

Prediction of large spatio-temporal data using machine learning methods

Brigitte Colin

Bachelor of Honours Cartography and Geomedia
Applied Sciences Munich University,
Bachelor Computer Science
Applied Sciences Pfungstadt University

under the supervision of

Principal Supervisor: Distinguished Prof Kerrie Mengersen
Associate Supervisors: Dr Alan Woodley, Dr Michael Schmidt, Dr Kim Lowell



Mathematical Sciences
Faculty of Science and Engineering
Queensland University of Technology

2019

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF
DOCTOR OF PHILOSOPHY

Abstract

This dissertation reports the methods and results of an interdisciplinary research approach of applied statistics, a supervised machine learning method and remotely sensed data to develop a prediction model that addresses the challenging nature of spatial dependencies and autocorrelation in spatial data sets. The project begins within a suitability study to investigate if a Boosted Regression Tree (BRT) model can accommodate the characteristics of the data, achieve a good model fit and yield a high prediction accuracy. In a comparison with other regression tree methods namely Least Absolute Shrinkage and Selection Operator, (LASSO) and Random Forest (RF), BRT outperformed those and showed the highest prediction accuracy and best model fit.

To address the high data volume, spatial autocorrelation and local dependencies of remotely sensed data we investigated four spatial data aggregation and spatial smoothing resolutions that simultaneously reduce computational cost and maintain local characteristics of the underlying land cover information. Our aim was to identify how spatial aggregation and spatial smoothing affects the intrinsic characteristics of green vegetation cover and the prediction accuracy of a BRT model.

By analysing prediction accuracy, computational speed and prediction raster maps we identified one resolution that best addressed the three criteria specified above. In the next step, the best spatial resolution was used for a long-term study covering 30 years of raster data on heterogeneous grazing land to investigate localised spatio-temporal trends in green vegetation cover. The conclusion of the dissertation is that BRT is a robust and accurate non-linear supervised machine learning method that addresses data-driven challenges, offers a wide range of interpretation, visualisation and variable selection tools and can deal with missing data by default, which is especially crucial in processing spatial and raster data sets.

Declaration

I hereby declare that this submission is my own work and to the best of my knowledge it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at QUT or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by colleagues, with whom I have worked at QUT or elsewhere, during my candidature, is fully acknowledged.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signature [QUT Verified Signature](#) Date 18.07.2019

Acknowledgements

I would like to thank my supervisors, Kerrie Mengersen, Michael Schmidt, Alan Woodley and Kim Lowell, for their mentorship and feedback throughout my candidature. Their expertise and guidance has had a great impact on this thesis and on my professional and personal development. The satellite images in this thesis were supplied by DSITI. Thanks particularly to Michael Schmidt for his assistance. I gratefully acknowledge the financial support of ACEMS, QUT, and the Australian Agriculture Company that provided the funding for our consultancy project at the beginning on my PhD. Thanks also to past and present members of the BRAG research group and ACEMS for your camaraderie and for fostering a great work environment at QUT. Last but by no means least, I express my sincere gratitude to my family and friends for their support and patience.

List of Publications Arising from this Thesis

- Chapter 3: **Student Brigitte Colin**, Samuel Clifford, Paul Wu, Samuel Rathmanner, Kerrie Mengersen (2017). Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making. *Open Journal of Statistics*. vol 7, number 5:page 1 - 17. doi:10.4236/ojs.2017.75061.
- Chapter 4: **Student Brigitte Colin**, Michael Schmidt, Samuel Clifford, Alan Woodley, Kerrie Mengersen (2018). Influence of Spatial Aggregation on Prediction Accuracy of Green Vegetation using Boosted Regression Trees. *Remote Sensing*. vol 10, number 8:page 1260. doi:10.3390/rs10081260.
- Chapter 5: **Student Brigitte Colin**, Alan Woodley, Kerrie Mengersen (2018). Analysis of spatial smoothing effects on green vegetation to improve prediction accuracy using Boosted Regression Trees.
- Chapter 6: **Student Brigitte Colin**, Kerrie Mengersen (2018). Estimating Spatial and Temporal Trends in Environmental Indices Based on Satellite Data: A Two-Step Approach. *Sensors, Computational Intelligence in Remote Sensing*. vol 19, number 2:page 361. doi.org/10.3390/s19020361

Contents

Abstract	i
Declaration	iii
Acknowledgements	v
List of Publications Arising from this Thesis	vii
Chapter 1 Introduction	5
1.1 Research aim and objectives	7
1.2 Thesis outline	8
1.3 Scope of the study	10
Chapter 2 Literature Review	11
2.1 Big data characteristics	11
2.2 Monitoring land use and land change with Landsat data	12
2.3 Statistical machine learning models	14
2.4 Benefits of using BRT	16
2.5 Boosted Regression Trees	17
2.6 Applications of BRT in agriculture and ecology	21
2.7 Review of Software	22
Chapter 3 Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making	25
Chapter 4 Effect of Spatial Aggregation on Prediction Accuracy of Green Vegetation using Boosted Regression Trees	45
Chapter 5 Analysis of spatial smoothing effects on green vegetation to improve prediction accuracy using Boosted Regression Trees	77
5.1 Introduction	79
5.2 Material	80
5.2.1 Case Study	80
5.2.2 Fractional Cover Data	80
5.3 Research design	81
5.3.1 Neighbourhood analysis	82
5.3.2 Gaussian Processes	82
5.3.3 Boosted Regression Tree	84
5.4 Results	85
5.4.1 Gaussian Kernel smoothing	85
5.4.2 Effect of smoothing on BRT prediction results	88

5.5	Discussion	91
5.6	Future Work	93
Chapter 6	Estimating Spatial and Temporal Trends in Environmental Indices Based on Satellite Data: A Two-Step Approach	95
Chapter 7	Discussion	115

List of Figures

1.1	Overview of thesis objectives and chapters.	8
2.1	Sub-figure (a) shows a graphical demonstration of the hierarchical regression and binary splitting process at the nodes of the BRT and how the observed values will be transported along the tree branches. In sub-figure (b) we demonstrate the ensemble approach of the boosting algorithm as part of the BRT. Binary splits indicated as red straight lines separate the data in grey and white sections and so called weak learners are created as seen in Equation (2.1). The combination of weak learners, to form one strong prediction rule is managed by the boosting algorithm. The BRT method yield a more accurate prediction accuracy through generating flexible boundaries and therefore allowing the identification of small areas of interest. Adapted from (Matteson, 2013).	19
5.1	Spectral unmixing approach explained graphically using the Instantaneous Field of View (IFOV) as the geometric resolution of one FCover pixel where the spectral information and the fractions of three objects on the ground are combined together. The spectral unmixing approach aims to separate the unique reflected or emitted radiations and to derive three map layers for each endmember, here photosynthetic/green vegetation (PV), non-photosynthetic vegetation (nPV) and bare soil (BS) (Kamal & Phinn, 2011).	81
5.2	The magnitude of the smoothing is depicted in red on the original data in black using four kernel sizes from ranging from $\sigma = 0.2$ (most smoothing) at the top to $\sigma = 20$ (least smoothing) on the bottom. The y-axis shows the green vegetation fractions and the x-axis demonstrates their unique location ID's plotted as a 1-dimensional vector from top left to bottom right.	85

-
- 5.3 Original Raster image (not smoothed) showing the green vegetation fractions where higher values represent a higher fraction of green vegetation represented in the individual pixel. 86
- 5.4 Plots of the Gaussian smoothing showing (a) smoothed raster maps on the left and (b) raster maps of the distribution of the residuals on the right. The top panel shows the kernel size $\sigma = 0.2$ (maximum smoothing), followed by $\sigma = 1$, the third shows the smoothing using $\sigma = 10$ and the last panel was performed using $\sigma = 20$ (the least smoothing). 87
- 5.5 Boxplots of the four different Gaussian smoothing kernels and the original values. From left to right: $\sigma = 0.2$, $\sigma = 1$, $\sigma = 10$, $\sigma = 20$ and the original values of not smoothed green vegetation fractions. 88
- 5.6 Marginal plots of the four different Gaussian kernels show that BRT underpredicts the green vegetation fractions shown on the y-axis in comparison to the observed values on the x-axis in (a) $\sigma = 0.2$ (maximum smoothing); (b) $\sigma = 1$; (c) $\sigma = 10$ (d) $\sigma = 20$ (the least smoothing). The distribution of the observed values are positively skewed and show a long tail starting at about 130 up to 140 and higher whereas the predicted values are less skewed and reach the maximum of 130 in histogram of the margin of the plot. 90
- 5.7 Relative Influence plots of the four Gaussian smoothing kernels showing the influence of the three covariates in predicting the green vegetation fractions (response variable) on the x-axis where (a) used $\sigma = 0.2$ (maximum smoothing); (b) $\sigma = 1$; (c) $\sigma = 10$ (d) $\sigma = 20$ (the least smoothing). . . . 91

List of Tables

5.1	BRT prediction results on two scenarios. Scenario 1: only latitude and longitude have been used as covariates. Scenario 2: latitude, longitude and the smoothed values have been used to predict the observed green vegetation fractions (original values). The best RMSE is printed in bold numbers and the significance of the covariate in predicting the response and their importance in the splitting process of the BRT is depicted as their relative influence in %	89
-----	---	----

1 Introduction

Statistical machine learning methods play an important role in analysing complex relationships, in extracting meaningful information and in providing predictive capabilities and results that can assist in a better informed decision making process in real world applications. The opening of the Landsat archive and a new open data policy have revolutionised the use of Landsat data (Wulder, Masek, Cohen, Loveland, & Woodcock, 2012). The Fractional Cover (FCover) (Scarath, 2012) product is a derived product from Landsat imagery and provides fractional cover representation of the proportions of green or photosynthetic vegetation, non-photosynthetic vegetation, and bare surface cover across the Australian continent in three separate layers (Guerschman et al., 2015; Muir et al., 2011). Modelling spatial data for land use and land cover (LULC) analyses using Landsat imagery has great potential since the Landsat data are freely available, cover a wide geographical area, and it avoids expensive, extensive and often impractical in-situ measurements (Irons, 2018, December 10; J.Walsh, Crawford, Welsh, & A.Crews-Meyer, 2001). In addition, the temporal resolution of Landsat imagery, meaning that every 16 days new raster data are available, enables us to perform extensive spatio-temporal analysis on LULC (Irons, 2018, December 10). In this thesis, attention is focused on the use of Landsat imagery for estimation and prediction of green vegetation, with the aim of providing insight and evidence about its change, quality and quantity.

A common challenge in dealing with satellite imagery is its sheer data volume. A FCover scene is about 350 MB and consists of 50 million pixels. One way of reducing the data volume is to derive descriptive statistics from regularly or irregularly shaped polygons comprising homogeneous pixels contained within the given extent. Spatial data possess autocorrelation in that areas that are close together share similar characteristics. In this thesis, this intrinsic characteristic is incorporated in a data reduction approach in which we aggregated pixels contained in an evenly spaced spatial grid overlaid on our FCover scene to delineate the arithmetic mean of the individual grid cells. Each spatial grid cell accounts for spatial autocorrelation effects in remotely sensed data in four different spatial resolutions in a given extent. To further refine the modified raster data and to address issues of the signal-noise ratio, a spatial filter such as a moving kernel can be used to smooth over the imagery to reduce the variance and improve the overall prediction

accuracy. This results in a substantial volume reduction, but can still account for local vegetation characteristics.

As noted above, there is a variety of machine learning methods that can be employed for estimation and prediction of green vegetation and LULC analysis using Landsat data (DeFries & Chan, 2000; Irons, 2018, December 10; Robinzonov, 2013; Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012). In this thesis, a Boosted Regression Tree (BRT) is adopted as the main methodology to address the challenges in predicting green vegetation in this thesis (De'ath, 2007; De'ath & Fabricius, 2000; Elith & Leathwick, 2017; Elith, Leathwick, & Hastie, 2008). BRT is a popular statistical, hierarchical and supervised machine learning approach that has been applied to remotely sensed data in various studies (De'ath, 2007; Elith et al., 2008). BRTs have only recently been extended to incorporate spatial and temporal data features that are characteristic of remotely sensed data (Elith & Leathwick, 2017; Elith et al., 2008; Emelyanova, McVicar, Van Niel, Tao Li, & Van Dijk, 2013). A BRT consist of two algorithms, namely a binary Regression Trees approach and a Boosting component and arguably yield higher prediction accuracy than simple tree-based methods such as a Classification and Regression Trees (CART) (Elith et al., 2008).

There are two major advantages of using BRT over more traditional regression methods. First, it allows a more flexible partition of the feature space that is not as rigid as using a simple linear regression. BRT combines simple binary partitions to form a complex prediction rule that can more accurately identify small areas of interest. Second, it can deal with missing values by default like masked out areas (clouds and cloud shadows), water bodies or the Scan Line Error of Landsat 7 ETM+ (Irons, 2018, December 10).

In this thesis, we focus on a single study area to demonstrate and evaluate the proposed methodology. The study area is located in the Northern Territory, Australia. The location of the FCover scenes at the Landsat footprint of path 102 row 72 on the Worldwide Reference System-2 (WRS-2) and covers an area of 185km x 185km. Our study area is defined as “dry” with variations of “desert, hot arid” and “dry summer, hot arid” (BWh and Bsh) based on the Koeppen-Geiger scheme and is very vulnerable with regard to climate variability (Chen, 2017). A quantitative estimation of green vegetation in semi-arid grazing land is our primary interest in this case study and the results are important for agricultural managers. Our study area is a heterogeneous region with a complex topography of native grass types. To give insight in the topology we created a Digital Elevation Model (DEM) using freely available Shuttle Radar Topography Mission (SRTM) data. The highest point is 255m and the lowest is located at 23m above mean sea level (MSL) referenced to the Australian Height Datum (AHD).

This research will bring together the different domains of remote sensing, statistical methods, supervised machine learning algorithms, big data, geoscience, remote sensing and agricultural and environmental sciences. We aim to develop a computationally efficient modeling approach with the focus on prediction accuracy. The intent is to make processes, patterns and relationships more transparent and enable more confident decision making based on the improved predictions.

1.1 Research aim and objectives

The overarching aim of this PhD project is to develop spatio-temporal decision tree models using big spatial data, with applications to environment and agriculture. In particular, the methodological focus will be Boosted Regression Trees (BRT), the data will be derived from remotely sensed satellite images, and the applied focus will be on the estimation and prediction of active photosynthetic land cover using geographic coordinates.

In order to achieve this aim, the project has three main methodological and four applied objectives.

M1: Development of a customised BRT model that incorporates spatial data sources with different granularities, data characteristics, noise and missing data following a big data approach.

- A1: Application of the BRT model to Landsat image data and other environmental and non-spatial agricultural data for estimation of grass biomass using a surrogate variable

M2: Development of a customised BRT models using spatial aggregation and spatial smoothing on aggregated green vegetation to assess the influence on prediction accuracy.

- A2: Application of the BRT model by including aggregated spectral information of remotely sensed green vegetation land cover data to assess the influence of a spatial aggregation scheme on prediction accuracy, using centroid coordinates as surrogate variables.
- A3: Analysis of spatial smoothing effects on green vegetation to improve prediction accuracy using Boosted Regression Trees.

M3: Extension of BRT models developed in M2 to predict spatio-temporal green vegetation trends as our newly created response variable in a long-term time series approach on aggregated green vegetation fractions.

- A4: Extension of the model to include location based slope regression trends to understand the seasonal trends on aggregated green vegetation averaged over 30 years based on a spatial grid.

The aims M1 and A1 are addressed in chapter 3. Aim M2 and A2 are addressed in chapter 4. Aim M3 and A3 are addressed in chapter 5, and Aim M3 and A4 will be addressed in chapter 6. The chapter overviews are displayed in Figure 1.1.

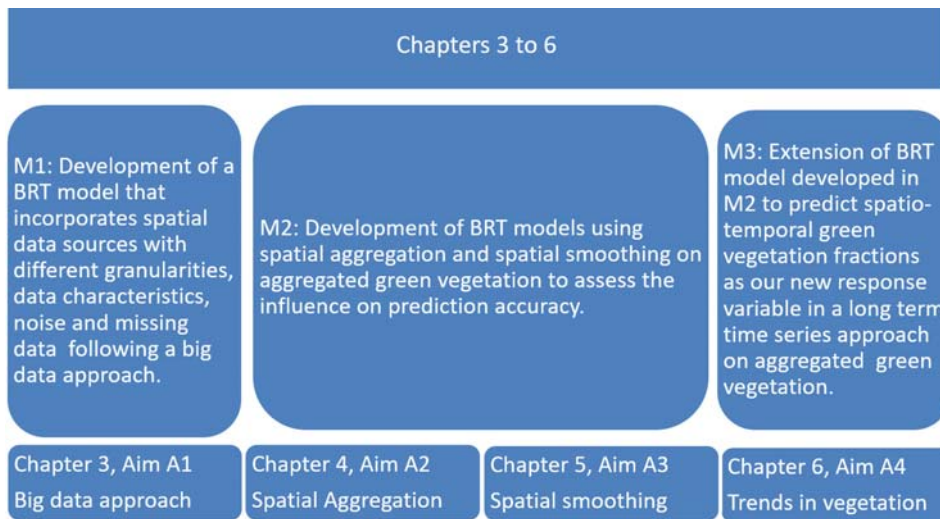


Figure 1.1: Overview of thesis objectives and chapters.

The results of this thesis will fill these significant gaps: 1.) Methods: Contribute to statistical and machine learning methods for big data showing different granularities, data structures, data characteristics, spatial and temporal patterns and missingness. 2.) Application: Analysis of suitability of using Landsat data and derived products for predicting green vegetation cover in semi-arid heterogeneous land.

1.2 Thesis outline

This thesis is presented in the style of a traditional monograph where the chapters are structured in the style of journal articles. Because the thesis is structured in this way, there is some overlap in the introductions and concept descriptions within the chapters. A separate set of references is provided for each chapter. A more comprehensive review of the literature and set of references is presented in Chapter 2 and at the end of the thesis.

The objectives of the thesis are addressed as follows: Chapter 3 addresses the applied aims A1 and M1 and resulted in my first published paper. Chapter 4 address the aim A2

and M2 and resulted in my second published paper. Chapter 5 addresses aim A3 of M3 and resulted in my third published paper and Chapter 6 addresses the aim A4 of M3 and is presented as a draft paper.

The literature review in Chapter 2 provides a summary of the current relevant literature on big data challenges, remotely sensed data like Landsat imagery and its derived product fractional cover data. Further, it addresses monitoring of spatio-temporal land use, land change, and supervised statistical machine learning methods on large spatial data sets.

In Chapter 3 the goal was to establish a supervised statistical machine learning model which incorporates various types of heterogeneous data, showing very different characteristics and data granularities, directly observed or measured, e.g. stocking data, Landsat Imagery, derived from existing data, e.g. Vegetation Indices, Fractional Cover, and Data as an output from a model, e.g. AussieGRASS (Carter & Bruget, 2015). We investigated the relationship between the variables in the data set that best predicts green vegetation. In the exploratory data analysis we analysed the different data sets in order to summarise their main characteristics. In further steps this new knowledge was combined with previous knowledge in order to refine the model successively and to gradually identify strong predictors and remove weaker ones to decrease computational processing time. Altogether four scenarios have been created to identify the best covariates for the following model building process covering a time frame from 2006 to 2012.

Chapter 4 presents the development and application of a data reduction scheme applied on FCover green vegetation data. Working with big spatial data processing is computationally expensive. The proposed data reduction scheme accounts for reduction scheme that account for spatial autocorrelation by combining nearby pixels together to create new FCover data in four different aggregation resolutions. Data reduction steps without loss of information are a scientific and computational challenge but are critical to enable effective data processing and information delineation in data-rich studies. We used all four spatial aggregation resolutions in our BRT modelling approach to investigate the influence of the aggregation and impact on prediction accuracy of green vegetation.

Chapter 5 presents a two step approach in using a combination of a simple linear regression model to extract slope parameters that represent green vegetation trends and those were used as our new response variable in the BRT to predict location based green vegetation trends. The geographic coordinates of the extracted slope coefficients were used to predict, visualise and understand the temporal and spatio-temporal variability on aggregated green vegetation in a long-term time series approach on aggregated green vegetation fractions covering 30 years.

Chapter 6 proposes that applying a spatial filter reduce the noise in the imagery and therefore spatial smoothing enhance the predictive performance of BRT. It is currently set up as a draft version.

Chapter 7 provides a summary of the key implications from Chapters 3 to 6 and will draw together these separate chapters and the body of research therein. The chapter also discusses the strengths and limitations of the work presented, and outlines the significance of the research completed and its implications for the field. Finally, the questions raised as a result of this work will be outlined, presenting future avenues for research answering these questions.

1.3 Scope of the study

The scope of the research presented in this thesis was determined largely by three factors: the nature of the case study, the available remote sensing data and the chosen statistical machine learning methodology. As described above, the case study focuses primarily on estimating green vegetation cover. For the first paper we are using an hierarchical and supervised machine learning method, namely Boosted Regression Tree. For the first paper the study area spread from the Northern Territory to Central Queensland and showed disjunct regions. For the remaining three papers the study was conducted on the case study area in the Northern Territory, described above. This study area provides sufficient challenges in methodology and variety in landscape to allow clear development of the aims and objectives described above.

As discussed above, attention is confined to BRTs as the analytic tool. Comparison of BRTs with other candidate statistical and machine learning methods is undertaken as part of the literature review in Chapter 2, but is not a focus of the substantive research in this thesis. Similarly, we focus on geographic coordinates as covariate information to account for spatial dependencies and to obtain predictions that are location based. Since the interest is on the utility of the geographic coordinates alone, other covariates such as elevation, rainfall, temperature, etc are not considered.

The spatial data we used are satellite data such as MODIS (Wolfe, 2018, December 10) and Landsat (Irons, 2018, December 10) and derived products such as Vegetation Indices, and fractional cover data based on spectral values of optical Landsat imagery. Other spectral data from optical sensors might also be suitable but have not been considered in our study. Our focus was on prediction green vegetation such as grass and pasture.

