance, W. F., Albernaz, A. K. M., Schroth, G., Fearnside, P. M., Bergen, S., Venticinque, E. M. and Da Costa, C. (2002). "Predictors of deforestation in the Brazilian Amazon." Journal of Biogeography **29**(5-6): 737-748.
Liedloff, A. C. and Smith, C. S. (2010). "Predicting a 'tree change' in Australia's tropical savannas: Combining different types of models to understand complex ecosystem behaviour." Ecological Modelling **221**(21): 2565-2575.
Lobo, J. M., Jiménez-Valverde, A. and Real, R. (2008). "AUC: a misleading measure of the performance of predictive distribution models." Global Ecology and Biogeography **17**(2): 145-151.
mapAbility. (2012). "VMAP0." Retrieved 20/07, 2015, from http://www.mapability.com/index1.html?http&&&www.mapability.com/info/vmap0_index.html .
Mas, J. F. (2005). "Assesing protected area effectiveness using surrounding (buffer) areas environmentally similar to the target area." Environmental Monitoring and Assessment **105**: 69-80.
Mas, J. F., Puig, H., Palacio, J. L. and Sosa-López, A. (2004). "Modelling deforestation using GIS and artificial neural networks." Environmental Modelling & Software **19**(5): 461-471.
McConnell, W. J., Sweeney, S. P. and Mulley, B. (2004). "Physical and social access to land: spatio-temporal patterns of agricultural expansion in Madagascar." Agriculture, Ecosystems & Environment **101**(2-3): 171-184.
McGarigal, K., SA Cushman, and E Ene (2012). FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps.**:** Computer software program produced by the authors at the University of Massachusetts, Amherst.
Moore, D. S. M., G.P. (1996). Introduction to the Practice of Statistics. New York, W. H. Freeman and Company.
Müller, R., Müller, D., Schierhorn, F. and Gerold, G. (2011). "Spatiotemporal modeling of the expansion of mechanized agriculture in the Bolivian lowland forests." Applied Geography **31**(2): 631-640.
NASA. (2015, June 4, 2015). "Landsat 8." Retrieved 10/06/2015, from landsat.gsfc.nasa.gov/?p=3186.
NE. (2013a). "Admin 1 – States, Provinces." Retrieved 06/06/2013, from http://www.sciencemag.org/citmgr?gca=sci%3B342%2F6160%2F850.
NE. (2013b). "Public domain map dataset." Retrieved 29/04/2014, from http://www.naturalearthdata.com/.
Newman, M. E., McLaren, K. P. and Wilson, B. S. (2014). "Assessing deforestation and fragmentation in a tropical moist forest over 68 years; the impact of roads and legal protection in the Cockpit Country, Jamaica." Forest Ecology and Management **315**: 138-152.
Norsys Software Corp (2013). Netica Bayesian Belief Network software. https://www.norsys.com.
Olson, D. M. and Dinerstein, E. (2002). "The Global 200: Priority Ecoregions for Global Conservation." Annals of the Missouri Botanical Garden **89**(2): 199-224.
Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F.,

Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P. and Kassem, K. R. (2001). "Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity." BioScience **51**(11): 933-938.

Pérez-Vega, A., Mas, J.-F. and Ligmann-Zielinska, A. (2012). "Comparing two approaches to land use/cover change modeling and their implications for the assessment of biodiversity loss in a deciduous tropical forest." Environmental Modelling & Software **29**(1): 11-23.

Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I. and Dholakia, M. B. (2016). "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)." Environmental Modelling & Software **84**: 240-250.

Platts, P. J., McClean, C. J., Lovett, J. C. and Marchant, R. (2008). "Predicting tree distributions in an East African biodiversity hotspot: model selection, data bias and envelope uncertainty." Ecological Modelling **218**(1–2): 121-134.

Python Software Foundation (2012). Python Language Reference. http://www.python.org

Rasmussen, C. and Williams, C. (2006). Gaussian Processes for Machine Learning. Massachusetts, MIT Press.

Rasmussen, W. and Nickisch, H. (2010). "Gaussian Processes for Machine Learning (GPML) Toolbox." Journal of machine learning research **11**.

Reuter, H. I., Nelson, A. and Jarvis, A. (2007). "An evaluation of void-filling interpolation methods for SRTM data." International Journal of Geographical Information Science **21**(9): 983-1008.

Robinson, B. E., Holland, M. B. and Naughton-Treves, L. (2014). "Does secure land tenure save forests? A meta-analysis of the relationship between land tenure and tropical deforestation." Global Environmental Change **29**: 281-293.

Rogers, H. M., Glew, L., Honzák, M. and Hudson, M. D. (2010). "Prioritizing key biodiversity areas in Madagascar by including data on human pressure and ecosystem services." Landscape and Urban Planning **96**(1): 48-56.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). "Learning representations by back-propagating errors." Nature **323**(6088): 533-536.

Tan, C. O., Özesmi, U., Beklioglu, M., Per, E. and Kurt, B. (2006). "Predictive models in ecology: Comparison of performances and assessment of applicability." Ecological Informatics **1**(2): 195-211.

Tayyebi, A., Pijanowski, B. C., Linderman, M. and Gratton, C. (2014). "Comparing three global parametric and local non-parametric models to simulate land use change in diverse areas of the world." Environmental Modelling & Software **59**: 202-221.

The MathWorks Inc (2014). MATLAB and Statistics Toolbox Release 2014b. Massachusetts, United States., The MathWorks Inc.

UT-Battelle, L. (2013). LandScan (2013)™ High Resolution global Population Data Set O. R. N. Laboratory.

Uusitalo, L. (2007). "Advantages and challenges of Bayesian networks in environmental modelling." Ecological Modelling **203**(3-4): 312-318.

Vaca, R. A., Golicher, D. J., Cayuela, L., Hewson, J. and Steininger, M. (2012). "Evidence of Incipient Forest Transition in Southern Mexico." PLOS one **7**(8).

van Vliet, J., Bregt, A. K., Brown, D. G., van Delden, H., Heckbert, S. and Verburg, P. H. (2016). "A review of current calibration and validation practices in land-change modeling." Environmental Modelling & Software **82**: 174-182.

Venables, W. N., Ripley, B. D. (2002). Modern Applied Statistics with S. New York.

Wang, G., Oyana, T., Zhang, M., Adu-Prah, S., Zeng, S., Lin, H. and Se, J. (2009). "Mapping and spatial uncertainty analysis of forest vegetation carbon by combining national forest inventory data and satellite images." Forest Ecology and Management **258**(7): 1275-1283.

WDPA. (2010). "World Database on Protected Areas." Retrieved 06/06/2013, from http://www.wdpa.org/Default.aspx.

Yan, J., Li, K., Bai, E.-W., Deng, J. and Foley, A. M. (2016). "Hybrid Probabilistic Wind Power Forecasting Using Temporally Local Gaussian Process." IEEE Transactions on Sustainable Energy $V 7(1): 87-95.