2 Literature Review

In this literature review, the necessary background and state of current methods will be detailed and critically reviewed to motivate the objectives previously defined in Section 1.1. The review is organised in three main sections. Section 2.1 will give an introduction to big data, their properties, challenges and why we need to find non-traditional methods to address those challenges to extract knowledge out of data. A literature review has been undertaken in order to identify a method which can deal with big data challenges and missing values by default without the need for infilling or interpolation. Data gaps are common in satellite imagery and usually this is a result of data refinement, when obscuring elements have been filtered or masked out. Section 2.2 will give background information of Landsat imagery and the derived product, namely fractional cover (FCover) and how this data can be used to address our defined aims. We justify the suitability in using Landsat imagery and FCover for predicting green vegetation and additionally demonstrate that we yield satisfying prediction accuracy. Section 2.3 explains the importance of spatial modelling such as LULC studies for environmental purposes and identify potential methods. The discussion of the current methodology and review of software will be outlined in Section 2.5 and in Section 2.7 according to the aims explained in the Introduction and in Figure 1.1.

2.1 Big data characteristics

“We are drowning in information but starved for knowledge” said by John Naisbitt in 1982 (Cressie, Shi, & Lang, 2010). This sentence is still valid and gets more important with the beginning of processing large and diverse datasets, so called big data. “Advances in data collection (...) and computerization of many businesses and government transactions have flooded us with data and generated an urgent need for new techniques and tools that can intelligently and automatically assist in transforming this data into useful knowledge” (Banerjee, Gelfand, Finley, & Sang, 2008).

There are many advantages in having access to those large data sets, but there are also challenges in processing these data. One is dealing with the difficult demand to manage,

process, analyse, extract, visualize and publish useful information out of large, diverse heterogeneous and often widely distributed datasets. Many experts are convinced that the era of big data has arrived and new approaches along with it. The expression “Big Data” could be found in the 1990’s in formal literature and it refers not specifically to large datasets only, but massive data collections and consolidations from multiple data sources and even to the techniques to manage and analyse the data. However, not only the volume is one of the challenging properties which define big data, also velocity, veracity and variety will be taken into account for processing large datasets. Data scientists break big data into four dimensions: Volume (scale of data), Velocity (refers to the speed of data processing), Veracity (biases, noise and abnormality in data), Variety (different forms of data). Moreover, (Rubin, 2014) claims that veracity is now considered the biggest challenge which has not been acknowledged so far as important for successful applications of big data. Spatial data like remotely sensed imagery possess the above mentioned four dimensions of big data. We need to investigate new non-traditional approaches to overcome the challenges of dealing with large and complex data. We are only at the beginning of the new so called era of big data and much more can be expected in the next decades (Wu & Kumar, 2009). A meaningful extraction of information out of rich and diverse datasets to create new knowledge is limited and challenging. Therefore, we need new tools to manage and process big data which are beyond the ability of existing software tools to manage and process them within a tolerable elapsed time (Wu & Kumar, 2009) and to gain knowledge that is needed for the future.

2.2 Monitoring land use and land change with Landsat data

Landsat optical imagery has been extensively used for environmental monitoring (Guerchman et al., 2015; Lindquist & D’Annunzio, 2016; Reiche, de Bruin, Hoekman, Verbesselt, & Herold, 2015; Rigge, Smart, Wylie, & Kamp, 2014; Sarker, Alvarez, & Woodley, 2016; Schmidt, Thamm, Menz, & Bénes, 2003). With remote sensing data it is possible to objectively observe and monitor land use and land change (LULC). Landsat imagery is quantised in 8 bit, meaning that there are 256 different grey values for each pixel ranging from black (0 – max absorption) to white (255 – max reflection). Behind each pixel is a numerical integer value showing the individual reflectance of objects on the ground recorded by the sensor. Landsat offers several spectral bands in the electromagnetic spectrum where the reflectance of objects on the earth surface will be recorded. Since different object reflect their properties differently, characteristic spectral curves can be plotted throughout the electromagnetic spectrum and through all spectral channels. The spatial resolution of the Landsat sensor is 30 x 30 m and the reflectance of all recorded objects will be combined into one pixel. This leads to a mixture of spectral information represented in one pixel and is called a mixed pixel, or Mixel (Schmidt et al., 2003). Fractional Cover data is derived from Landsat imagery and provides fractional cover representation

of three ground cover classes across Australia (Scarth, 2012; Scarth, Röder, & Schmidt, 2010). The fractional cover spectral unmixing algorithm breaks the spectral information stored in each pixels up into three parts of fractions represented as percentages. Those are a.) Photosynthetic vegetation - includes leaves and grass, b.) Non-photosynthetic vegetation - includes branches, dry grass, and dead leaf litter c.) Bare surface cover - bare soil or rocks. Further descriptions of how FCover fractions can be derived can be found in (Guerschman et al., 2015; Scarth et al., 2006, 2010; Schmidt, Carter, Stone, & O'Reagain, 2016).

Analysing Landsat time series data can give insight into how knowledge of how the land use has changed in the last years. (Schmidt et al., 2003) showed that the analysis of Landsat data is suitable to describe changes in the LULC pattern for three decades. However this approach has its limitations. Therefore, (Pugh & Waxman, 2006) argues that multi-spectral imagery from earth observation satellites, like Landsat, has been widely used for land cover classification, but the outcomes in form of land use classifications have generally been limited to broad categories. The author (Pugh & Waxman, 2006) further argues that a reliable classification of sub-categories in the monitoring of land cover is in high demand. Nevertheless, there is a big advantage of using remotely sensed Landsat imagery and applied spectroscopy for land use monitoring because the data are quickly available and require no further costs in the acquisition since Landsat opened the archive for the public. Therefore, it can be expected that sequential images of the Landsat satellite time series data can be used as a basis for an assessment and estimation of existing green vegetation. Moreover, using freely available Landsat imagery avoids in-situ observation and measurement. However, there is a trade-off between using Landsat imagery which shows a moderated resolution and in-situ measurements which are expensive, time consuming but offer a high localised accuracy. In conclusion, using Landsat imagery for monitoring LULC is a valuable data source that can be used to address many ecological questions, cover large regions and time periods and is freely available.

The importance of spatial modelling for environmental purposes (Mazzotti, Hughes, & Harvey, 2007) can be demonstrated in revealing short and long term trends that can lead to new knowledge, and a better understanding in complex relationships and dependencies. This new knowledge can be used to revise policies as part of an iterative learning development. Adaptive Management assists in learning of the effectiveness of management decision by monitoring its outcomes (McCarthy & Possingham, 2007). It is used with applications which possess a high degree of complexity since they involve economic, institutional, and ecological linkages across large landscapes with high heterogeneity (B. K. Williams, 2011). One of the challenges of adaptive management can be addressed with remote sensing techniques. Earth observation satellites like Landsat are recording environmental information in a constant time interval for over four decades now. Before management practices can be judged in regards of their effectiveness, a thorough

data gathering of the ecological system needs to be conducted first. (Schmidt et al., 2003) demonstrated on time series data that Landsat imagery is suitable to describe changes in land-cover in Morocco, Africa. In addition, the Australian Collaborative Land Use and Management Program (ACLUMP) highlights the value of monitoring land management over time with imagery to observe ground cover and land management practices in Australia (Forward, 2009). Further, a study in China describes the retrieval of grassland aboveground biomass based on satellite imagery in the time frame from 2006 to 2010 (Jin et al., 2014). Moreover, (Bastin, Denham, Scarth, Sparrow, & Chewings, 2014) demonstrated in a Landsat time series the change of rangeland due to grazing effects in Queensland between 1988 and 2005. In conclusion, it can be said that remotely sensed imagery used for ecological LULC studies can assist in adaptive management practices to ensure an iterative learning development based on monitoring green vegetation in time series.

2.3 Statistical machine learning models

In the era of big data, statistical analysis of data has become an increasingly important tool to quantitatively analyse complexity in spatial data, their relationships and dependencies using statistical machine learning methods. Given this challenging need to interpret data and create new knowledge, we use statistical concepts to learn from data and their properties. All data have intrinsic characteristics (Jain, 2010) that can be described as statistical concepts such as their distribution, data range or extreme values. Consequently, when applying supervised machine learning methods we build a model that learns from the data and use this model to predict events using data that have been excluded in the model building process. Supervised machine learning methods progressively improve performance without being explicit programmed what to do since the analytical model building process is based on the underlying data structure and the data characteristics. (Box, 1979) stated that "All models are wrong but some are useful". There is the trade-off with regards of over- and under-simplifying the complexity of certain local characteristics described by the model. It is critical to keep certain local details, but also it is important to create a model that ensures its transferability/generalizability. Further, it should provides a level of detail and not exclude too much variability just to enhance the goodness of model fit metrics when assessing the model. Generalisability describes to which extent research findings can be applied to other settings and involves drawing broad inferences in quantitative research (Brown & Raymond, 2007). If relevant components are excluded the model is too simple and we don't develop the understanding of local characteristics. Whereas, if there is too much detail the model becomes too complex and lacks transferability and generalizability to different regions and data sets.

There are many suitable statistical machine learning methods that address environmental and ecological complexity. For example, (T. J. Hastie & Tibshirani, 1986) developed a

Generalized Additive Model (GAM) that links a univariate response variable with predictor variables and allows exponential families distribution to unite characteristics of a Generalised Linear Model with additive models. It is considered as a nonparametric extension of Generalised Linear Models and also an extension of classical linear models (Moisen et al., 2006). GAM's have been widely used in ecological applications (Moisen et al., 2006). In a study of (Moisen et al., 2006) GAM has been used to compile step-wise unique models based on environmental covariates predicting tree species in Utah, USA, and he further compared the results achieved by GAM with Stochastic Gradient Boosting (SGB). The model performance was evaluated using independent test data set including specificity, sensitivity, Kappa, and area under the curve (AUC). (Moisen et al., 2006) concluded that SGB had higher values for the majority of species for naïve accuracy, specificity and kappa; while GAMs had higher values for a majority of the species for sensitivity. Alternatively, (Leathwick, Elith, Francis, Hastie, & Taylor, 2006) used GAM and BRT in comparison for analysis of relationships between demersal fish species richness, environmental characteristics and trawl data around New Zealand. BRT are the extension of the functionality of the regression trees by adding a second algorithm called Boosting. Boosting is a machine learning algorithms used for reducing bias and variance in supervised learning (Breiman, 1996). Based on the results he concluded that BRT is a powerful analysis tools, which showed substantially superior predictive performance over generalised additive models. Further, the BRT model had a greater predictive power and explained 6 % more deviance than the GAM model (Leathwick et al., 2006). He used mostly k-fold cross-validation as an evaluation method and demonstrated clearly that BRT's outperformed substantially in comparison with GAM.

We demonstrated that a BRT as a supervised machine learning approach achieves good prediction results and is suitable for spatial and non-spatial data. In addition to BRT there are many other supervised and unsupervised machine learning approaches that perform well in using satellite imagery for prediction purposes such as crop yield (Pantazi, Moshou, Alexandridis, Whetton, & Mouazen, 2016) using the NDVI and understanding variations in the yield based on soil data, drought assessments that causes crop failure (Park, Im, Jang, & Rhee, 2016) using MODIS imagery, quantify aboveground biomass (Dube, Mutanga, Elhadi, & Ismail, 2014) using RapidEye data and classify local forest communities (Li, Im, & Beier, 2013) using Landsat imagery. (Pantazi et al., 2016) used three different Self Organizing Map models, namely, Counter-propagation Artificial Neural Network (CPANN), Supervised Kohonen Network (SKN) and XY-fusion network (XYF) to predict the field variation in wheat yield using the NDVI derived out of GeoEye imagery and concluded that the SKN model demonstrated the best overall performance. In a different study of (Park et al., 2016) the influence of temperature and evapotranspiration in crop failure is investigated by monitoring meteorological and agricultural drought factors in a temporal approach for different climate regions in the USA. The authors used three machine learning approaches, namely Random Forest, Boosted Regression Trees, and Cubist and concluded that the temperature and evapotranspiration showed higher

relative influence for short-term meteorological drought while vegetation-related factors contributed towards long-term effects. (Li et al., 2013) investigated in the identification of forest type change using multi-temporal Landsat data covering a 20-year timeframe using Random Forest, Decision Trees and Support Vector Machine approaches and concluded that Support Vector Machine and Random Forest outperformed Decision Trees. Dube et al. (2014) also used Stochastic Gradient Boosting and Random Forest to predict non-linear intra- and inter- species biomass. The Stochastic Gradient Boosted outperformed Random Forest and the authors emphasised the relevance of stochastic prediction models in predicting aboveground biomass.

2.4 Benefits of using BRT

BRT consist of two algorithms: Decision Trees and Boosting. One of the advantages is, that they are highly customisable to the specific requirement of the application (Natekin & Knoll, 2013). BRT can perform regression (Elith et al., 2008) and classification (De'ath, 2007) and can deal with many data types such as categorical variables for classification (De'ath, 2007; Jafari, Khademi, Finke, Van de Wauw, & Ayoubi, 2014; Rizzo, Martin, & Wohlfahrt, 2014) or numeric variables for regression (De'ath, 2007; Elith & Leathwick, 2017), and the use of different loss functions (De'ath, 2007; De'ath & Fabricius, 2000). BRT is beneficial for processing large spatial data sets for estimation and prediction purposes (Elith & Leathwick, 2017; Leathwick et al., 2006) and is a statistical model that built a series of regression trees, minimise errors through cross validation and avoids over fitting which allows for more flexibility in the selection of environmental variables (Humphries, 2015). Further, BRT can account for non-linearities and interactions between variables (Müller, Leitão, & Sikor, 2013). This makes BRT an appealing method that combine high predictive accuracy where interactions can be quantified and visualised and offers easy ways to interpret and diagnostics of the results (De'ath, 2007; Müller et al., 2013). Moreover, a case study of (Cruse, Liedloff, & Wintle, 2012) demonstrated that BRT can account for spatial autocorrelation. Another feature the diagnostic capabilities of BRT is a relative influence plot, listing all covariates in a descending order showing their contribution as a percentage and their predictive power towards the response variable. Relative influence plots enable a more confident variable selection to create a subset of strong covariates without minimising the predictive capabilities of BRT. This is an important analysis tool since variables that do not contribute significantly can be removed from the modelling process and collinearity can therefore be avoided or minimised (George, 2000). In comparison with other popular machine learning methods such as Artificial Neural Networks (ANN) and SVM those methods don't provide relative influence plots due their multi- or high dimensional analysis. A basic linear separation of results provided by regression analysis such as BRT, is not possible and therefore ANN and SVM lose their ability of visualising dependencies and relationships of data.

In a study of (Rizzo et al., 2014) they assessed the BRT predictions on location specific features and location probabilities using relative influence plots to compare the strength of the covariance with each other and concludes three variables were most influential in successfully predict a special crop type used for bioenergy production. The case study of (Jafari et al., 2014) evaluated the suitability and performance of BRT for soil mapping, where added covariates were investigated with regard of their improvement of prediction results. The authors concluded that by removing strong covariates the prediction accuracy strongly decreased, whereas adding strong covariates contributed importantly and resulted in a better prediction accuracy. A similar study has been conducted by (Humphries, 2015) where the identified top predictor allowed a better understanding in seabird distribution on a long term population monitoring in oceanographic regions. There are many applied studies where BRT showed considerable advantages in showing the differences between cropland based on the influence of topography in agricultural cropland (Müller et al., 2013), identify important regions using long-term time series on variations of seabird population and success of population using gridded spatial data (Humphries, 2015) and pixel-based classification of soil mapping (Jafari et al., 2014). (Leathwick et al., 2006) analysed the relationship between fish richness, trawl characteristics and oceanic environment and concluded that BRT is a powerful tool, with superior predictive performance to GAM. In Chapter 3 we investigated the suitability of BRT on spatial and non-spatial data and we demonstrate the superior performance of BRT in comparison with two other regression models, namely RF and LASSO. BRT outperformed RF and LASSO in the predictive performance and visualisation of results and interpretation of results in general. Another advantage in using BRT is the flexibility of performing a uni- or multivariate analysis and the individual interpretation of their strength as a covariate on the response variable or as a comparison of existing relationships or dependencies amongst each other.

2.5 Boosted Regression Trees

The principles behind a Decision Tree method is to divide the space of input covariates through a binary rule based system. Those splitting rules use operators at the nodes of the tree to split the data in if-else paths along tree branches. The data is split at several nodes and result in most homogeneous groups in the leaves of the tree. At each node, the "error" between the predicted value and the actual values is squared to get a sum of squared errors. At the first iteration, the split point errors across all the variables are compared and the variable yielding the lowest sum of squared errors is chosen as the root node point (J. Friedman, 2006; Tarling, 2009). In a regression tree, the standard deviation (Robinzonov, 2013) is used to make that decision in place of information gain.

The disadvantage in using a regression tree method is the rigid binary partition of the feature space. In order to capture features and their local characteristics better, a more flexible approach is needed. Boosting performs binary splits also referred to as weak

learners/binary splits, and it combines them in an ensemble approach to create one complex prediction rule. The partition of the feature space is based on the covariates. All following partitions focus on the residual errors of the previous steps to iteratively create new trees. In this way observations which are predicted poorly are given a higher weight for the next iteration in creating a new tree. It goes on successively to generate each new set of residuals until all trees have been created. The result from the single steps in creating those trees is summed up to give a successive accumulative model. Boosting is primarily used for reducing bias and variance in supervised learning and it converts weak learner to strong ones. The idea behind boosting is that through combing the predictive power of many weak classifiers, a classifier of arbitrary accuracy can be created (Breiman, 1997). This implies that even the simplest classifiers, with prediction only slightly better than random, could be combined to create an excellent predictor.

BRT belongs to the family of greedy algorithms that aim to find the best choice in an iteration by finding the local optima assuming that this will lead to a global solution. To achieve a more stable convergence towards correct values, we can make smaller steps by only adding a fraction of the result given by each tree. By forcing smaller convergence steps, we slow down the greedy learning rate and this is controlled by the shrinkage hyperparameter (Sancetta et al., 2016). Other important hyperparameters are the total number of trees in the final model (tree complexity), interaction between different nodes along the branch (interaction depth), and the minimum number of training set samples in a node to commence splitting (minimum observations in node). Those four hyperparameters are the most influential ones that prevents the BRT from overfitting, regulates the bias-variance trade-off and ensures a good model generalisation ability to other data sets. A feature of the BRT algorithm is that the performance can be tuned to accommodate specific data structures and characteristics through specification of hyperparameters. For our BRT model, the *carat* package (Kuhn, 2008) was employed to find optimal values for the hyperparameters.

BRT also has limitations, namely overfitting and ensuring the generality capabilities of the model through regularization (Natekin & Knoll, 2013). Empirically, it has been found that using a small value for shrinkage results in impressive improvements in a model's generalisation ability (T. Hastie, Tibshirani, & Friedman, 2009). The drawback of a lower learning rate is that more trees need to be generated, resulting in increased computational time.

BRT, also known as Gradient Boosted Machine (GBM) or Stochastic Gradient Boosting (SGB), is a non-parametric regression technique that combines a Regression Tree with a boosting algorithm (J. Friedman, 2006) and is graphically explained in Figure 2.1. This

extension to the classical regression tree allows greater flexibility and predictive performance in modelling the data.

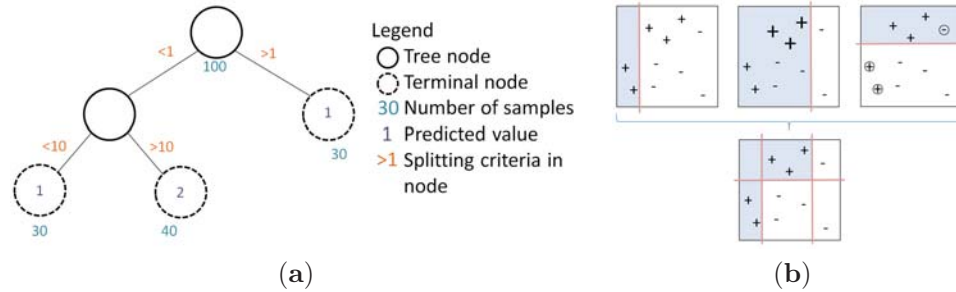


Figure 2.1: Sub-figure (a) shows a graphical demonstration of the hierarchical regression and binary splitting process at the nodes of the BRT and how the observed values will be transported along the tree branches. In sub-figure (b) we demonstrate the ensemble approach of the boosting algorithm as part of the BRT. Binary splits indicated as red straight lines separate the data in grey and white sections and so called weak learners are created as seen in Equation (2.1). The combination of weak learners, to form one strong prediction rule is managed by the boosting algorithm. The BRT method yield a more accurate prediction accuracy through generating flexible boundaries and therefore allowing the identification of small areas of interest. Adapted from (Matteson, 2013).

A regression tree partitions multivariate data with a hierarchy of binary splits that define regions of the covariate space in which the response variable has similar values. The partitions of the feature space are defined by splitting rules in the nodes of the tree, a distance metrics or by information gain. The partition can be illustrated as a tree-like structure, comprising nodes representing the selected factors, branches acting as if-else connectors between the nodes, and leaves representing terminal nodes containing the subsets of responses as depicted in Figure 2.1 (Robinzonov, 2013; Tarling, 2009).

Boosting improved the performance, whereby a sequence of trees is grown, such that in each subsequent tree a greater focus will be applied where observations showed a greater prediction error by giving the observations a higher weight. This results that misclassified or large residual errors in the current iteration will get prioritized in the next iteration. The variance can be captured by growing the tree deeper, meaning that the tree accommodates more segments. The motivation behind boosting is that each tree can be quite shallow (a weak classifier) and thus fast to estimate, but by combining the predictive power of many weak classifiers, a classifier of arbitrary accuracy and precision can be created (Breiman, 1998; Freund & Schapire, 1996; J. H. Friedman, 2002).

Here, we summarise the method, following Friedman (J. Friedman, 2006). Consider a response variable y and a vector of predictor variables \mathbf{x} that are connected via a joint probability distribution $P(\mathbf{x}, y)$. Using a training sample $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ of known values of \mathbf{x} and corresponding values of y , the goal is to find an approximation $F(\mathbf{x})$ to a

function $F^*(\mathbf{x})$ that minimises the expected value of a loss function $\psi(y, F(\mathbf{x}))$. Boosting approximates $F^*(\mathbf{x})$ by an additive expansion. The parameters $\{\mathbf{a}_m\}_0^M$ and the expansion coefficients are jointly fitted to the training data. This is done in a forward stage wise manner. Gradient Boosting (J. Friedman, 2006) approximately solves differentiable loss functions $\psi(y, F(\mathbf{x}))$ with a two step procedure. First, the function $h(\mathbf{x}; \mathbf{a})$ is fit by least squares to the current pseudo-residuals which represent the residuals from the given stage of the tree building.

First, the model will be initialized with the mean of the training set that is defined through $\{y_i, \mathbf{x}_{i1}\}_i^N$. Then we specify the number of trees/iterations shown as m in the for-loop control structure. Friedman (J. Friedman, 2006) added a stochastic element by proposing to draw a random subsample from the full training data set without replacement. This subsample is then used to fit the base learners and compute the model update for the current iteration. The random subsample of size $\tilde{N} < N$ is given by $\{y_{\pi(i)}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}}$. Adding randomness to the algorithm in this way has been shown to improve the performance of Gradient Boosting (J. H. Friedman, 2002). In the last step of the algorithm the current approximation of F_{m-1} is updated in each corresponding region R_{lm} .

Next, the current approximation $F_{m-1}(\mathbf{x})$ is individually updated in all of the corresponding regions

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} \mathbf{1}(\mathbf{x} \in R_{lm}). \quad (2.1)$$

The shrinkage parameter, ν , ranges from 0 to 1 and controls the learning rate γ , so each gradient step is reduced by some factor between 0 and 1 of the learning rate. The value of γ is influenced by the choice of loss function ψ .

Then, given $h(\mathbf{x}; \mathbf{a}_m)$, the optimal value of the coefficient β_m is calculated via

$$\beta_m = \arg \min_{\beta} \sum_{i=1}^N \psi(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a}_m)). \quad (2.2)$$

Thus, at each iteration m , the tree partitions the feature space into L disjoint regions $\{R_{lm}\}_{l=1}^L$ and predicts a constant value, \bar{y}_{lm} , in each region. Gradient Boosting proceeds in this way until the base learner $h(\mathbf{x}; \mathbf{a})$ is an L terminal node regression tree.

The parameters of the estimated tree are the splitting variables and corresponding split points that define the tree, and this defines the corresponding regions $\{R_{lm}\}_1^L$ of the partition at each iteration. These are accomplished in a hierarchical top-down approach using a least squares splitting measure (J. Friedman, 2006). Equation 2.2 can be solved individually within each region, R_{lm} defined by the corresponding terminal node l of the

m th tree. Because the tree predicts a constant value \bar{y}_{lm} within each region, R_{lm} , the solution to 2.2 reduces to a simple location estimate based on the criterion ψ

$$\gamma_{lm} = \arg \min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma). \quad (2.3)$$

The model fit is primarily analysed on the basis of the root mean square error (RMSE). The RMSE is a measure of how well our model performs when using new data and measures the difference between values predicted by a model and the values actually observed that is being modelled on the test dataset. In general, the RMSE is best when it is small, but there is no absolute good or bad threshold.

2.6 Applications of BRT in agriculture and ecology

In a time series study of (Müller et al., 2013) BRT has been successfully used to identify land use change of cropland abandonment in Albania and Romania based on Landsat imagery. They concluded with analysing the spatial data mining pattern of change that cropland abandonment was largely determined by the underlying topography in the two countries. (Rizzo et al., 2014) proved that BRT can spatially map miscanthus, an emerging bioenergy crop in France. The aim of this study was to detect regions of miscanthus based on real spatial distribution data in order to identify areas where miscanthus is present or absent. (Humphries, 2015) investigated in Generalized Boosted Regression modelling to infer important oceanographic regions for seabirds using breeding sooty shearwaters in a long time series in New Zealand. The goal was to identify the population and distribution of Archival Geolocation (GLS) tagged seabirds in important regions around the islands. He used a kernel density tool to delineate important foraging locations for the birds during the breeding season. Where the recordings were dense he assumed that those areas are important. He concluded that it is possible to combine spatial techniques with long-term data sets to infer potential foraging areas using ecological data with ecological niche modelling techniques via BRT. In addition, (Leathwick et al., 2006) used spatio-temporal and spectral information out of Landsat imagery in BRT to identify high concentration of chlorophyll which can be interpreted as primary productivity sites which attracts fish in oceanic waters. The study concluded that BRT provide a powerful analysis tool, which superior predictive performance over generalized additive models (Leathwick et al., 2006). He demonstrates that the BRT is suitable for complex analysis. To sum up, BRT has been successfully used in spatial, spatio-temporal and spectral studies which aim to address specific ecological needs and yields superior results.

In a very recent study of (Pourghasemi & Rahmati, 2018) compared the accuracy of 10 advanced machine learning techniques (Artificial Neural Networks (ANNs), Boosted Regression Tree (BRT), Classification and Regression Trees (CART), Generalized Linear Model (GLM), Generalized Additive Model (GAM), Multivariate Adaptive Regression Splines (MARS), naïve Bayes (NB), Quadratic Discriminant Analysis (QDA), Random Forest (RF), and Support Vector Machines (SVM) for modeling landslide susceptibility and evaluating the importance of environmental variables. The aim of the study was to identify landside prone areas in an semi-arid region of Ghaemshahr Region, Iran in order to control land degradation. They concluded that RF and BRT have the best performances comparison to other machine learning techniques used in this study by using the Receiver Operating Characteristic curve (ROC) method to illustrate the ability to separate classification results. Whist it has been shown that BRT perform well in general studies (Jafari et al., 2014) and (S. E. Wang, 2013) further demonstrate successful applications and high prediction accuracy achieved from BRT by using Landsat imagery in semi-arid regions such as Iran, Australia and China. Surprisingly, spatial predictions using FCover data in Australia using BRT can only be found in a study of (B. Wang et al., 2018) where they used the FCover band "bare soil" for predicting changes in soil organic carbon in semi-arid rangelands in eastern Australia. They concluded that the results of the study are important in Australian rangeland because they provide a statistical basis for producing maps using remotely sensed data and have potential for further use in similar rangeland condition across the globe.

2.7 Review of Software

There are many software products available for working with spatial, non-spatial and raster data formats to perform tasks such as data pre-processing, storage and data management. To achieve our results we used the freely available software R and its extensive libraries offering built-in functions of supervised machine learning algorithms, various geo-spatial tools and geo-spatial drivers that can read and write raster and vector data extensions, and extensive and flexible visualisation options. R is an open source scripting language supported by the R Foundation for Statistical computing and is widely used for data analysis and for developing statistical software. It is available for Linux, Mac (OS X), Windows (R Core Team, 2013; *R Development Core Team*, 2008; Wikipedia, 2018, December 26).

One of the biggest advantages in using R is, that the R environment and its libraries can be extended through user-created software packages public available and published through GitHub or the Comprehensive R Archive Network (CRAN). This allows the user to apply specialised statistical tools, function and algorithms to existing data and also enables to create new R packages to share with the R community. The packages are developed mostly in R, but are also available in more efficient programming languages

such as Java, C, C++ and Fortran (Wikipedia, 2018, December 26). R can be used with the graphical user interface RStudio and the R functionally is accessible through several object orientated and high-level programming such as Python that is predominantly used in the proprietary ArcGIS suit. To combine the power of R and ArcGIS we used the package R-ArcGIS Bridge that serves as an Application Programming Interface (API) to read and write data to and from ArcGIS and R and run scripts within ArcGIS. By combining R and ArcGIS we were able to use the statistical library of R and the extensive geo-spatial data analysis capabilities of the ArcGIS suit. Other software available to analyse and process big spatial data are Hadoop, System for Automated Geoscientific Analyses (SAGA), SAS, SPSS and Stata (Wikipedia, 2018, December 26).

3 Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making

Preamble

The primary purpose of this chapter is to introduce Boosted Regression Trees as our main modelling technique and its suitability for our research aims.

In presenting this paper we contribute to our first aim A1, namely the suitability of BRT of big noisy data showing missingness, different granularities and data characteristics stored in monolithic hardware components and therefore imply the lack of interoperability. Further, we demonstrate that BRT offer a wide range of visualisation possibility that allow for a good interpretability of results. This is especially useful for a wide range of data-driven decision for real word applications and in assisting in a better informed decision making.

This chapter has been prepared as a paper and has been published.

Statement for Authorship

This chapter has been written as a journal article. The authors listed below have certified that:

- (a) They meet the criteria for authorship as they have participated in the conception, execution or interpretation of at least the part of the publication in their field of expertise;
- (b) They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (c) There are no other authors of the publication according to these criteria;
- (d) Potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of the journals of other publications and the head of the responsible academic unit; and
- (e) They agree to the use of the publication in the student's thesis and its publication on the Australian Digital Thesis database consistent with any limitations set by publisher requirements.

The reference for the publication associated with this chapter is; **Student Brigitte Colin**, Other supervisors (2017). Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making. This chapter has been prepared as a paper and has been published.

Contributor	Statement of contribution
Student Brigitte Colin	Conduct the research, develop code for the statistical approach, write the manuscript, make modifications to the manuscript as suggested by coauthors and reviewers.
Signature and date:	
Samuel Clifford	Contributed text and comments on manuscript.
Paul Wu	Contributed text and comments on manuscript.
Samuel Rathmanner	Comments on manuscript.
Kerrie Mengersen	Propose and supervise research, comments on manuscript.

Principal Supervisor Confirmation: I have sighted email or other correspondence for all co-authors confirming their authorship.

Name: K Mengersen Signature: **QUT Verified** Signature Date: _____

Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making

Brigitte Colin¹, Samuel Clifford¹, Paul Wu¹, Samuel Rathmanner¹, Kerrie Mengersen¹

¹ARC Centre of Excellence for Mathematical and Statistical Frontiers, School of Mathematical Sciences QUT, Brisbane, Australia

Email: b.colin@qut.edu.au

How to cite this paper: Colin, B. et al. (2017) Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making, Open Journal of Statistics, 5, *-*.

Received: **** **, ***

Accepted: **** **, ***

Published: **** **, ***

Copyright © 2017 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Abstract

Challenges in Big Data analysis arise due to the way the data are recorded, maintained, processed and stored. We demonstrate that a hierarchical, multivariate, statistical machine learning algorithm, namely Boosted Regression Tree (BRT) can address Big Data challenges to drive decision making. The challenge of this study is lack of interoperability since the data, a collection of GIS shapefiles, remotely sensed imagery, and aggregated and interpolated spatio-temporal information, are stored in monolithic hardware components. For the modelling process, it was necessary to create one common input file. By merging the data sources together, a structured but noisy input file, showing inconsistencies and redundancies, was created. Here, it is shown that BRTs can process different data granularities, heterogeneous data and missingness. In particular, BRTs have the advantage of dealing with missing data by default by allowing a split on whether or not a value is missing as well as what the value is. Most importantly, the BRT offers a wide range of possibilities regarding the interpretation of results and variable selection is automatically performed by considering how frequently a variable is used to define a split in the tree. A comparison with two similar regression models (Random Forests and Least Absolute Shrinkage and Selection Operator, LASSO) show that BRT outperforms these in this instance. BRT can also be a starting point for sophisticated hierarchical modelling in real world scenarios. For example, a single or ensemble approach of BRT could be tested with existing models in order to improve results for a wide range of data-driven decisions and applications.

Keywords

Boosted Regression Trees, remotely sensed data, Big Data modelling approach, missing data.

1. Background

Data are typically stored in various ways and various formats, mostly in monolithic software architectures which do not allow for interoperability. Analysis of data across multiple data sources is thus difficult, since the functionality of the single data sources with respect to input and output, maintenance, data processing, error handling and user interface are all interwoven and act as architecturally separate components. In order to create a basis for analysing the data considered here, it was required to extract the datasets from their original databases and combine them to form a common input file for the modelling process. It was therefore inevitable that this resulted in a data file structure which showed missing data, inconsistencies, duplicates and redundancies.

A case study is presented here to examine land use data sourced from a GIS, direct observations from an agricultural company, and remotely sensed data. The data were extracted from a relational database, Excel spreadsheets, remotely sensed imagery stored as raster data, and vector data from a Geographic Information System (GIS), directly observed and measured data in real-time and interpolated data. By combining these data sources to form one common basis for our analysis, issues of data volume, variety and veracity were encountered. Big Data research clearly deals with issues beyond volume and belongs not only to the ongoing digital revolution, but to the scientific revolution as well. The question posed of Big Data and illustrated in the case study presented here, is whether new knowledge can be extracted from various data sources that haven't been analysed in combination before, and can thus assist in a better and more confident decision making.

2. Introduction

There is an exponential increase in interest in the use of digital data to improve decision making in a range of areas such as human systems, urban environments, agriculture and national security. For example, decisions in the agricultural domain may require information based on vegetation or land use change, estimation of crops or biomass, distribution of native or exotic species, livestock or weed assessment and so on. One source of digital data that has generated intense interest over the past decades is remotely sensed imagery. These data are available from a wide range of sources, ranging from satellites to drones, and have been used for a very wide range of environmental applications [1–8].

The availability and resolution of these data, combined with improved computer storage and data management facilities, have greatly increased the opportunity for mathematicians and statisticians to utilise this information in their models and analyses. The challenge in linking remotely sensed data to decision-making is that there are multiple steps in the process. Here, we focus on an exemplar real-world problem in the livestock industry: deciding on the allocation of animals to different paddocks and potentially different grazing properties based on the predicted availability of grass over the year. This problem arose in the context of collaboration between statisticians at the Queensland University of Technology and a large livestock organisation in Australia. The specific aim of the project was to develop an ensemble of mod-

els to predict the carrying capacity, that is, the number of animals that can be sustained on a paddock. In order to achieve this goal we utilised remote sensing data and supporting information about climate and paddock characteristics. Further, it was important to present the results in a form that is useful for the agricultural decision makers.

Difficult or challenging decisions demand a thorough consideration and even then they imply uncertainty, complexity and different levels of risk. Making the right decisions at the right time can lead to success, increase of profit or minimisation of risk. It is thus important that thoughtful considerations are put into each decision. Figure 1 demonstrates the workflow following a Big Data approach for our case study. Here, we use structured but heterogeneous data sources that showed characteristics like missing data, noise and redundancies. In the Appendix in Figure 6 we show a plot that demonstrates the data structure and missing values. All the data sources were used to create a BRT model via an ensemble approach. The resulting model and its output serves as a foundation for a better decision making. The steps involved in the process are depicted in Figure 1. Due to commercial confidentiality concerns, the final results of the modelling workflow are not presented here.

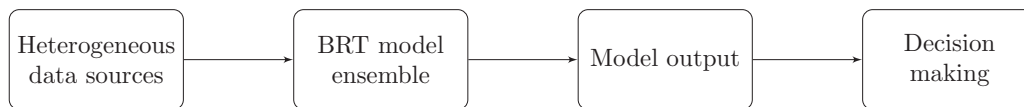


Figure 1: Modelling process for case study.

In this article we focus on one component of the ensemble modelling approach employed in the project, namely the use of BRT to estimate so-called animal equivalents per paddock. Since calves, cows and bulls of different ages consume different amounts of grass, these animals are standardised to a reference animal which can then be used as a common response variable in the analysis. An interesting conundrum is that one of the major inputs into such a model is the amount of grass, or more generally the biomass, in a paddock. This can potentially be estimated directly from remote sensing, but is confounded by the fact that animals are on the paddock eating the very thing that is being measured by the sensor. Moreover, the decision maker may be interested in the biomass estimates themselves, either directly via the remotely sensed measurements or indirectly via the animal equivalents based on animal weight and metabolic formula.

A BRT is a popular statistical and machine learning approach that has not yet seen much application in the analysis of remotely sensed data. Indeed, although they were first defined two decades ago, BRTs have only recently been extended to deal with the types of features that are characteristic of remotely sensed data, in particular its spatial and temporal dynamics. Most of the activity around the use of BRT for agricultural and environmental applications does not appear in the mainstream mathematical and statistical literature.

2.1. Case Study

The study area is located in the Northern Territory, Australia. The main climate zone is identified as grassland with hot dry summers and mild winters [9]. It is a heterogeneous region with a complex topography and land cover and type of grassland. Identification, differentiation and quantitative estimation of biomass is of primary interest in this case study. A range of data from different sources was required for this problem. In this section, we describe the information derived from Landsat imagery and comment briefly on other data. The reflectance recorded by the Landsat sensor is stored as an 8 bit value, resulting in a scale of 256 different grey values ranging from black (0 – max absorption) to white (255 – max reflection). The electronically recorded data appear as an array of numbers in digital format. In addition to the 8 bit quantisation, Landsat offers several spectral bands in the electromagnetic spectrum in which each individual pixel shows different values across different bands. This means that each pixel has a different dimension and therefore will be represented differently in each spectral band. Raster data are becoming increasingly common and increasingly large in volume, although it is possible to reduce file size with compression functions.

There is a strong advantage in using remotely sensed Landsat imagery and applied spectroscopy for these types of analyses because the data are freely available, the imagery covers a wide geographical range, and it avoids expensive, extensive and often impractical in-situ measurement. However, the trade-off is in resolution: in-situ measurements provide highly localised accuracy whereas a pixel in a Landsat image covers an area of $30\text{ m} \times 30\text{ m}$. It is noted that other satellites are now able to provide higher resolution, but these are not yet freely available for the areas of interest in this case study.

Estimation of biomass using satellite data is of ongoing global interest. Grass biomass estimation is challenging since the phenological growing cycle of naturally existing grass is a dynamic process influenced by many complex parameters, including grass type, soil, climate, topography and land use. With the spectral information of remotely sensed imagery it is possible to detect green vegetation, which is driven by the photosynthetic biochemical process of grass biomass. However, since raster imagery is only a two dimensional representation of the land cover it is difficult to derive the quantity of the vertical grass biomass directly.

Fractional cover [10] data are often available as derived products; for example Geoscience Australia (GA) who provides an Australian Reflectance Grid 25 (ARG25) product which gives a 25m scale fractional cover representation of underlying vegetation across Australia or Tern - Auscover in 30m resolution of Landsat 5 and 7 covering the temporal extent from 2000 – 2011. Fractional cover unmixing algorithms use the spectral reflectance of a Landsat scene for a pixel to break it into three fractions represented as percentage values. These are photosynthetic vegetation (includes leaves and grass), non-photosynthetic vegetation (includes branches, dry grass, and dead leaf litter) and bare surface cover (bare soil or rocks) [11].

In addition to fractional cover Vegetation Indices (VI) are commonly

used to extract meaningful information out of the imagery through image analysis techniques. To calculate VIs it is common to apply arithmetical methods in order to create additional artificial channels using existing spectral bands of the imagery. Other related data were also available to support the analyses. For example, SILO (Scientific Information for Land Owners) is a database of historical climate records for Australia. SILO provides daily datasets for a range of climate variables and in formats suitable for a variety of applications. In addition, SILO datasets are constructed from observational records provided by the Bureau of Meteorology (BOM). As another example, the AussieGRASS spatial framework includes inputs of key climate variables (rainfall, evaporation, temperature, vapour pressure and solar radiation), soil and pasture types, tree and shrub cover, domestic livestock and other herbivore numbers. The derived results of AussieGRASS data are spatially interpolated to construct gridded datasets on a regular grid (approximately $5\text{ km} \times 5\text{ km}$) across Australia [12, 13].

2.2. Data-related challenges

The analysis of relationships in ecological data sets is not trivial [14]. In addition to the complexity of the processes being modelled, there is the challenge of dealing with data dimensionality since it is often necessary to combine various data sources. Moreover, the scale of spatial data needs to be considered when there are differing granularities of spatial and temporal data. For example, SILO rainfall data are reported at a $5\text{ km} \times 5\text{ km}$ grid, whereas a Landsat pixel covers an area of $30\text{ m} \times 25\text{ m}$. The SILO data are stored in a tabular data base format and the single measurement points to record the precipitation independently from each other. In contrast, the derived VI cover a whole Landsat scene of $185\text{ km} \times 185\text{ km}$ and are highly correlated. All our environmental data have been provided from the Department of Science, Information Technology and Innovation (DSITI). In addition to the environmental data we used operational data provided by a commercial entity under a confidential agreement.

Another challenging characteristic of remotely sensed data is missing information. There are two major considerations in dealing with this issue. The first is dealing with the missing values. Common options are to filter them out [15, 16], interpolate them or increase the spatial aggregation. There are advantages and disadvantages to each of these approaches in terms of computational resources, inferential capability, and precision and bias of the resultant estimates [17]. The second consideration is whether to undertake the chosen method as part of the pre-processing or post-processing steps.

For our case study we performed a number of pre-processing steps to prepare our data for the modelling process, namely data aggregation and data reduction for our predictor variables, as well as calculation of the response variable. Instead of working with single pixel values we reduced the volume of data by deriving descriptive statistics from Landsat, MODIS and SILO data, thereby obtaining paddock specific means, medians, first quartile, third quartile, variance and Shannon Entropy. With respect to our response variable, we aggregated real-time measurements to a monthly mean. In the next step we created a test

and a training data set by partitioning the data to 20% and 80% respectively. The training set was used to estimate the model parameters. The test set was used for model performance evaluation on unseen data.

3. Boosted Regression Trees

Boosted Regression Trees (BRT), also known as Gradient Boosted Machine (GBM) or Stochastic Gradient Boosting (SGB), are non-parametric regression techniques that combine a regression tree with a boosting algorithm [18]. This extension to the classical regression tree allows greater flexibility and predictive performance in modelling the data. The implementation of these methods used in this study can be found in the `gbm` R package.

A regression tree partitions the data with a hierarchy of binary splits that define regions of the covariate space in which the response variable has similar values. These splits are defined by rules, distance metrics or information gain. The choice of variables and the value at which the split point occurs is determined in a recursive manner at each stage of the tree construction. The segmentation can be depicted as a tree-like structure, comprising nodes representing the selected factors, branches acting as if-else connectors between the nodes, and leaves representing terminal nodes containing the subsets of responses [19, 16, 20].

Boosting improves the performance of a simple base-learner by reweighting observations that were misclassified or had large residual errors in the previous iteration. The deeper we grow the tree, the more segments we can accommodate and thus more variance can be explained. This results in higher model complexity and therefore higher risk of overfitting the model to the data.

The motivation behind Boosting is that each tree can be quite shallow (a weak classifier) and thus fast to estimate, but by combining the predictive power of many weak classifiers, a classifier of arbitrary accuracy and precision can be created [21–23].

3.1. Gradient Boosting

In this section we give a brief summary of the method, following Friedman [18]. This supervised machine learning approach deals with a response variables y and a vector of predictor variables \mathbf{x} that are connected via a joint probability distribution $P(\mathbf{x}, y)$. Using a training sample $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ of known values of \mathbf{x} and corresponding values of y , the goal is to find an approximation $F(\mathbf{x})$ to a function $F^*(\mathbf{x})$ that minimises the expected value of a loss function $\psi(y, F(\mathbf{x}))$, i.e.

$$F^*(\mathbf{x}) = \arg \min_{F(\mathbf{x})} E_{y, \mathbf{x}} \psi(y, F(\mathbf{x})). \quad (1)$$

Boosting approximates $F^*(\mathbf{x})$ by an “additive” expansion in the form of

$$F(\mathbf{x}) = \sum_{m=0}^M \beta_m h(\mathbf{x}; \mathbf{a}_m), \quad (2)$$

where the functions $h(\mathbf{x}; \mathbf{a})$ are generally simple functions of \mathbf{x} with parameters $\mathbf{a} = \{a_1, a_2, \dots\}$. The parameters $\{\mathbf{a}_m\}_0^M$ and the expansion coefficients $\{\beta_m\}_0^M$ are jointly fit to the training data. This is done in a forward stage wise manner. Gradient Boosting [18] approximately solves differentiable loss functions $\psi(y, F(\mathbf{x}))$ with a two step procedure. First, the function $h(\mathbf{x}; \mathbf{a})$ is fit by least squares to the current “pseudo”-residuals

$$\tilde{y}_{im} = - \left[\frac{\partial \psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})} \quad (3)$$

which represent the residuals from the given stage of the tree building.

Then, given $h(\mathbf{x}; \mathbf{a}_m)$, the optimal value of the coefficient β_m is calculated via

$$\beta_m = \arg \min_{\beta} \sum_{i=1}^N \psi(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a}_m)). \quad (4)$$

Gradient Tree Boosting performs this with a base learner $h(\mathbf{x}; \mathbf{a})$ of an L terminal node regression tree. A regression tree partitions the feature space into L disjoint regions $\{R_{lm}\}_{l=1}^L$ and predicts a separate constant value at each iteration m .

$$h(\mathbf{x}; \{R_{lm}\}_1^L) = \sum_{l=1}^L \bar{y}_{lm} 1(\mathbf{x} \in R_{lm}). \quad (5)$$

The parameters of the base learner are the splitting variables and corresponding split points that define the tree, and this defines the corresponding regions $\{R_{lm}\}_1^L$ of the partition at each iteration. These are accomplished in a top-down “best-first” approach using a least squares splitting measure [18]. Equation 4 can be solved individually within each region R_{lm} defined by the corresponding terminal node l of the m th tree. Because the tree in Equation 5 predicts a constant value \bar{y}_{lm} within each region R_{lm} , the solution to 4 reduces to a simple location estimate based on the criterion ψ

$$\gamma_{lm} = \arg \min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma). \quad (6)$$

Next, the current approximation $F_{m-1}(\mathbf{x})$ is individually updated in all of the corresponding regions

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} 1(\mathbf{x} \in R_{lm}). \quad (7)$$

Friedman [18] added a stochastic element to the above boosting algorithm by proposing to draw a random subsample from the full training data set without replacement. This subsample is then used to fit the base learner and compute the model update for the current iteration. By adding randomness to the algorithm the performance of gradient boosting was improved and this resulted in the stochastic Gradient Boosting machine (GBM) [23]. The Stochastic Gradient Boosting algorithm is summarised as pseudo code below [15, 23]. The input training data is defined through $\{y_i, \mathbf{x}_{i1}\}_i^N$ and $\{\pi(i)\}_i^N$ is the random permutation of the integers $1, \dots, N$. The random subsample of size $\tilde{N} < N$ is given by $\{y_{\pi(i)}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}}$.

Algorithm 1 Stochastic Gradient Boosting algorithm

Training data $\{y_i, \mathbf{x}_{i1}\}_i^N$

Initialization

$$F_0(\mathbf{x}) = \arg \min_{\gamma} \sum_{i=1}^N \psi(y_i, \gamma)$$

for $m = 1$ to M **do**

$$\{\pi(i)\}_1^N = \text{randperm} \{i\}_1^N$$

Compute pseudo-residuals

$$\tilde{y}_{\pi(i)m} = - \left[\frac{\partial \psi(y_{\pi(i)}, F(\mathbf{x}_{\pi(i)}))}{\partial F(\mathbf{x}_{\pi(i)})} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, \tilde{N}$$

Fit a base learner to pseudo-residuals

$$\{R_{lm}\}_1^L = L \text{-terminal node tree} \left(\{\tilde{y}_{\pi(i)m}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}} \right)$$

Compute multiplier γ_{lm} by solving optimization problem

$$\gamma_{lm} = \arg \min_{\gamma} \sum_{\mathbf{x}_{\pi(i)} \in R_{lm}} \psi(y_{\pi(i)}, F_{m-1}(\mathbf{x}_{\pi(i)}) + \gamma)$$

Update the model

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} \mathbf{1}(\mathbf{x} \in R_{lm})$$

4. Results

The data were presented as a set $\{(x_i, y_i) \mid 0 \leq i < n_{\text{samples}}\}$ with feature vector $\mathbf{x}_i \in \mathbb{R}^{n_{\text{features}}}$, and the response $y_i \in \mathbb{R}$. All the data we used for our case study were combined into a structured comma-separated values (CSV) file that consisted of 209 observations and 141 covariates. The machine friendly notation of our covariates are generated in the following manner. There are in total 5 different components for creating the covariate names. The first shows whether the calculated summary statistics are for monthly values of EOLW/D = end of last wet/dry, or WS = wet season; these are then followed by whether it is an aggregated mean, minimum or maximum monthly values, followed by the nature of the descriptive statistic: first quartile, median, mean, third quartile, variance and Shannon Entropy; next comes the name of data source (e.g. rain = SILO data), and lastly the corresponding area in proximity to water (3km, 5km, 99km = whole paddock). The covariate name of paha.99km/5km stores values for the whole paddock area measured in hectare and the proximity of water e.g. 5km radius or 99km for the whole extent of the paddock. As described in section 2.2, the data set was partitioned by treating 80% as training data and the remaining 20% as test data, resulting in 167 training and 42 test observations.

The computational environment was the R statistical modelling software version 3.3.3 [24] running inside Windows 7 SP1 (64-bit) on a 2.60 GHz Intel i7 CPU with 16GB of RAM. All of the results and illustrations were created in the R programming language. The GBM model implementations for this article were taken from the gbm packages. Table 1 show the distribution of the response variable and the most influential covariates. Please see Figure 3 as a further reference in

regards of their individual contribution in the splitting process.

Table 1: Distribution of the response variable and key predictors. Predictor names are described in text.

Covariates	Min	Median	Mean	Max	Std Dev.
Response variable	8.33	7323.89	11 830.00	87 549.92	13 612.75
1st: paha.99km	310.30	11 400.00	12 670.00	43 710.00	10 856.50
2nd: paha.5km	310.30	7347.00	8569.00	28 200.00	7097.22
3rd: EOLW.q3.abrad.3km	34.56	235.40	235.10	374.80	94.81
4th: EOLD.mean.lgcg.99km	0.00	0.05	0.06	0.33	0.04
5th: WS.max.var.rain.99km	0.00	16.69	33.99	412.10	50.45

One way of showing the complex relationships of the joint probability and contribution of each covariate in describing the response is through a relative influence plot. Relative influence measures are calculated by averaging the number of times a covariate is used for splitting, weighted by the squared improvement to the model as the result of each split. It is then scaled so the values sum to 100.

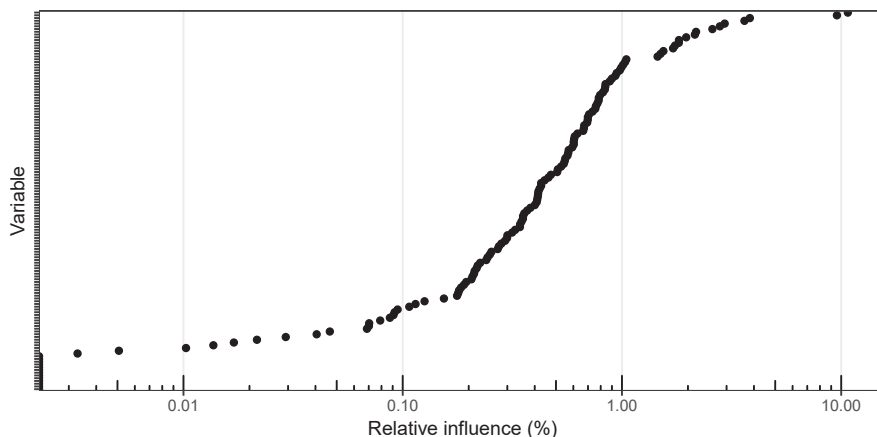


Figure 2: Relative influence plots of all 141 covariates showing their contribution in the splitting process. The horizontal axis indicates the frequency of the contribution with the maximum of 10.8%.

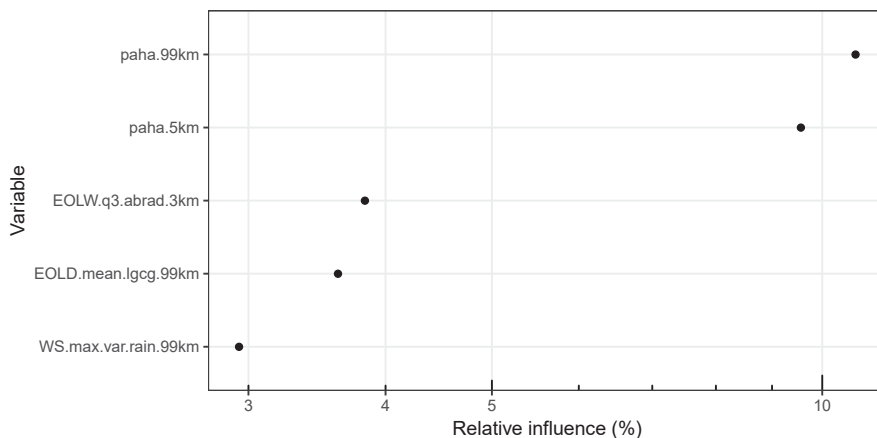


Figure 3: Subset of a relative influence plots of covariates with a contribution greater than 2.9%. (log scale)

In Figure 2 we present a relative influence plot for all of the available variables. The relative influence of the 141 variables varies considerably, with some never contributing (0%) and only 20 variables having relative influence greater than 2.9% as depicted in Figure 3. The two variables that contribute the most are paha.99km at 10.8%, followed by paha.5km with 9.56%. The third strongest variable is EOLW.q3.abrad.3km which contributes with only %3.83. Figure 3 shows the top five contributors on a log scale plot.

Regularisation methods are used to constrain the fitting procedure so that it balances model fit and predictive performance [15]. Regularisation is particularly important for BRT because its sequential model fitting allows trees to be added until the data are completely overfitted [25]. As discussed in section 3, introducing some randomness into a boosted model usually improves accuracy and speed and reduces overfitting [23].

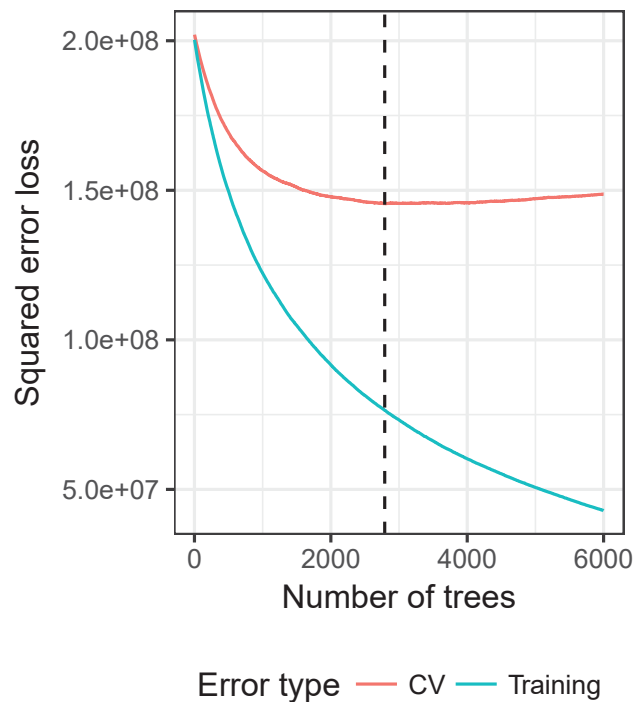


Figure 4: Squared error loss for the training (blue) and test (red) data as the number of trees in the ensemble increases to a maximum of 6000. The optimal tree size (2784) is shown with the dashed black line.

Figure 4 describes the effect of regularisation on the squared error loss. The blue line is the error in the training data, the red line in the test data. The vertical dashed line indicates the optimal number of iterations/trees provided by the gbm model where the test data reaches its minimum, here at 2784 trees. After reaching the minimum, the graph of the squared error loss starts to increase again. This change of direction indicates the start of the model overfitting the training data and therefore poorly explaining the variation seen in the test data. The bias-variance trade-off goal is to find the optimal number of trees where the bias and the variance are balanced and the error is minimised, since both under- and overfit-

ting will have a negative effect on the predictive performance of the model.

Histograms of the residuals for the test and training sets are shown in Figure 5. In comparison to the training data the test data does not have multiple peaks – which often indicate that important variables are not yet accounted for – but there are some large positive outliers in the training data, beyond 50000.

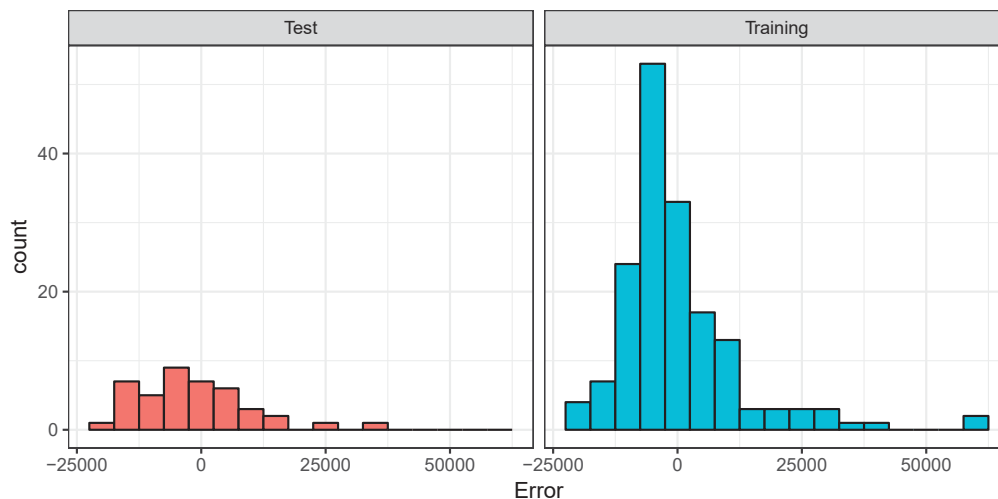


Figure 5: Histogram of residuals in the test and training sets at the optimal tree size.

Table 2 shows the results of the comparison of BRT and other methods. It is seen that the BRT performed best in fitting the data according to the RMSE.

Table 2: Overall model average prediction performance, based on 500 cross-validations.

Method	RMSE
Random Forest	0.48
BRT	0.38
LASSO	0.84

One of the biggest advantages in using a BRT is that it can handle missing values in the predictors by default. As part of the model diagnostics, we can plot how the data have been split, to which node they have been assigned, and the reduction in error for this single iteration/tree. If the tree is challenged with data that are missing a variable, the split is decided based on a surrogate variable, typically one that has a high correlation with non-missing observations.

Table 3: Summary of gbm single tree prediction in pretty.gbm.tree.

ID	SplitVar	SplitCodePred	Left	Right	Miss.	Err.Red	Wt	Prediction
0	84	$3.0117 \times 10^{+2}$	1	2	3	29.72521	466	-1.9659×10^{-5}
1	-1	1.8441×10^{-3}	-1	-1	-1	0.00000	6	1.8441×10^{-3}
2	-1	-1.5669×10^{-4}	-1	-1	-1	0.00000	274	-1.5669×10^{-4}
3	437	8.8800×10^{-1}	4	5	6	31.31934	186	1.2208×10^{-4}
4	-1	7.7070×10^{-5}	-1	-1	-1	0.00000	116	7.7070×10^{-5}
5	-1	3.3260×10^{-3}	-1	-1	-1	0.00000	3	3.3260×10^{-3}

The R function `pretty.gbm.tree()` returns a data frame in which each row corresponds to a node in the tree (Table 3). Here, the root node (indicated by the row number 0) is split by the 84th SplitVar (splitting variable). Since the numbering starts with 0 the split variable is the 85th column in the training set of our case study. Rows in the table with a SplitVar of -1 are terminal nodes. A SplitCodePred value of 301.171 denotes that all points less than 301.17 were allocated to the left node 1 (and hence all points greater than 301.17 were allocated to the right node 2). All points that had a missing value in this column were assigned to the missing node 3. If the node is a terminal node then this is the prediction. The error reduction (29.73) indicates the reduction in the loss function as a result of splitting this node and there were 466 weights (weights will be on each node) in the root node. The weight indicates the total weight of observations in the node. The last column prediction of -0.000019659 denotes the value assigned to all values at this node before the point was split. The prediction column refers to individual trees and they are fit to predict the gradient of the loss function evaluated in the current prediction and the response. This is the gradient part of Gradient Boosting.

5. Discussion

In this case study we demonstrated that BRTs are able to address Big Data challenges, produce satisfying results and can deal with missing values by default. In addition, we obtained in-depth knowledge of the diverse and heterogeneous data sources used in this study, and identified key covariates that were most influential in describing the response variable. Further, descriptive statistics has been used to quantitatively describe our data and basic features of it by providing summaries that enables us to present our results in a meaningful way and therefore allowed for a simpler interpretation. The histograms of training and test data showed us the underlying frequency distribution of our continuous data. In this case both histograms are left skewed and demonstrate that the majority of data can be found on the left hand side. Because histograms use bins to display data it is not possible to see exactly what the specific values are for the minimum and maximum. However, we can see an approximation of the range of values, see how spread out the data are and that there are not outliers that we need to take care of. One of the biggest disadvantages of BRT is that they are prone to overfit the data thus appropriate settings for the hyperparameters need to be set in order to control the model building process. It is therefore advisable to tune the model hyperparameters as part of a pre-processing step in an iterative manner prior to performing the final mod-

elling.

There are many features of BRT that are advantageous for the problem considered here. In addition to computational speed and accuracy of estimation, they can describe complex nonlinearities and interactions between variables, accommodate missing data, include different types of input variables without the need for transformations, perform well in high-dimensional problems, and allow for different loss functions such as accurate identification of small areas of interest. Moreover, they can be visualised and interpreted easily, thus facilitating the translation of the analytic results to decision makers [18]. BRT have also been compared favourably with other flexible regression approaches such as generalised additive models [14]. An example of BRT models helping in developing an understanding of missingness structure in the data is given by [26]. In this study Tierney [26] concluded that more knowledge was gained about the origins of the data and the data collection process, as well as the handling of missing values for future analysis. In another study [14], the author took a different approach to deal with missing values by taking summary values such as the mean over grouped data.

There are several challenges in using BRT for this case study. First, the volume of one single satellite imagery is quite high even without aggregating or combining them in a dense time series. One Landsat satellite scene covers $185\text{ km} \times 185\text{ km}$ of land and has a file size of about 300 MB. The temporal resolution of Landsat is on average 16 days; thus, in one year there are 22 scenes of the same area to computationally process, analyse and store, a data volume of about 6.6 GB. Examination of several years of satellite imagery yields in enormous geotemporal datasets. Given these specifications, a substantive challenge is the storage, processing and management of massive volumes of raster data information. This challenge is exacerbated when the other input variables are also considered, especially since these are of different data formats, sources, structure and spatial granularities. In order to decrease the volume we calculated descriptive statistics based on individual paddock information instead of using pixel information for our analysis.

The second challenge is determining the geographic area to include in the statistical models. The region of interest is spread over multiple stations, with multiple paddocks per station. However, not all of the land in a paddock is grazable. Jansen [27] investigated the quantification of livestock effects on the scalable, season specific metric of Landsat imagery and biomass identification and development of a model assessing spatial relationships between spectral indices and ruminants over a growing season. The focus was on finding significant correlations between existing biomass, vegetation metrics and management practices to quantify changes in vegetation due to grazing. Changes can be caused not only through overgrazing and loss but also due to changes in phenology caused by climate variability and also availability of water. The spatial distribution of animal impacts becomes organised along a utilisation gradient termed a piosphere [28]. Moreover, since animals need access to water, concentric rings can be calculated based on the distance from naturally occurring water points in the paddocks. In the case study these were of order 3km, 5km and the size of the whole paddock. The area around those water locations is then deemed to be the available foraging area. In addition to the concentric rings there are also natu-

ral water streams which attract the animals and provide biomass along those linear features. So-called linear buffer zones can be calculated along the streams to indicate grazing areas nearby the water, like the concentric rings around the water points. The quantity and quality of biomass types can be extracted through the spectral values of the Landsat pixels and with additional spatial data, in particular fractional cover which identifies three categories of ground cover percentage (photosynthetic vegetation, non-photosynthetic vegetation and bare soil).

The third challenge is incorporating spatial information with disjoint geographic areas (agricultural properties or stations), each of which comprises regions (paddocks) of varying sizes. In the case study, the provided information was typically in the form of summary values per paddock per month. Seasonal (wet and dry) indicators were also used to help quantify the biomass [29] and define the spatial extent of the area due to varying rainfall. The beginning of the dry season is a critical time stamp in terms of predicting the amount of grass that will be available during the dry season and the corresponding decision regarding the number of animals to be placed in paddocks to avoid the negative impact of over- or under grazing.

There is a large literature on the predictive, methodological and computational properties of decision trees, including the Random Forest (RF) and Boosted Regression Tree (BRT) models used in this paper. The predictive accuracy of these methods has been investigated both theoretically [30–33] and in various applications [34]. The latter authors also compared modelling approaches considered in this paper in the analysis of a large epidemiological dataset and concluded that RF, BRT and LASSO outperformed the conventional logistic regression framework. Methodologically, decision tree approaches belong to the family of greedy algorithms and select variables in a forward selection manner. Both of these features strongly influence the convergence speed and computational time [18, 23]. The computational time is also influenced by the choice of model parameters such as the learning rate and tree complexity [25]. For example, while a smaller shrinkage parameter slows down the learning rate and results in better predictive performance, the trade-off is a larger number of iterations in order to converge to a local minimum and therefore a longer computational time. The total running time also depends on the choice of loss function, regularisation method and the measure of convergence [31]. Empirical comparisons of the running time of different tree methods such as RF and BRT have also been published [35].

This article has focused on the use of a modern statistical machine learning technique, namely Boosted Regression Trees, to address a challenging real world problem in industry. We presented and demonstrated the efficiency of BRT for addressing Big Data properties with environmental data, specifically remotely sensed data for decision making. There are, of course, other methods that could be used for this type of problem. An appealing alternative that also deals with big, noisy and spatial data is the Bayesian additive regression model [36], a Bayesian sum-of-tree model that generates samples from a posterior. Further, a sum-of-trees model is an additive model with multivariate components. Compared to generalized additive models based on sums of low dimensional smoothers [37, 38], these multivariate components can more naturally incorporate interac-

tion effects. This approach enables full posterior inference including point and interval estimates of the unknown regression function as well as the marginal effects of potential predictors. Gathering large and diverse environmental data is essential in this field and analysing those covariates is challenging. Big data has notable effects on predictive analytic, knowledge extraction and interpretation tools [39] and appropriate methods need to be applied in order to gain new knowledge of data-driven discoveries that assist in decision making.

Acknowledgement

This research was undertaken through the Australian Cooperative Research Centre for Spatial Information with the Project number P4.101. We thank our colleagues for their insight and expertise, particularly Dr. Michael Schmidt (Department of Science, Information Technology and Innovation) for providing key publications of the Wambiana Grazing Trial and helpful discussions.

Appendix

As part of the review process, we added all requested modifications to the Appendix since the paper was published before I submitted my work for external revision.

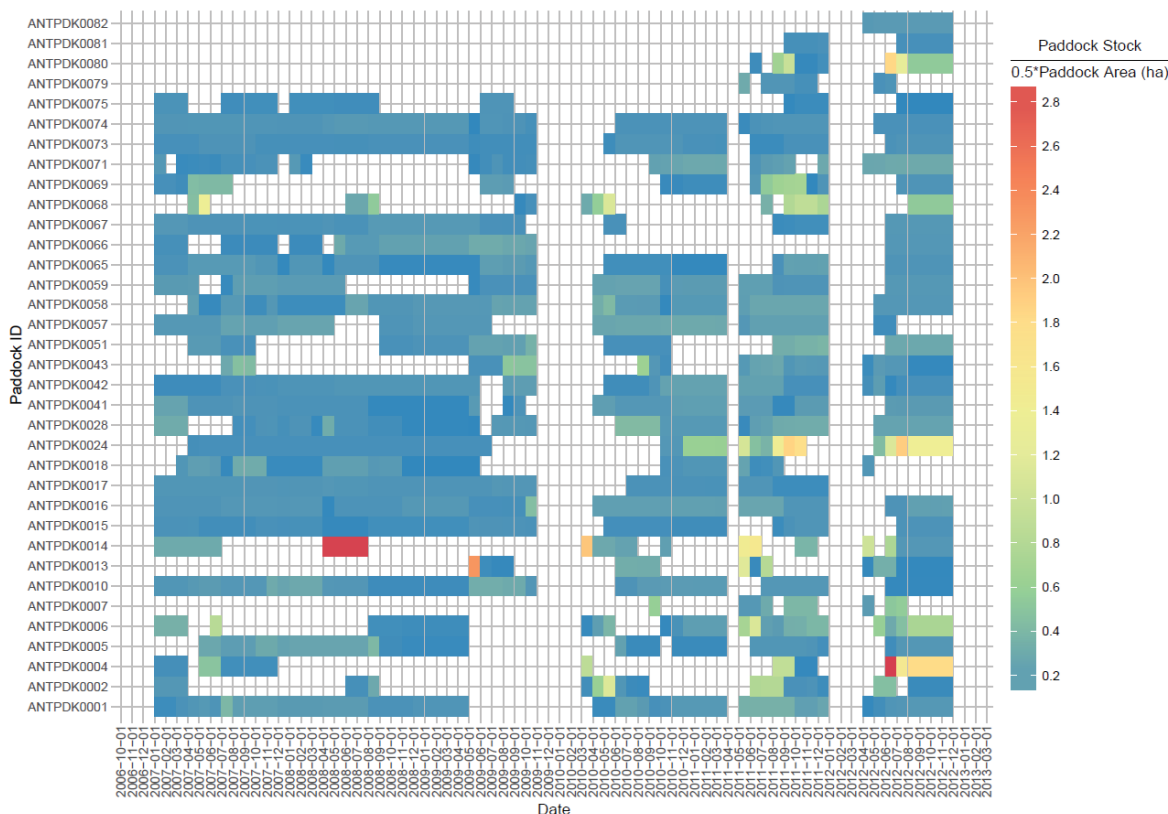


Figure 6: Missing data structure.

5.1. Implementation

At first we split the data into training and test datasets using a random partition that assigns 80% of the data in a training set and the remaining 20% to a test set. In the next step we specified the hyperparameters for BRT modelling. Typical hyperparameters include the

- the total number of trees in the final model, where each tree represents an iteration
- interaction between the nodes along the tree branches
- shrinkage rate of how quickly the algorithm learn and reaches its local minimum
- minimum number of training set samples in a node to commence splitting.

The outcome of the tuning process for the model was a recommendation of number of trees = 2500, interaction depth = 3, shrinkage = 0.01, and minimum observations in node = 10. Those hyperparameters were then used to estimate the coefficients using the training data. The prediction results are based on the test data set and goodness of model fit was determined by the RMSE. Cross-Validation (CV) methods were used for the tuning process to help identify the hyperparameters and to restrict the number of iterations to avoid overfitting when the local minimum has been reached. For resampling the CV method was applied.

References

- [1] Schmidt, M. and Thamm, H.P. and Menz, G. (2003) Long term vegetation change detection in an arid environment using Landsat data, *Geoinformation for European-wide integration*, Millpress, Rotterdam
- [2] Marsett, R.C. and Qi, J. and Heilman, P. and Biedenbender, S. and Watson, M.C. and Amer, S. and Weltz, M. and Goodrich, D. and Marsett, R.C. (2006) Remote sensing for grassland management in the arid southwest, *Rangeland Ecology & Management*, **59**, 530–540. <http://doi.org/10.2111/05-201R.1>
- [3] Huete, A. and Ponce-Campos, G. and Zhang, Y. and Restrepo-Coupe, N. and Ma, X. and Moran, M.S. (2015) Land resources monitoring, modeling, and mapping with remote sensing, monitoring photosynthesis from space, CRC Press.
- [4] Jafari, A. and Khademi, H. and Finke, P.A. and Van de Wauw, J. and Ayoubi, S. (2014) Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran, *Geoderma*, **232–234**, 148–163. <http://doi.org/10.1016/j.geoderma.2014.04.029>
- [5] Anderson, M.C. and Allen, R.G. and Morse, A. and Kustas, W.P. (2012) Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources, *Remote Sensing of Environment*, **122**, 50–65. <http://doi.org/10.1016/j.rse.2011.08.025>
- [6] Washington-Allen, R.A. and Van Niel, T.G. and Ramsey, R.D. and West, N.E. (2004) Remote sensing based phosphorus analysis, *GIScience and Remote Sensing*, **41**, 136–154. <http://doi:10.2747/1548-1603.41.2.136>
- [7] Stohlgren, T.J. and Ma, P. and Kumar, S. and Rocca, M. and Morisette, J.T. and Jarnevich, C.S. and Benson, N. (2010) Ensemble habitat mapping of invasive plant species, *Risk Analysis*, **30**, 224–235. <http://doi.org/10.1111/j.1539-6924.2009.01343.x>

-
- [8] Sarker, C. and Mejias Alvarez, L. and Woodley, A. (2016) Integrating recursive Bayesian estimation with support vector machine to map probability of flooding from multispectral Landsat data, International Conference on Digital Image Computing: Techniques and Applications <https://doi.org/10.1109/DICTA.2016.7797054>
- [9] Australian Bureau of Meteorology (2016) Climate classification of Australia, http://www.bom.gov.au/climate/averages/climatology/gridded-data-info/metadata/md_koppen_classification.shtml
- [10] Scarth, P. (2017) Fractional cover – Landsat, Joint remote sensing research program algorithm, Australia coverage, <http://data.auscover.org.au/xwiki/bin/view/Product+pages/Landsat+Fractional+Cover>
- [11] Scarth, P.F. and Röder, A. and Schmidt, M. (2010) Fractional cover, Proceedings of the 15th Australasian Remote Sensing and Photogrammetry Conference. <https://doi.org/10.6084/m9.figshare.94250>
- [12] DSITI (2015) AussieGRASS Environmental calculator. https://www.longpaddock.qld.gov.au/about/publications/pdf/agrass_user_guide.pdf
- [13] Bastin, G. and Denham, R. and Scarth, P. and Sparrow, A. and Chewings, V. (2014) Remotely-sensed analysis of ground-cover change in Queensland's rangelands, 1988-2005, Rangeland Journal, **36**, 191–204. <http://doi.org/10.1071/RJ13127>
- [14] Leathwick, J. R. and Elith, J. and Francis, M. P. and Hastie, T. and Taylor, P. (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees, Marine Ecology-Progress Series, **321**, 267–281. <http://doi.org/10.3354/meps321267>
- [15] Hastie, T. and Tibshirani, R. and Friedman, J. (2009) The Elements of statistical learning, Springer Series in Statistics, New York. <https://doi.org/10.1007/b94608>
- [16] Tarling, R. (2009) Statistical modelling for social researchers: Principles and practice, Taylor & Francis Group.
- [17] De'ath, G. and Fabricius, K.E. (2000) Classification and regression trees: A powerful yet simple technique for ecological data analysis, Ecology, **81**, 3178–3192. [http://doi.org/10.1890/0012-9658\(2000\)081\[3178:CARTAP\]2.0.CO;2](http://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2)
- [18] Friedman, J.H. (2001) Greedy function approximation : A Gradient Boosting Machine, Institute of Mathematical Statistics, **29**, 1189–1232. <http://www.jstor.org/stable/2699986>
- [19] Robinzonov, N. (2013) Advances in boosting of temporal and spatial models, Ludwig-Maximilians-Universität München. <http://edoc.ub.uni-muenchen.de/15338/>
- [20] James, G. and Witten, D. and Hastie, T. and Tibshirani, R. (2013) An Introduction to statistical learning, Springer Texts in Statistics, 856–875. <http://doi.org/10.1007/978-1-4614-7138-7>
- [21] Breiman, L. (1998) Arcing classifiers, Annals of Statistics, **26**, 801–849. <http://doi.org/10.1214/aos/1024691079>
- [22] Freund, Y. and Schapire, R.E. (1996) Experiments with a new Boosting algorithm, International Conference on Machine Learning, 148–156. <http://dl.acm.org/citation.cfm?id=3091696.3091715>
- [23] Friedman, J.H. (2002) Stochastic gradient boosting, Computational statistics and data analysis, **28**, 367–378. [http://doi.org/10.1016/S0167-9473\(01\)00065-2](http://doi.org/10.1016/S0167-9473(01)00065-2)

-
- [24] R Core Team (2017) R: A language and environment for statistical computing. <https://www.R-project.org/>
- [25] Elith, J. and Leathwick, J.R. and Hastie, T. (2008) A working guide to boosted regression trees, *Journal of Animal Ecology*, **77**, 802–813. <http://doi.org/10.1111/j.1365-2656.2008.01390.x>
- [26] Tierney, N.J. and Harden, F.A. and Harden, M.J. and Mengersen, K.L. (2015) Using decision trees to understand structure in missing data, *BMJ Open*, **5**, 1–12. <http://doi.org/10.1136/bmjopen-2014-007450>
- [27] Jansen, V. and Kolden, C. and Taylor, R. and Newingham, B. (2016) Quantifying livestock effects on bunchgrass vegetation with Landsat ETM+ data across a single growing season, *International Journal of Remote Sensing*, **37**, 150–175. <http://doi.org/10.1080/01431161.2015.1117681>
- [28] Derry, J.F. (2004) Piospheres in semi-arid rangeland: Consequences of spatially constrained plant-herbivore interactions, The University of Edinburgh, 1–305. <http://hdl.handle.net/1842/600>
- [29] Humphries, G.R.W. (2015) Estimating regions of oceanographic importance for seabirds using a-spatial data, *PLoS ONE*, **10**, 1–15. <http://doi.org/10.1371/journal.pone.0137241>
- [30] Schapire, R. (2003) The boosting approach to machine learning – an overview, MSRI Workshop on Nonlinear Estimation and Classification, Springer, New York.
- [31] Chipman, H.A. and George, E.I. and McCulloch, R.E. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York.
- [32] Bell, J.F. (1999) *Tree-based methods, Machine Learning Methods for Ecological Applications*, Kluwer, Dordrecht, 89–05.
- [33] Breiman, L. and Friedman, J. and Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*, Chapman and Hall, Wadsworth, New York.
- [34] Tsangaratos, P. and Ilia, I. (2016) Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece, *Landslides*, **13**, 305–320.
- [35] Cormen, T.H. and Leiserson, C.E. and Rivest, R.L. and Stein, C. (2009) *Introduction to Algorithms*, The MIT Press, Cambridge, Massachusetts.
- [36] Chipman, H.A. and George, E.I. and McCulloch, R.E. (2010) BART: Bayesian additive regression trees, *Annals of Applied Statistics*, **4**, 266–298. <http://doi.org/10.1214/09-AOAS285>
- [37] Hastie, T.J. and Tibshirani, R. (1986) Generalized additive models, *Statistical Science*, **1**, 297–310. <http://doi:10.1214/ss/1177013604>
- [38] Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, *Journal of the Royal Statistical Society (B)*, **73**, 3–36.
- [39] Lary, D.J. and Alavi, A.H. and Gandomi, A.H. and Walker, A.L. (2016) Machine learning in geosciences and remote sensing, *Geoscience Frontiers*, **7**, 3–10. <https://doi.org/10.1016/j.gsf.2015.07.003>

4 Effect of Spatial Aggregation on Prediction Accuracy of Green Vegetation using Boosted Regression Trees

Preamble

This paper focuses on aim A2 in which we investigate in four different spatial aggregation scales and how the spatial aggregation effects the efficiency and prediction accuracy of green vegetation using aggregated FCover data. For the BRT modelling approach we used four spatial scales of four years that result in 16 different BRT models. The aggregated FCover information serves as our new response variable and is included in the modelling along with the latitude and longitude geographic coordinates. These corresponding centroid grid cell coordinates serve as surrogates variables showing the gradient in North-South and East-West direction.

This paper provides insight into the complexities and dependencies of smoothing features in bigger grid cells on heterogeneous land and the resulting predictive performance of the two gradients on green vegetation along with the recorded processing time. The result of this paper as the best identified spatial scale with regards of accuracy and processing time, is later used for chapter 5.

Statement for Authorship

This chapter has been written as a journal article. The authors listed below have certified that:

- (a) They meet the criteria for authorship as they have participated in the conception, execution or interpretation of at least the part of the publication in their field of expertise;
- (b) They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (c) There are no other authors of the publication according to these criteria;
- (d) Potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of the journals of other publications and the head of the responsible academic unit; and
- (e) They agree to the use of the publication in the student's thesis and its publication on the Australian Digital Thesis database consistent with any limitations set by publisher requirements.

The reference for the publication associated with this chapter is; **Student Brigitte Colin, Other supervisors (2018). Influence of Spatial Aggregation on Prediction Accuracy of Green Vegetation using Boosted Regression Trees.** This chapter has been prepared as a paper and has been published.

Contributor	Statement of contribution
Student Brigitte Colin	Conduct the research, develop code for the statistical approach, write the manuscript, make modifications to the manuscript as suggested by coauthors and reviewers.
Signature and date:	
Michael Schmidt	Comments on manuscript.
Samuel Clifford	Comments on manuscript.
Alan Woodley	Comments on manuscript.
Kerrie Mengersen	Propose and supervise research, comments on manuscript.


Principal Supervisor Confirmation: I have sighted email or other correspondence for all co-authors confirming their authorship.

Name: K MENSEN Signature: Signature Date: _____

QUT Verified

Article

Influence of Spatial Aggregation on Prediction Accuracy of Green Vegetation using Boosted Regression Trees

Brigitte Colin ^{1,†,‡}  <https://orcid.org/0000-0003-2866-3212>, Michael Schmidt ^{3,‡}, Samuel Clifford ^{1,‡}, Alan Woodley ^{2,‡} ^{1,‡} and Kerrie Mengersen ^{1,*}

¹ Affiliation 1; School of Mathematical Sciences, Queensland University of Technology, Brisbane QLD 4000, Australia, k.mengersen@qut.edu.au

² Affiliation 2; Institute for Future Environments, Queensland University of Technology, Brisbane QLD 4000, Australia; a.woodley@qut.edu.au

³ Affiliation 3; Department of Environment and Science (DES), Brisbane QLD 4102, Australia; michael.schmidt@des.qld.gov.au

* Correspondence: b.colin@qut.edu.com; Tel.: +61-3138-2019

Version July 25, 2019 submitted to MDPI

Abstract: Data aggregation is a necessity when working with big data. Data reduction steps without loss of information are a scientific and computational challenge but are critical to enable effective data processing and information delineation in data-rich studies. We investigated the effect of four spatial aggregation schemes on Landsat imagery on prediction accuracy of green photosynthetic vegetation (PV) based on fractional cover (FCover). To reduce data volume we created an evenly spaced grid, overlaid that on the PV band and delineated the arithmetic mean of PV fractions contained within each grid cell. The aggregated fractions and the corresponding geographic grid cell coordinates were then used for boosted regression tree prediction models. Model goodness of fit was evaluated by the Root Mean Squared Error (RMSE). Two spatial resolutions (3000m and 6000m) offer good prediction accuracy whereas others show either too much unexplained variability model prediction results or the aggregation resolution smoothed out local PV in heterogeneous land. We further demonstrate the suitability of our aggregation scheme, offering an increased processing time without losing significant topographic information. These findings support the feasibility of using geographic coordinates in the prediction of PV and yield satisfying accuracy in our study area.

Keywords: Boosted regression trees; green vegetation; fractional cover imagery; spatial aggregation; data reduction

1. Introduction

A spatial aggregation of remotely sensed data results generally in a loss of spatial detail. If the object of interest is, however bigger than the pixel resolution an optimal resolution needs to be identified. In the case of monitoring heterogeneous land with Landsat (30m pixels) a spatial aggregation and data reduction of remotely sensed information results in a smoothing effect and with increasing coarseness small features will become lost or mixed into neighbouring pixels. This is especially crucial when predicting green vegetation in different spatial aggregations resolutions where spectral information of green vegetation will be averaged over large areas. However, an increasing coarseness enables a faster processing time and is more efficient when dealing with big data challenges.

The Red and Near Infrared [1] spectral information of remotely sensed imagery have considerable potential for monitoring green vegetation on a regional or local scale. Remote sensing measurement devices are not in direct contact with the objects they sense and therefore offer great advantages and potential in recording large areas. Remotely sensed data are available from a wide range of sources, ranging from satellites to drones, and have been used for a very wide range of environmental applications [2–10].

There is a strong advantage in using remotely sensed Landsat imagery for land use and land cover (LULC) analyses [11,12]. Landsat data are freely available [13], the imagery covers a wide geographical area, and it avoids expensive, extensive and often impractical in-situ measurement. The spatial resolution of a satellite pixel combines the reflected or emitted radiation from different objects on the earth surface, and this spectral mixing effect results in a so-called mixed pixel [14] or Mixel. With the decrease of spatial resolutions, spectra from individual objects cannot be separated and linked to specific features on the ground anymore. There is a range of earth observation satellites available, with different spatial resolution, for example, MODIS (250, 500 and 1000m) with a high temporal resolution to monitor vegetation health, Sentinel-2A/2B that simultaneously records land surface reflectance with a spatial resolution starting from 10m up to 60m, Sentinel 3 (Full resolution: 300m and reduced resolution: 1.2km) primarily used for climate-related studies on sea-land-surface temperatures, AVHRR (1.1km) to monitor clouds and the thermal emission of terrestrial land, and SeaWiFS (1km) that can quantify chlorophyll produced by marine phytoplankton. We refer to high resolution as <15m, moderate resolution as 15-100m, and low resolution as > 100m. LULC analysis with low spatial resolution (hundreds of meters) is more suitable for studies related to climate change, climate variability and environmental degradation.

Fractional cover (FCover) is a derived product based on Landsat 5 Thematic Mapper (TM) imagery. In a spectral unmixing approach the Landsat mixel information is separated into assigned biophysical variables, here bare soil, photosynthetic vegetation (green vegetation) and non-photosynthetic vegetation [15–18]. A spectral unmixing technique was applied to estimate the proportion of green vegetation (PV), senescent or non-photosynthetic active vegetation (nPV) and bare soil (BS) represented in one pixel as percentages ranging from 0% (no representation of one ground cover type) to 100% (full representation) [15], [18]. However, spectral unmixing is not limited to these fractions and not to Landsat imagery. The spectral unmixing approach used for our FCover data is described in [15,19,20].

Using FCover provides a major advantage over using spectral bands and their derived vegetation indices like the NDVI. It is not required to perform an additional ground truth assessment since an extensive data collection has been conducted to collect samples of ground cover that are used for the spectral unmixing algorithm. The Australian FCover we are using for our case study is a standardised and validated product on similar LULC types on heterogeneous land and provided by a state government agency with an overall error of the fractional ground cover with an RMSE of 11.8% [21]. A description of how the ground cover samples were collected is given in [19].

FCover imagery is a fundamental site and landscape scale measurement required by landholders, non-government organizations and state and federal government departments in Australia [21]. PV, nPV and BS are calculated using spectral unmixing models linked to an intensive field sampling program whereby more than 600 sites covering a wide variety of vegetation, soil and climate types were sampled to measure over-storey and ground cover [22]. Fractional cover mapping has been applied in a number of rangeland systems [23–25].

In Australia, FCover products are routinely produced using Landsat imagery and are available at the Terrestrial Ecosystem Research Network (TERN) AusCover remote sensing data archive [26]. The AusCover Data portal aims to deliver consistent national time-series of remotely sensed biophysical parameters to support ecosystem research and natural resource management communities in Australia. These remote sensing products are based on past, current and future satellite image data sets with deliverables designed for Australian conditions. A similar and related product is persistent green vegetation fractions, that focus on woody and mostly vertical vegetation like trees, tree cover, tree density and canopy research [21].

One way to reduce data volume is to aggregate pixels, but this is at the potential expense of loss of accuracy in assigning LULC types based on the coarser FCover values. In this paper, we investigate this issue by creating four even spaced grids and overlaying these on the FCover scenes. All pixels contained within the cell extent are then aggregated by calculating the arithmetic mean representing the green vegetation of this specific grid cell. This aggregation adds an additional level of uncertainty in the estimation of the coefficients of the model. However, by aggregating the fractions of green vegetation we create a source of potential bias and uncertainty in statistical analyses at different spatial resolutions. The modifiable area unit problem (MAUP) occurs when continuous measures of spatial phenomena are aggregated into a higher order grid [27]. The association between variables depends on the size of the grid cell extent over which the FCover fractions are averaged.

Ershadi et al [28] investigated the effect of aggregating heat surface flux from fine (<100m) to medium (approx 1km) resolution using Landsat 5 imagery and indicated that aggregation using simple averaging methods have limited effect on land surface temperature compared to more sophisticated approaches. Moreover, by using the simple arithmetic mean to extract the required fraction of each grid cell we preserve the spatial distribution over the whole FCover scene [29].

In this paper, we use a boosted regression tree (BRT) to link the response variable (FCover) to the two covariates, namely latitude and longitude of the centroid of the area. A BRT is a popular statistical and supervised machine learning approach that has been readily applied to remotely sensed data. Indeed, although they were first defined two decades ago, BRTs have only recently been extended to deal with the types of features that are characteristic of remotely sensed data, in particular, its spatial and temporal dynamics. BRTs combine two algorithms (regression trees and boosting) and arguably yield higher prediction accuracy than simple tree-based methods such as a Classification and Regression Trees (CART) [30]. There are two major advantages of using BRT over more traditional regression methods. First, it allows a more flexible partition of the feature space that is not as rigid as using a simple linear regression. BRT combines simple binary partitions to form a complex prediction rule that can more accurately identify small areas of interest. Second, it can deal with missing values by default like masked out areas (clouds and cloud shadows), water bodies or the Scan Line Error of Landsat 7 ETM+. This is a great advantage especially when using remotely sensed imagery that has gone through several quality refinements and processing levels to filter out obscuring elements that leave data gaps behind.

The aggregated fractions of green vegetation derived from the FCover scene serve as our response variable. The delineated centroid coordinates from the midpoint of the spatial grid cells serve as surrogates for other spatial covariates and represent a north-south gradient shown as a vector of latitude coordinates and an east-west gradient shown as a vector of longitude coordinates. These surrogate variables will be used to statistically analyse the relationship to our response variable and the quantitative impact on prediction accuracy of different spatial aggregation schemes.

The use of latitude and longitude as surrogates for other covariates is not uncommon. For example, in a study of the geographic distribution of plant functional types [31] the authors examined the relationship of precipitation and temperature on C3 and C4 grass types and shrubs using latitude and longitude coordinates and concluded that latitude and longitude can be used as surrogate variables for the main climatic dimensions in North America. The latitude and longitude explained a substantial portion of the variability of the distribution of the relative abundance of shrubs, C3 grasses, and C4 grasses. Along a given longitude, C3 grasses increased with latitude. As one moves westward, C4 grasses are replaced by shrubs. In another study [32] the authors plotted latitude and longitude coordinates and included these as surrogate variables to account for variation in climate associated with geographic location within deciduous forested ecoregions. The response was an aggregated NDVI variable used as an on-site quantification of vegetation in North America.

In summary, the objective of our study was to analyse the statistical dependence between our two surrogates, the centroid coordinates in latitude and longitude, and their ability to predict the aggregated fractions of green vegetation delineated from the FCover scene. The focus is on the prediction accuracy achieved in four spatial resolutions and the preprocessing time needed to extract and aggregate the green fractions out of the FCover scene. We use a BRT to link FCover with the two covariates.

The paper is structured as follows. Section 2 presents the data and BRT methodology used for predicting green vegetation using geographic centroid coordinates of evenly spaced spatial grid cells, the relevance of the spatial aggregation measured as a model fit and a brief reminder about the principles of spectral unmixing approaches and its outcome. Section 3 presents the results structured in three groups: (1) the comparison of the model fit showing the distribution of the residuals around the mean, (2) the variable interactions as the relative influence and partial dependencies of the covariates on the response variable, the relationship and distribution of the predicted versus the observed test data set in marginal plots and model diagnostics and (3) the aggregation and scaling errors using different spatial resolutions. The outcome and the relevance of this work to real world scenarios and limitations of BRT are discussed in Section 4.

2. Material and Methods

2.1. Case Study

The study area used in our assessment is located in the Northern Territory, Australia. Figure 1 shows the location of the FCover scenes at the Landsat footprint of path 102 row 72 on the Worldwide Reference System-2 (WRS-2) and it is covering an area of 185 × 185 km. The geographic coordinates are given as centroids showing latitude -17.345 and longitude 135.587. For consistency over time, and because the FCover in the study area is dominated by wet and dry seasons, only December scenes indicating the very early period of the wet season have been used for this case study. Estimating FCover at this time of the year is important for agricultural managers. The study area is a heterogeneous region with a complex topography of native grass types.

Our study area is defined as "dry" with variations of "desert, hot arid" and "dry summer, hot arid" (BWh and Bsh) based on the Koeppen-Geiger scheme and is very vulnerable with regard to climate variability [33], [34]. The daily rainfall in December has been recorded as lowest at 16.2 mm in 1990 and highest at 96.4 mm in 1989, and the monthly total ranged from 38.0 mm in 1990 to 137.2 mm in 1987 [34].

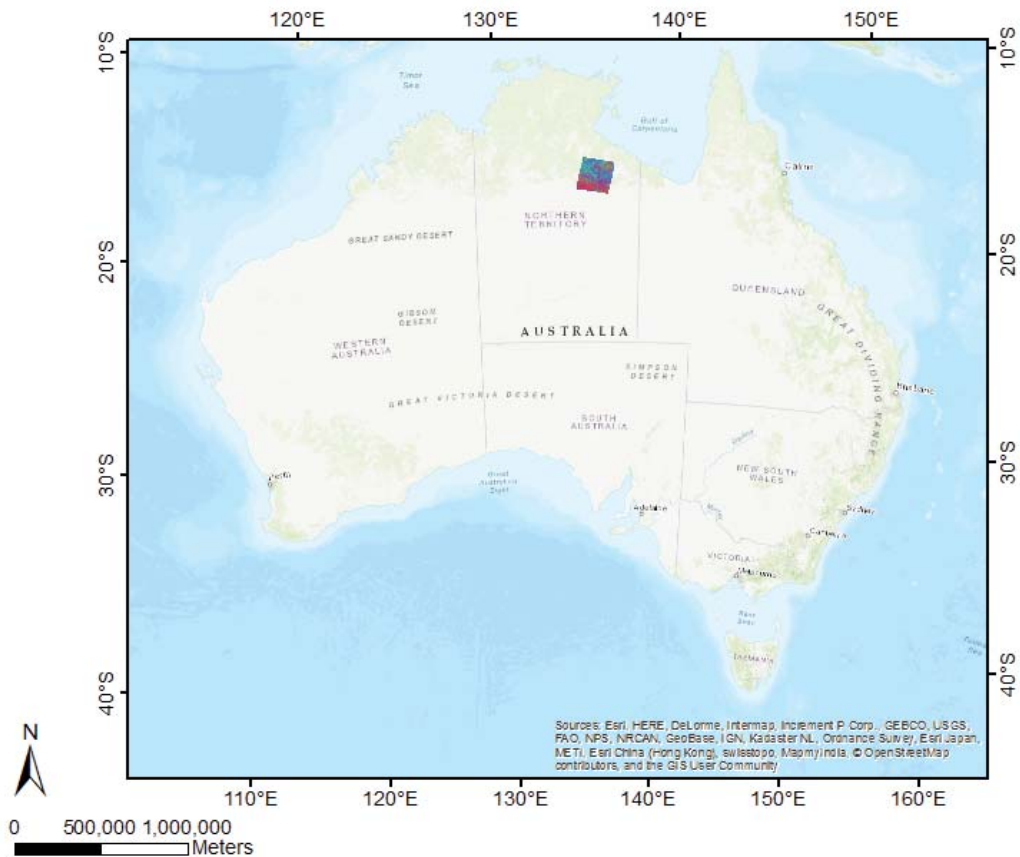


Figure 1. FCover scene in the Northern Territory showing the Landsat footprint of path 102 row 72 at the Worldwide Reference System-2 (WRS-2) and is covering an area of 185 x 185km.

We used a Digital Elevation Model (DEM) and generated equidistant contour lines in 50m intervals. Two DEMs were merged and clipped together to the full extent of our spatial raster grid. We used the freely available SRTM 90m resolution and used focal statistics on a 33 x 33 cell neighbourhood to smooth the surface so that the spatial resolution of the DEM represents our aggregated green vegetation fractions better. The highest point is 255m and the lowest is located at 23m ($\Delta 232m$) above mean sea level (MSL) referenced to the Australian Height Datum (AHD). In Figure 2 we can see a constant increase of about 15% of the elevation between the upper right corner (min 23m) to the lower left corner (max 255m).

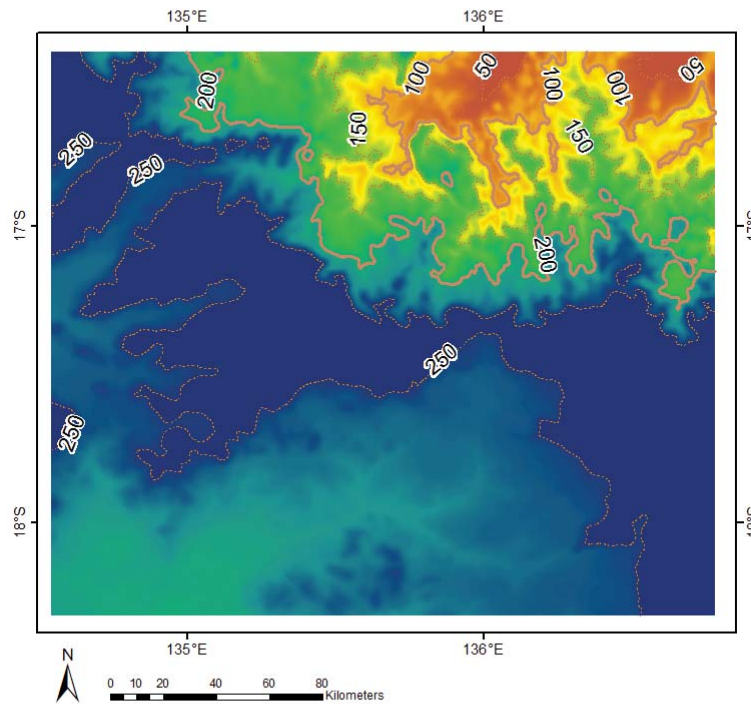


Figure 2. SRTM DEM and calculated contour lines in 50m intervals showing the Landsat footprint of path 102 row 72.

2.2. Data

2.2.1. Spectral unmixing approach

The opening of the Landsat archive and a new open data policy had have revolutionised the use of Landsat data [13]. The Fractional Cover product is derived from Geoscience Australia and the Australian Reflectance Grid 25 (ARG25) product and provides fractional cover representation of the proportions of green or photosynthetic vegetation, non-photosynthetic vegetation, and bare surface cover across the Australian continent. It is generated using the algorithm developed by the Joint Remote Sensing Research Program (JRSRP) and described in [19]. FCover is available for

Landsat Thematic Mapper (Landsat 5), Enhanced Thematic Mapper (Landsat 7) and Operational Land Imager (Landsat 8). FCover was made possible by new scientific and technical capabilities, the collaborative framework established by the Terrestrial Ecosystem Research Network (TERN) through the National Collaborative Research Infrastructure Strategy (NCRIS), and the leadership and capabilities of Geoscience Australia and the Joint Remote Sensing Research Program [35].

The spectral unmixing approach aims to separate the spectral reflectance of one pixel into its single ground cover components to determine the proportions of each of three classes PV, nPV and BS. The result of spectral unmixing is a series of three layers showing the fraction of each abundance images corresponding to each class and an image depicting the root mean square error (RMSE). Our FCover scene is located in the Northern Territory where the land is mainly used for grazing. It is rare to find a pure pixel in heterogeneous grazing land [19]. Further information about how the field data collection has been conducted and how to derive spectral endmembers using spectral unmixing approaches is provided in [36].

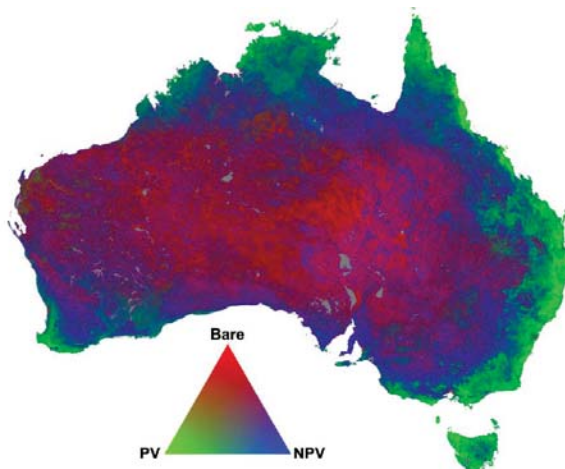


Figure 3. A multilayer FCover composite derived from Landsat 5 and available at the Terrestrial Ecosystem Research Network (TERN) AusCover remote sensing data archive [26].

Figure 3 shows a national FCover product for Australia. The triangular ternary diagram can be read anti-clockwise between PV, nPV and BS. The interpretation of the colour coded fractions is based on the additive colour coding principle showing the relationship between the three endmembers. A quantitative Attribute Accuracy Assessment of the spectral unmixing approach and the overall error of the fractional ground cover RMSE is 11.8%, while the error margins vary for the three different layers where green vegetation has an RMSE of 11.0%, non-green vegetation 17.4%, and bare soil 12.5%. The validated Landsat derived fractional cover products are now used as key indicators for a range of environmental monitoring and management activities [26].

2.3. Data exploration for FCover imagery

A FCover scene consists of three layers showing the fractions of each ground cover class in each layer. As part of our explanatory data analysis, we plotted histograms for all four years of the study period to review the distributions of PV, nPV and BS. Figure 4 shows the ground cover classes of the Landsat FCover bands combined in one histogram, along with the frequencies. We can clearly see that green vegetation has the least fractions but a high frequency, whereas non-green vegetation has higher fractions presented in one pixel.

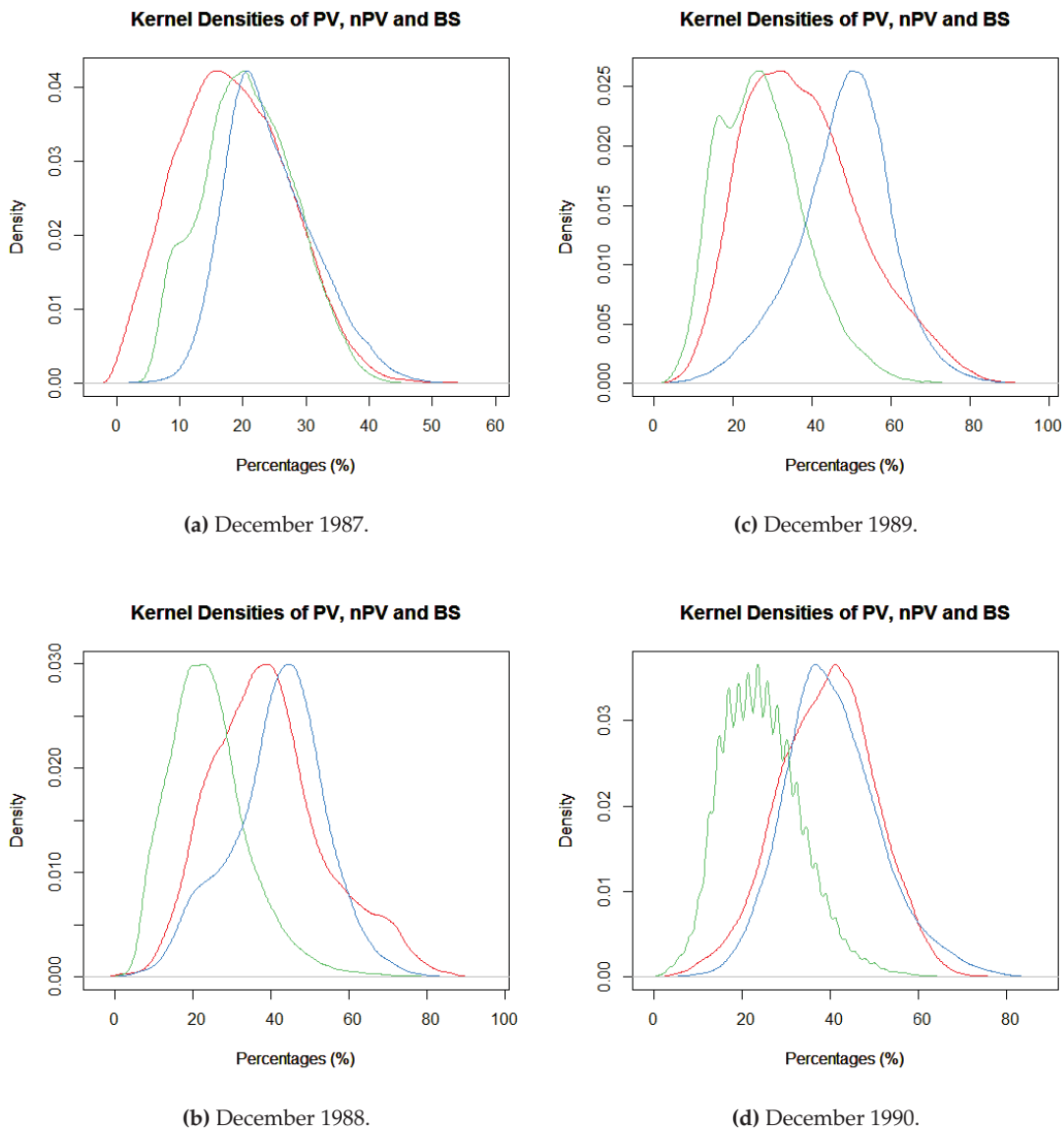


Figure 4. Kernel Density plots of all Landsat FCover bands representing the individual fractions of bare soil in red, green vegetation in green, not green vegetation in blue in one pixel. Subfigure a) December 1987, b) December 1988, c) December 1989 and d) December 1990.

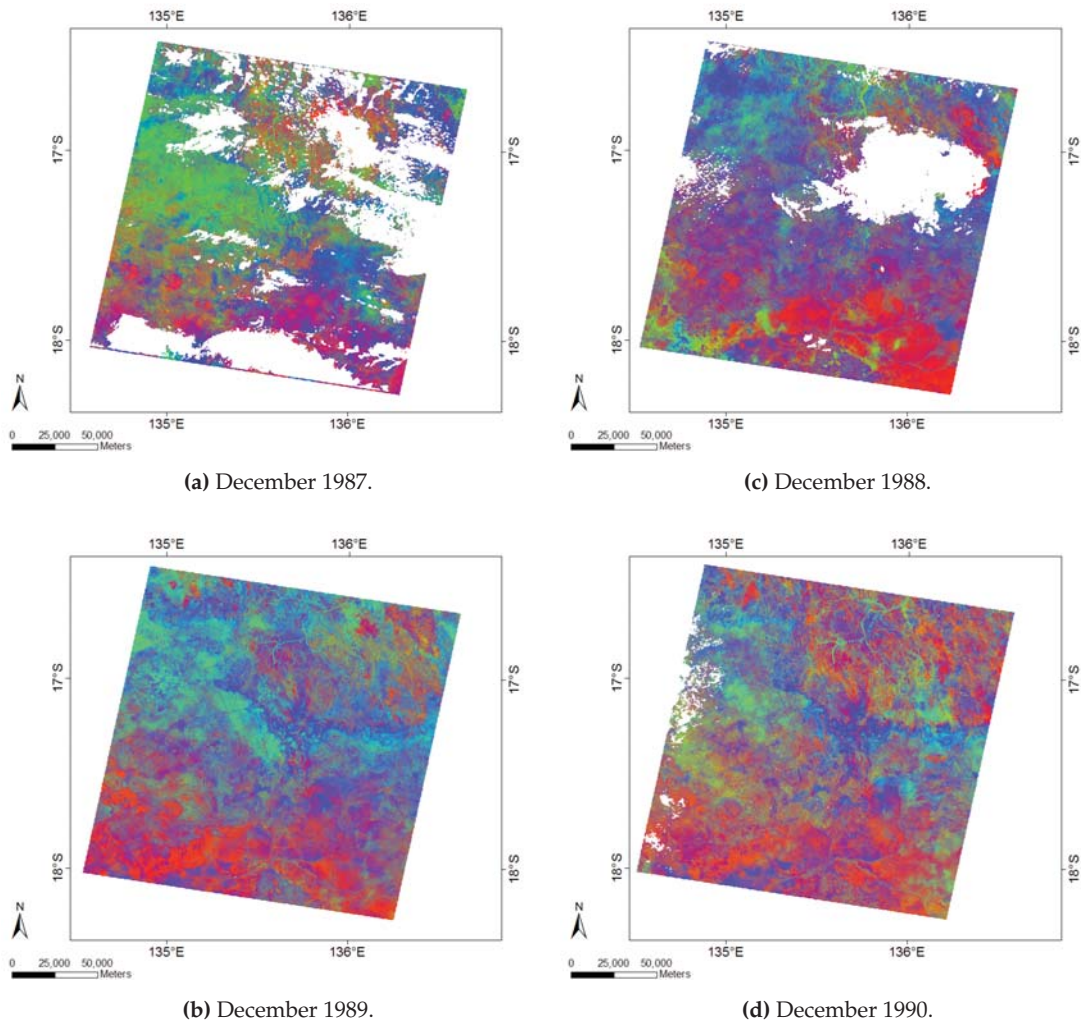


Figure 5. FCover of Landsat Thematic Mapper (Landsat 5) scenes of four years showing white data gaps caused by masking out clouds and clouds shadows.

The histograms indicated a roughly normal distribution for each of the classes. It can be seen that the green photosynthetic vegetation has the smallest fractions. In 1988, 1989 and 1990 the green photosynthetic vegetation was represented as the smallest fraction but with the highest frequency, except in 1987 where the bare soil has the lowest percentages and the mode (represented as the highest bar) of the green vegetation shifts towards 20% and higher. This is an indicator that in 1987 the PV is more strongly represented than in the rest of the three years and therefore, we can infer that December 1987 was our wettest month. This is in accordance with the recorded rainfall data, described in the case study, where the monthly total is the highest in all of our FCover scenes. Moreover, the mode of green vegetation is smallest (around 12.5%) in 1990 and this reflects the lowest recorded monthly total and the lowest recorded daily rainfall in December 1990 as described earlier. Hence, 1990 is described as our driest year [34]. Figure 5 shows the four FCover scenes and their masked out areas.

2.4. Data pre-processing and spatial aggregation

The aggregation involved several pre-processing steps. As one of the pre-processing steps, we created four evenly spaced spatial grid cell layers in four different spatial resolutions, showing the same geographic reference as the FCover scenes and overlaid this on the raster image. The spatial grid layers were used as a vector overlay on the FCover scene showing varying coarseness of the grid cells extents ranging from the spatial resolution of 12000m, 6000m, 3000m and 1500m. In addition, we ensured that the edges of the spatial grids lined up with the edges of the FCover pixels. Further, all missing values were removed and the arithmetic mean was calculated for each spatial grid cell. Figure 6 shows the spatial grid on top of the FCover scene at a resolution of 3000m. The figure also shows the extent of missing data, due to masking out obscuring elements such as cloud and cloud shadow.

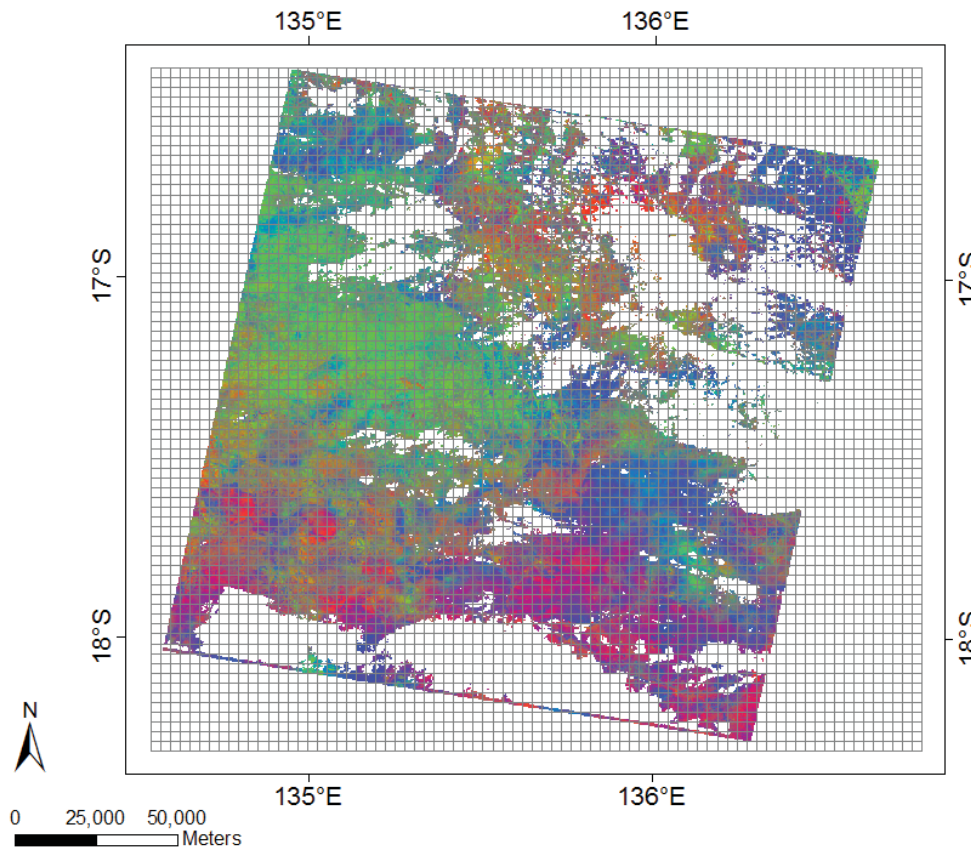


Figure 6. Spatial grid cells at a resolution of 3000m. The total number of cells 5530. The FCover data are mapped to a even spaced grid where each grid cell contains 100 x 100 pixels and covers an area of 3000 x 3000m. Please refer to Figure 3 for the triangular ternary diagram for the coloured relationships of the three ground cover types.

The spatial resolution determines the geographic extent of each spatial grid cell in the FCover scene. One spatial grid cell in 12000m contained 400 x 400 FCover pixels each having a geographic resolution of 30 x 30m and covering a total area of 12000 x 12000m on the ground ($400 \times 30\text{m} = 12000\text{m}$). In contrast, the spatial grid resolution of 1500m contains 50 x 50 pixels and covers an area of 1500 x 1500m within the spatial grid cell. Table 1 lists all the spatial resolutions used in this study, the number of pixels contained within a spatial grid cell as an overlay on the FCover scene, the total area covered on the ground and the total number of spatial grid cells in the overlay used for the proposed aggregation scheme. The choice of the spatial resolutions allows for consistent arithmetic averages of

FCover to be taken over the aggregated cells.

Table 1. Table of proposed data reduction scheme. The table shows the smallest resolution of 12000m up to the largest resolution of 1500m and resulting total number of spatial grid cells used for the following spatial aggregation steps. By proposing our data reduction scheme we are not dealing with the original number of 54 million pixels per FCover scene organised in about 7000 rows and 8000 columns.

Spatial resolution (m)	Number of pixels in grid each cell	Ground covered by each grid cell (m)	Total number of grid cells	Coloured outline of spatial grids
original	1 x 1	30 x 30	54 million	FCover pixel
12000	400 x 400	12000 x 12000	360	black
6000	200 x 200	6000 x 6000	1400	green
3000	100 x 100	3000 x 3000	5530	red
1500	50 x 50	1500 x 1500	21980	grey

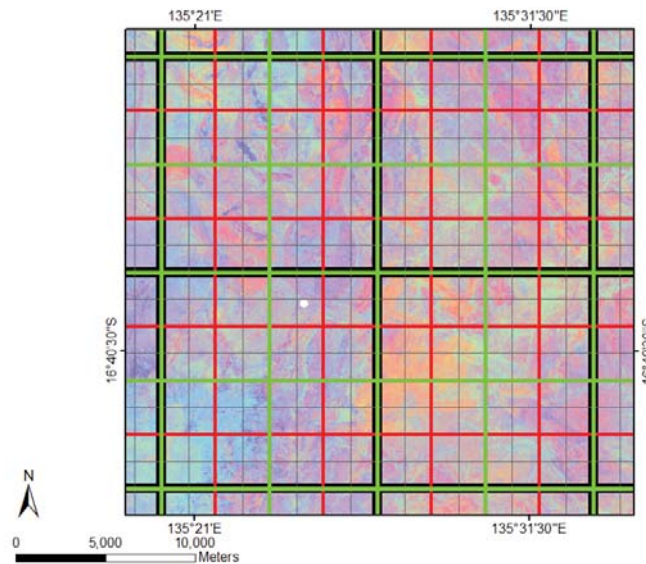


Figure 7. Combination of all four spatial grids used as an overlay for the data delineation of green vegetation fractions out of FCover scenes. The thick black outline shows the resolution in 12000m, green in 6000m, red in 3000m and thin grey in 1500m.

The four spatial grids demonstrated in Figure 7 were obtained using the open source software GME (Geospatial Modelling Environment). GME currently has dependencies on ArcGIS and R where it uses the statistical engine to drive some of the analysis tools.

Each individual grid cell was used to calculate the arithmetic mean as a measure of central tendency of all the pixels contained within the spatial grid cell extent. As a result, one aggregated value of all green vegetation fractions contained within the grid cell extent represented each individual grid cell with the aggregated PV fraction. Since the spatial grid cells line up with the edges of the FCover pixels, adjacent and overlapping pixels will not be considered in the aggregation process.

In addition to aggregating fractions of green vegetation spatial grid cells sizes we delineated the centroid coordinates as geographic latitude and longitude coordinates of each grid cell. The resultant

csv file contained the response variable of aggregated fractions of green vegetation and the centroid coordinates in latitude (North-South direction) and in longitude (East-West direction). As discussed in the Introduction, no additional environmental data were used for the following modelling process using BRT. Altogether 16 csv files were created representing four spatial aggregations scheme for four years. Table 2 shows further details.

Table 2. The size of the pre-processed data set varies according to coarseness of the spatial aggregation resolutions.

Spatial resolution (m)	Number of grid cells in overlay	Length of aggregated response variable
12000	360	360
6000	1400	1400
3000	5530	5530
1500	21980	21980

2.5. Boosted regression trees

A boosted regression tree (BRT), also known as gradient boosted machine (GBM) or stochastic gradient boosting (SGB), is a non-parametric regression technique that combines a regression tree with a boosting algorithm [37]. This extension to the classical regression tree allows greater flexibility and predictive performance in modelling the data. The implementation of these methods used in this study can be found in the *gbm* R package [38].

A regression tree partitions multivariate data with a hierarchy of binary splits that define regions of the covariate space in which the response variable has similar values. These splits are defined by rules, distance metrics or information gain. The choice of variables and the value at which the split point occurs are determined in a recursive manner at each stage of the tree construction. The segmentation can be depicted as a tree-like structure, comprising nodes representing the selected factors, branches acting as if-else connectors between the nodes, and leaves representing terminal nodes containing the subsets of responses [39–41].

The performance of the simple base learner is improved by boosting, whereby a sequence of trees is grown, such that in each subsequent tree greater attention is paid to observations with greater prediction error. This is achieved by iteratively shifting the focus towards those observations until a stopping rule is reached. The shift is effected by up-weighting observations that were misclassified or had large residual errors in the previous iteration. The deeper tree accommodates more segments and hence captures more variance. This results in higher model complexity but also higher risk of overfitting the model to the data. The motivation behind boosting is that each tree can be quite shallow (a weak classifier) and thus fast to estimate, but by combining the predictive power of many weak classifiers, a classifier of arbitrary accuracy and precision can be created [42–44].

Next, the current approximation $F_{m-1}(\mathbf{x})$ is individually updated in all of the corresponding regions

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + v \cdot \gamma_{lm} 1(\mathbf{x} \in R_{lm}). \quad (1)$$

The shrinkage parameter, v , ranges from 0 to 1 and controls the learning rate γ , so each gradient step is reduced by some factor between 0 and 1 of the learning rate. The value of γ is influenced by the choice of loss function ψ .

The Stochastic Gradient Boosting algorithm is summarised as pseudo code in algorithm 1 [44,45].

Algorithm 1 Stochastic Gradient Boosting algorithmTraining data $\{y_i, \mathbf{x}_{i1}\}_i^N$

Initialization

$$F_0(\mathbf{x}) = \arg \min_{\gamma} \sum_{i=1}^N \psi(y_i, \gamma)$$

for $m = 1$ to M **do**

$$\{\pi(i)\}_1^N = \text{randperm} \{i\}_1^N$$

Compute pseudo-residuals

$$\tilde{y}_{\pi(i)m} = - \left[\frac{\partial \psi(y_{\pi(i)}, F(\mathbf{x}_{\pi(i)}))}{\partial F(\mathbf{x}_{\pi(i)})} \right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, \tilde{N}$$

Fit a base learner to pseudo-residuals

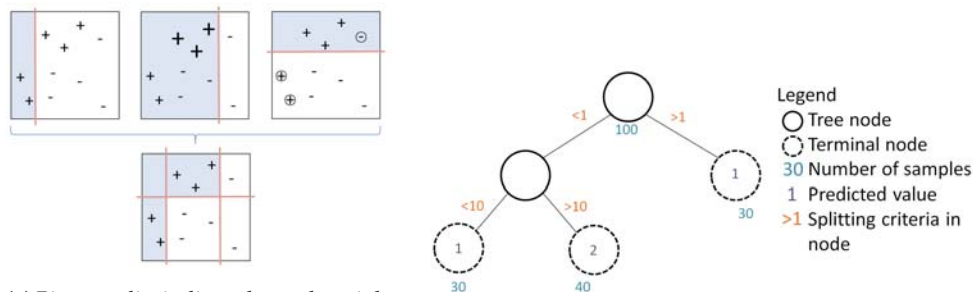
$$\{R_{lm}\}_1^L = L\text{-terminal node tree} \left(\left\{ \tilde{y}_{\pi(i)m}, \mathbf{x}_{\pi(i)} \right\}_1^{\tilde{N}} \right)$$

Compute multiplier γ_{lm} by solving optimization problem

$$\gamma_{lm} = \arg \min_{\gamma} \sum_{\mathbf{x}_{\pi(i)} \in R_{lm}} \psi(y_{\pi(i)}, F_{m-1}(\mathbf{x}_{\pi(i)}) + \gamma)$$

Update the model

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} \mathbf{1}(\mathbf{x} \in R_{lm})$$



(a) Binary splits indicated as red straight lines separate the data in grey and white sections and create weak learners as seen in equation (1). BRT as an ensemble approach combines them to create complex prediction rules. Adapted from [46].

(b) Hierarchical regression and binary splitting process showing observations in the nodes, predicted values in the terminal nodes and splitting criteria along the tree branches.

Figure 8. Figure 8a shows the combination of weak learners to one strong prediction rule used in the BRT ensemble approach and Figure 8b illustrated the hierarchical regression and binary splitting process along the branches of the decision tree.

Figure 8a shows four splits of the whole feature space of the data where the goal is to predict the plus symbols (+). The first three are binary splits that will be combined into one complex splitting rule (bottom). This yields a more accurate prediction result by separating the data allowing for flexible splitting boundaries. The first binary split (left) shown as the red vertical line has incorrectly predicted three observations indicated with a plus symbol (+). The misclassified observations get a higher weight to make sure those are favoured in the next splitting iteration (middle). The plot in the middle shows that three observations indicated with a minus symbol (-) are now misclassified. In the following step those misclassified observations will get higher weights again to be prioritised in the next splitting process. This time a horizontal line is generated. BRT is an ensemble approach and combines the first

three binary splits above into one in order to create a complex prediction rule to split data allowing for identification of small areas of interests. This is the boosting part of BRT.

2.6. Implementation

The R package *caret* [47] was used for two tasks. The first was to split the data into training and test datasets (random partition that assigns 80% of the data in a training set and the remaining 20% to a test set) and the second task was to tune the hyperparameters for BRT modelling.

Typical hyperparameters include the

- shrinkage; (how quickly the algorithm adapts)
- tree complexity; the total number of trees in the final model (number of iterations)
- interaction depth; interaction between different nodes along the branch
- minimum observations in node; minimum number of training set samples in a node to commence splitting.

A feature of the BRT algorithm is that the performance can be tuned to accommodate specific data structures and characteristics through specification of hyperparameters. For our BRT model, the *caret* package was employed to find optimal values for the hyperparameters listed above. We used the automatic grid search method for searching optimal parameters, combined with other methods for estimating the performance of our *gbm* model based on our aggregated FCover data.

The outcome of the tuning process for all the 16 models was a recommendation of number of trees = 2500, interaction depth = 5, (only data of 1987 in 12000m recommended 3), shrinkage = 0.01, and minimum observations in node = 10. Those hyperparameters were then used to estimate the coefficients using the training data, and the prediction results are based on the test data set. Cross-Validation methods were used for the tuning process to help identify the hyperparameters and to restrict the number of iterations (hyperparameter tree complexity) to avoid overfitting when the local minimum has been reached. Empirically, it has been found that using a small value for shrinkage results in impressive improvements in a model's generalisation ability [45]. The drawback of a lower learning rate is that more trees need to be generated, resulting in increased computational time. As described above, 16 BRT models were created showing four years in four spatial resolutions; see Table 1.

2.7. Quantitative assessment of the model fit

The accuracy of the 16 BRT models was primarily analysed on the basis of the root mean square error (RMSE), the mean absolute error (MAE) and the median absolute error (MDAE), where we measured the difference between values predicted by a model and the values actually observed from the environment that is being modelled on the test dataset. In general, the RMSE is best when it is small, but there is no absolute good or bad threshold. The RMSE ranging between 3.3 and 1.1 indicates a good model fit throughout all resolutions.

3. Results

The computational environment was the R statistical modelling software version 3.3.3 [48] running inside Windows 7 SP1 (64-bit) on a 2.60 GHz Intel i7 CPU with 16GB of RAM. All of the plots were generated in the R programming language [48] and maps throughout this paper were created using ArcGIS® software by Esri. The GBM model implementations were taken from the *gbm* package [38]. We structure our results in three main groups. Since we want to investigate prediction accuracy using different spatial aggregations we first checked the residuals and how they spread around the mean of the regression line and the model fit in all the 16 models. Second, we evaluated the influence of each covariate on the response, shown by relative influence plots, or the functional relationships between the covariates and the prediction outcome indicated by partial dependency plots. Further, we

investigated the relationship and distribution of the observed versus the predicted values in marginal plots. Last, we visualised the absolute error rate depending on the spatial resolution in all years and compared those with the elapsed time.

3.1. Comparison of model fit at different spatial resolutions

3.1.1. Deviation of residuals around the mean

Summary statistics and plots revealed that the residuals of the fitted models were relatively unbiased and homoscedastic. The residual plot of the worst model fit of the year 1988 in 1500m and 12000m showed a slight tendency to heteroscedasticity due to a larger variance of the fitted values towards the maximum number of observations and further the resolution 12000m showed an unbalanced spread around the regression line towards under-predicted values shown in Figure 9. These effects were not visible in any of the residual plots for the best model fit in the year 1990 demonstrated in Figure 10.

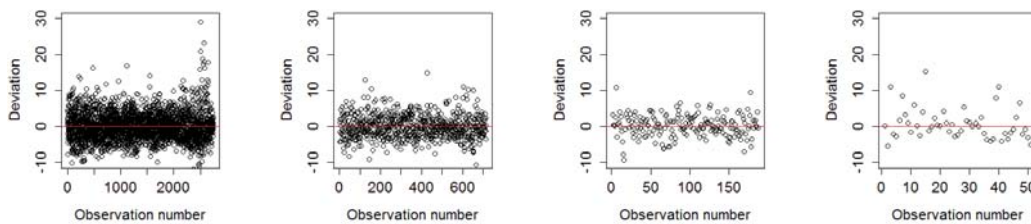


Figure 9. Deviation of residuals around the mean of 1988 in all four resolutions as worst model fit.

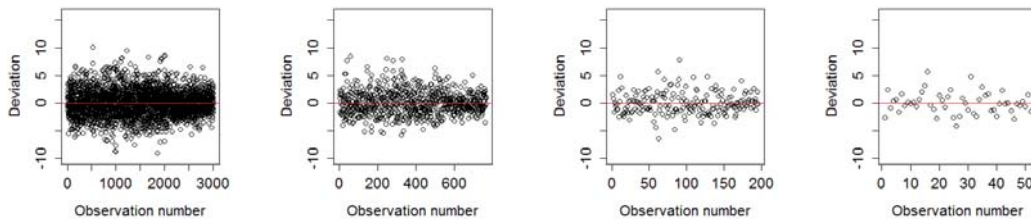


Figure 10. Deviation of residuals around the mean of 1990 in all four resolutions as best model fit.

Figure 11 shows the combined residuals over all years for all resolutions on the left and the corresponding box plots on the right.

The box plots show that the deviation of the residuals within the Inter Quartile Range (IQR), indicated as the white box around the zero line, is similar regardless of the spatial aggregation. However, there is more variation in the resolution of 1500m than in any other resolutions. This can be explained by the argument that the loss function ψ used in the BRT and the weighting of problematic observations result in a similar deviation of the residuals at all aggregated spatial resolutions.

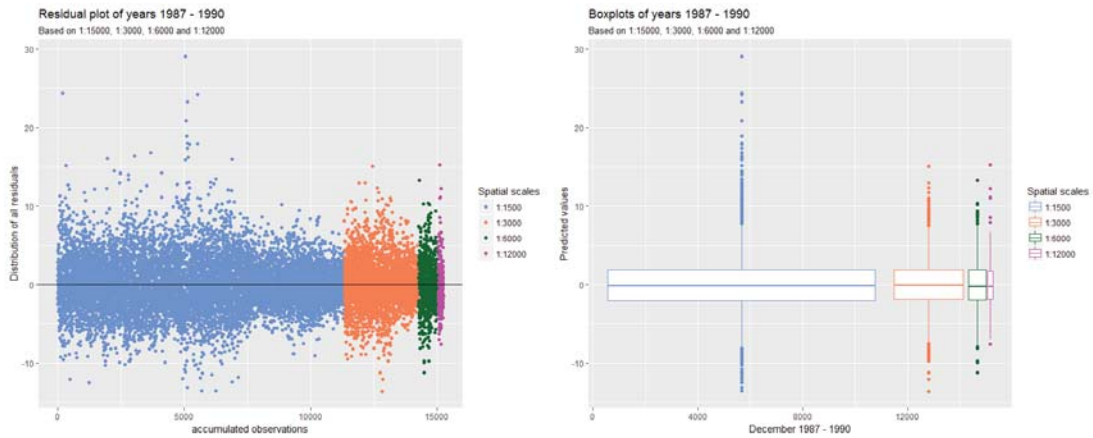


Figure 11. Deviation of residuals in all resolutions combined in one plot (left) and corresponding box plots (right).

We can see that the 6000m resolution has the least error rates and is most symmetrically distributed around the black line showing the mean of the residuals. We conclude that aggregating from an initial geographic resolution of 30 × 30m to 6000m resolution results in the largest reduction in data volume without sacrificing precision of the prediction.

Table 3 shows the RMSE error rates for the four resolutions and four years. In general the smaller the RMSE error, the better the model fit.

Table 3. Comparison of the RMSE in all four years and resolutions.

Spatial resolution (m)	Year	RMSE
12000	1987	3.0583
	1988	3.9691
	1989	3.0056
	1990	1.6151
6000	1987	2.8583
	1988	3.1428
	1989	3.1591
	1990	1.9577
3000	1987	3.1120
	1988	3.2134
	1989	3.1543
	1990	2.0731
1500	1987	3.4241
	1988	3.8306
	1989	3.4500
	1990	2.3348

3.1.2. RMSE Comparisons between BRT and Linear Model (LM)

In order to evaluate the comparative performance of the BRT results, the data were also analysed using a linear regression model. The R package `lm.br` [49] was used to fit the model. We assume that green vegetation, denoted as Y_i , is linearly related to the covariates latitude and longitude, denoted as X_1 and X_2 respectively, and the residuals ε_i are distributed $N(0, \sigma^2)$. The LM was formulated as follows: $Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \varepsilon_i$.

Table 4. Comparison of RMSE of LM and BRT on worst (1988) and best (1990) model fit.

Spatial resolution (m)	1988		1990	
	Linear Model	BRT	Linear Model	BRT
12000	4.0551	3.9691	2.7933	1.6151
6000	4.9710	3.1428	3.0449	1.9577
3000	5.3688	3.2134	3.3028	2.0731
1500	5.5863	3.8306	3.5676	2.3348

The comparative goodness of fit of the LM and the BRT is shown in Table 4. It is clear that under all four spatial resolutions, the BRT delivers a smaller RMSE. Based on this measure of performance, the BRT is argued to be an attractive alternative to the more common LM approach for analysing these types of data.

3.1.3. Mean absolute error (MAE) and median absolute error (MDAE)

In addition to the RMSE, we calculated the mean absolute error (MAE) and the median absolute error (MDAE) shown in Table 5. MAE computes the average absolute difference between observed and predicted values as the vertical or horizontal distance between each point in a scatter plot. MDAE computes the median absolute difference between the two variables. In section 3.2.4 we see in the marginal plots that BRT under-predicts peak values. In section 3.3 we use the absolute error to quantitatively assess the difference between observed and predicted values for all four spatial resolutions and all four years.

Table 5. MAE and MDAE of the worst (1988) and best model fit (1990) in four resolutions.

Spatial resolution (m)	Mean Absolute Error (worst/best)	Median Absolute error (worst/best)
12000	2.752/1.236	2.236/0.836
6000	2.370/1.500	1.909/1.185
3000	2.489/1.613	2.053/1.305
1500	2.925/1.808	2.398/1.467

3.2. Variable importance

3.2.1. Relative influence of covariates at different resolutions

One way of showing the relationships of the joint probability and contribution of our geographic coordinates in describing the response is through a relative influence plot. The relative influence is calculated by averaging the number of times a covariate is used in the tree building process, weighted by the squared improvement to the model as the result of each split. It is then scaled so the values sum to 100 [50]. Relative influence plots were used to compare the covariates with respect to their explanatory power. Regardless of the spatial resolution, among the two covariates used in the BRT model, the latitude (CenterY) is always more dominant than the longitude (CenterX). Moreover, the influence of the longitude (East/West direction) reduces as the spatial resolution is decreased towards 12000m. However, this is not a consistent reduction. In Figure 12 we demonstrate the influence of CenterX and CenterY covariates and their contribution towards predicting the aggregated green vegetation in the year 1989. The plots show the contribution at the best-estimated number of trees of 2500 iterations starting at 73.91% in 1500m and reaching the maximal influence of 83.15% in 12000m. The relative influence of latitude (CenterY) dominates considerably over longitude (CenterX).

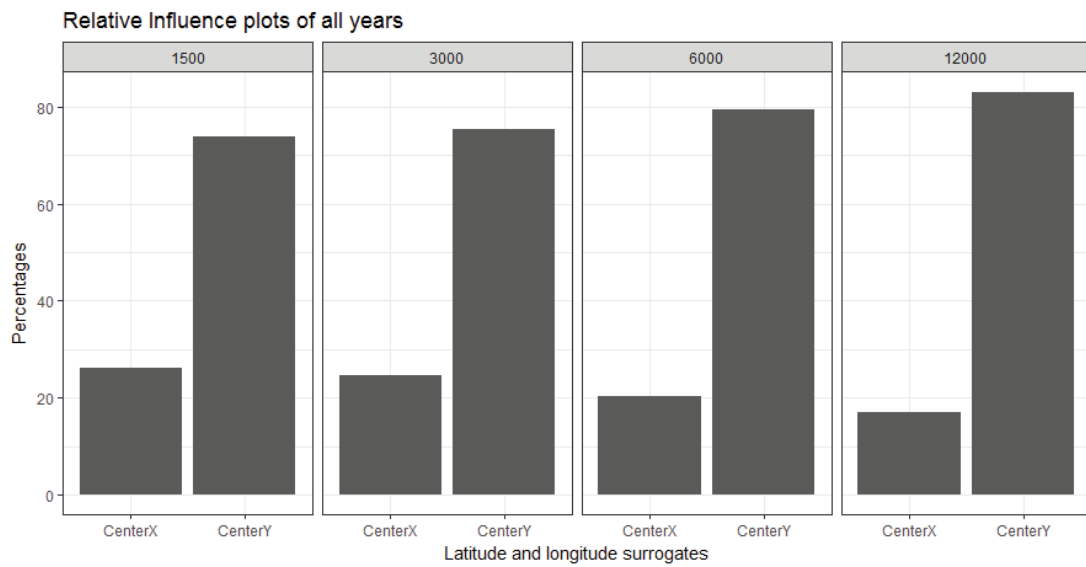
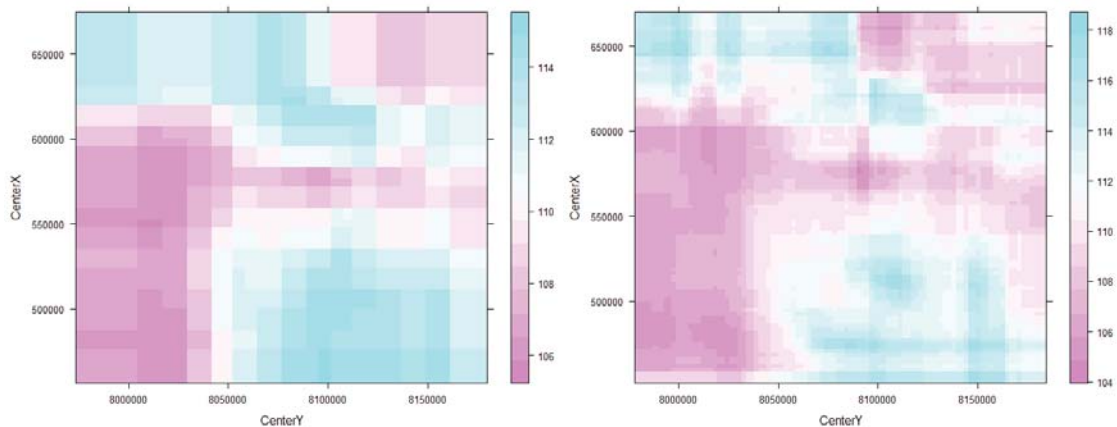


Figure 12. Relative influence plots of December 1989 in all four resolutions showing the contribution of the centroid coordinate of the latitude (CenterY) and longitude (CenterX).

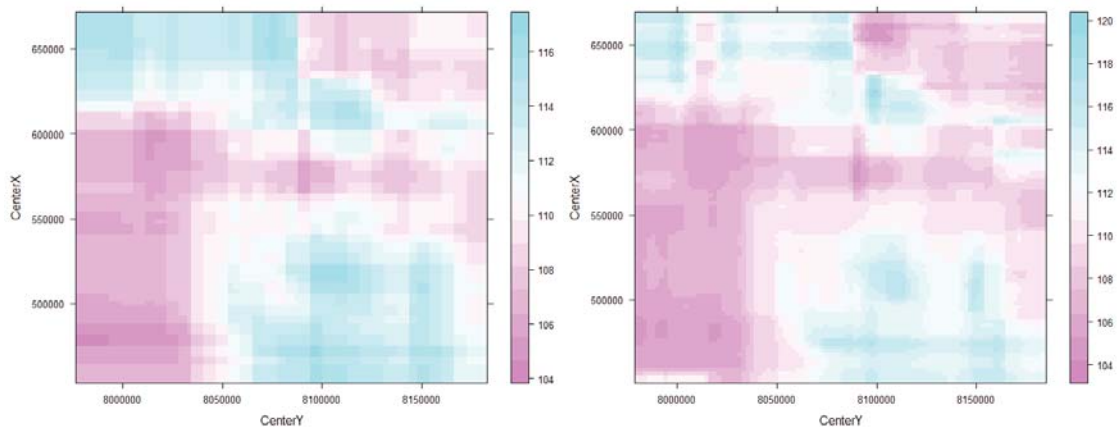
3.2.2. Prediction raster maps

The Prediction Raster Maps clearly demonstrate a change in the marginal effect across spatial resolutions, seen as a smoothing effect towards the 12000m resolution; see Figure 13.



(a) December 1990 in the resolution of 12000m.

(c) December 1990 in the resolution of 3000m.



(b) December 1990 in the resolution of 6000m.

(d) December 1990 in the resolution of 1500m.

Figure 13. Prediction raster maps for the year 1990. a) 12000m, b) 6000m c) 3000m and d) 1500m.

3.2.3. Prediction Surface Plots

As fractional cover varies with the geographic coordinates, the partial dependence can be shown as a prediction surface plot. Here, the independent variables CenterX and CenterY are plotted against the model outcome \bar{y} after considering the average effect of the other independent variable in the model. Since we only have geographic coordinates as covariates we get a prediction surface plot showing the comparative influence of the latitude and the longitude as seen in Figure 14.

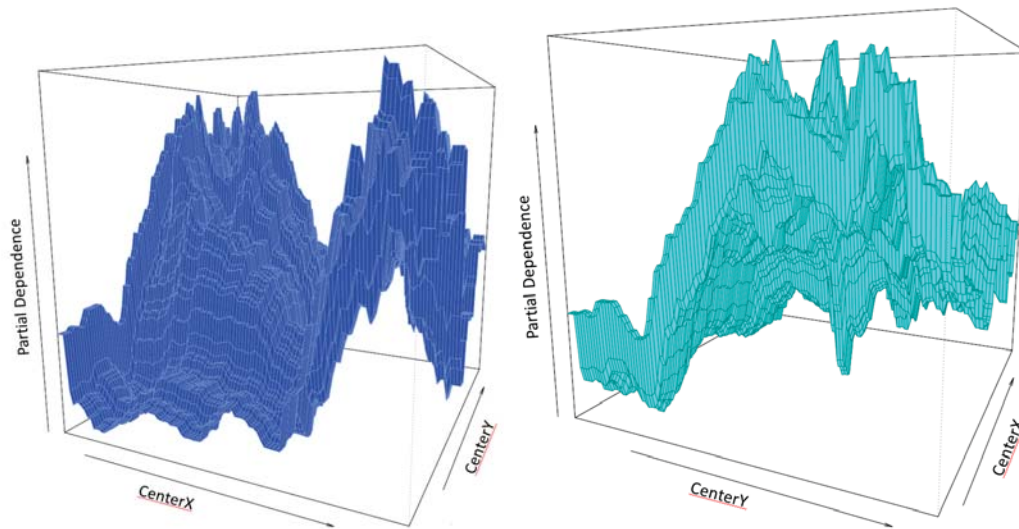


Figure 14. Prediction surface of 1990 in the resolution of 3000m.

3.2.4. Marginal influence plots

Marginal plots help in understanding the interaction effects of two variables by displaying the marginal relationship between the predicted aggregated fractions and the observed values of the test data set. Marginal plots also provide useful diagnostic information about the fitted model.

Figure 15 shows the marginal plots for the best model fit in the year 1990. The plots indicate that the BRT model under-predicts high observed values throughout all resolutions. This is especially apparent in the longer tails of the right-skewed histogram and density curves shown on the observed axis. In general, all plots exhibit a positive and relatively strong relationship, with a tendency towards clustering at the predicted values as seen by the vertical multi-modal histogram and density plot on the predicted axis. This is especially true in the resolution of 12000m where three clusters are evident, whereas in the resolution of 1500m it seems there is more smoothing present. This is a feature of the BRT design, as described in Section 2.5.

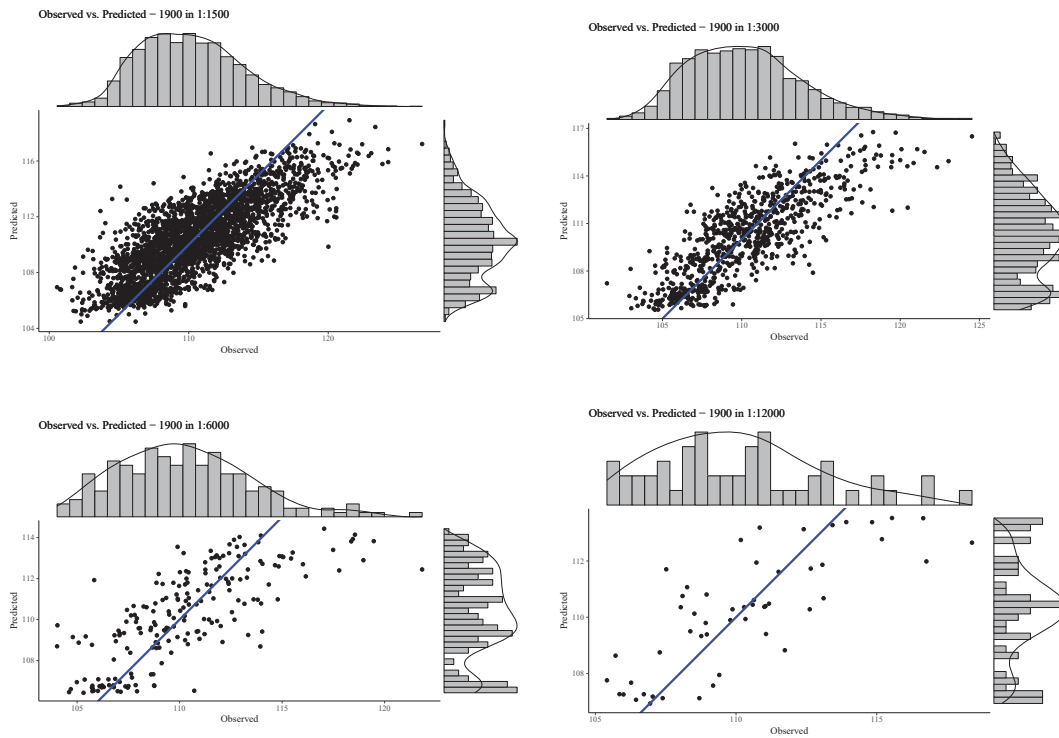


Figure 15. Marginal plots in the four different spatial aggregation resolutions showing the predicted PV fractions on the y-axis and the observed values on the x-axis.

3.3. Aggregation and scaling error

We investigated the effect of spatial aggregation on prediction accuracy and compared the predictive outcome to the computational time to extract aggregated means out of the FCover band for green vegetation. We argue that a full FCover scene is not required in order to achieve satisfying prediction results and therefore investigated finding a threshold of a spatial resolution that yields acceptable results but is also computationally inexpensive. For this, we recorded the elapsed time to generate the mean of the spatial grid cells and the time required to write the calculated mean to a csv file.

Figure 16 provides comparative information about computational time for the different spatial resolutions. The dominant factor in computing time was extracting the aggregated means from the FCover band for PV. The resultant values are and depicted in Figure 16.

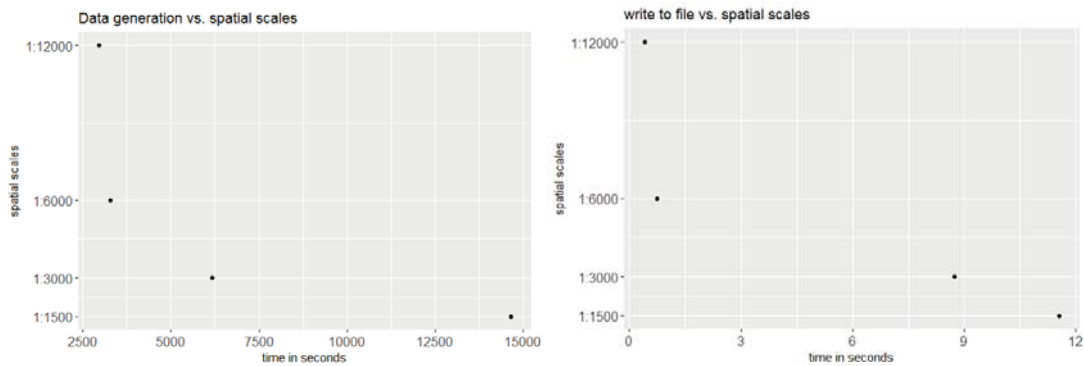


Figure 16. Comparative information about computational time on a) the delineation of green vegetation out of FCover imagery and b) writing to a csv file that will be used as an input file for BRT modelling.

The effect of the four aggregation resolutions on the prediction accuracy is depicted in Figure 17. For this plot, we calculated the absolute difference between the observed and the predicted values for all the years present in this case study. The largest and smallest error rates were observed at the resolution of 12000m, whereas the resolutions of 3000m and 6000m showed the most stable performance. It should be noted that the resolution of 1500m also yielded lower absolute error rates, but there was a trend towards higher rates in 1988 and 1990. Overall, the 3000m resolution showed the best error rates, followed by 6000m. These resolutions also have a reasonable processing time as shown in Figure 16.

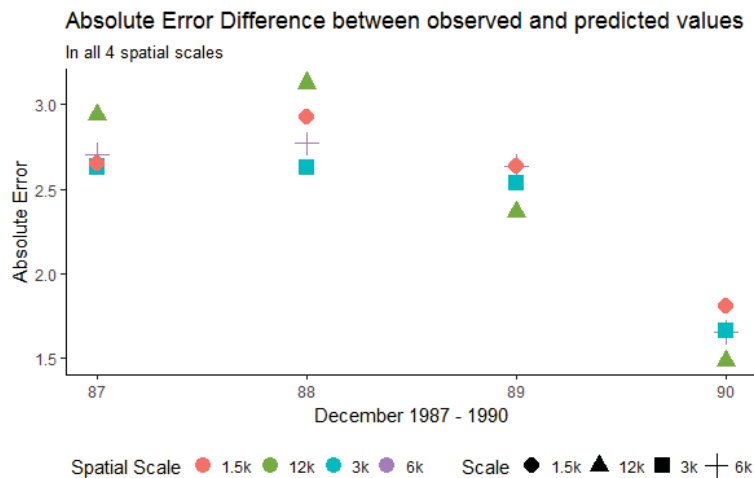


Figure 17. Quantitative assessment of absolute error rates in the four spatial aggregations.

The overall conclusion based on the inspection of the times is that the resolution of 3000m is best, followed by 6000m with regards of processing time, prediction accuracy, the strong and positive interaction effect shown in the marginal plots and a significant relative influence of the contribution of CenterY in the splitting process of min 52% and max 75%.

4. Discussion

The goal of this paper was to investigate how spatial aggregation affects prediction accuracy of green vegetation using a BRT model. We focused our evaluation on a case study and chose four aggregation schemes that follow a linear scale. Aggregating the fractions of green vegetation

and calculating the mean does alter the original fractions of PV. This alteration can be seen as the consequence of data compression. In our case, we introduced a compression that causes loss since the original fractions cannot be recovered by decompression. The results show that it is not necessary to compute FCover at full (30m) spatial resolution to obtain satisfactory predictions. This is an important outcome since the computational time will be significantly reduced by spatially aggregating the fractions of the FCover scene. Figure 16 shows the reduction of time needed for the data generation process in extracting the means. This is especially important when more than one FCover scene is used. However, comparisons between aggregations are not straightforward, particularly because the data quality (large data gaps) and green vegetation cover differs between the scenes. Further investigation is still necessary to test BRT on homogeneous land to assess whether the best spatial aggregation resolution identified here as 6000m is still the same and whether the prediction accuracy is affected by a different topography. We demonstrated that the BRT outperformed the LM by achieving much better RMSE rates.

In this paper, we first demonstrated that the distribution of residuals around the mean is relatively consistent throughout the resolutions. Moreover, in our study latitude and longitude coordinates alone were shown to be able to effectively predict FCover. We showed the strong relationship between latitude and longitude in the marginal plots in Figure 15.

In the relative influence plots we demonstrated that the centroid of the latitudes (indicating North-South direction) are far more dominant in describing the aggregated FCover mean values. For reasons discussed in Section 2.2 it is not surprising that the latitude dominates over longitude with regard to green vegetation. What is surprising though is the high contribution and very strong influence of around 80% as shown in Figure 12.

The marginal plots illustrate that BRT under-predicts high peak values throughout all resolutions. We argue that 72 FCover mean values of 12000m can represent the existing green vegetation in one scene. Further, we could demonstrate that the scene offered enough heterogeneous land cover and the Landsat footprint of 185 x 185km was sufficient to show the targeted and generalisable results.

Another interesting investigation would be to use multi-sensory imagery and multi-granularity pixel sizes as additional covariates in the modelling process. We focused on the exact alignment of pixel edges to the spatial grid by choosing a resolution that incorporates full pixels. However, before we used the resolutions here, we had more common ones in 1000m, 5000m and 10000m and the data extraction time was significantly increased due to the effect of incorporating adjacent pixels that overlapped with the spatial grid cell.

Limitations of our approach can be found in the aggregation scheme using the arithmetic mean. When extracting the mean of a spatial grid cell we don't know the distribution of fractions within the grid cell since we only obtain one value representing the aggregated fractions. Different methods of aggregating values may provide better capture of cell statistics and data structure within a spatial grid cell.

An alternative way of using FCover fractions is to sum all the vegetation (nVP and VP) and compare those values with the fraction of bare soil in a presence/absence study. This could be useful in time series analysis, such as an investigation of an increase or decrease of vegetation versus bare soil. This is of particular interest with ongoing climate change towards desertification in arid or semi-arid areas. A potential approach is to use indicator functions that can encode logical and simple calculations by defining thresholds in order to investigate if fractions of the combined vegetation versus bare soil represent values greater than the set threshold of both classes. Depending on the

magnitude of a fraction, the pixel could be mapped to categorical values used in the modelling process instead of our approach using continuous values. BRT can deal with continuous and categorical response values.

There are many features of BRT that are advantageous for the problem considered here. The BRT model itself comprises a flexible regression structure with improved predictive performance effected through boosting. Boosting is an adaptive method for combining many simple models to give an improved predictive performance. In addition to the computational speed and accuracy of estimation, they can describe complex non-linearities and interactions between variables, accommodate missing data, include different types of input variables without the need for transformations or elimination of outliers, perform well in high-dimensional problems, and allow for different loss functions such as accurate identification of small areas of interest. Moreover, they can be visualised and interpreted easily, thus facilitating the translation of the analytic results to decision makers [44]. The predictive accuracy of BRT has been investigated both theoretically [42,43] and in various applications [51]. Although BRT models are complex, they can be summarized in ways that give powerful ecological insight, and their predictive performance is superior to most traditional modelling methods.

To sum up, BRT is a very flexible, statistical and hierarchical machine learning approach that can be used in various remote sensing aspects. In a study by Kotta [52] the author combined hyperspectral remote sensing and BRT to test their ability to predict macrophyte and invertebrate species cover in the optically complex seawater of the Baltic Sea and concluded that there is a strong potential for BRT in modelling aquatic species. Further, Jafari et al [5] evaluated the suitability and performance of BRT for soil mapping using a limited point dataset in an arid region of Iran. The performance was tested in two scenarios: (i) using only the DEM and remote sensing covariates and (ii) additionally using the geomorphology map. Results showed that the geomorphology map contributed importantly to the prediction accuracy. In addition, Colin et al [50] combined a collection of GIS shapefiles, remotely sensed imagery, and aggregated and interpolated spatio-temporal information to one input file that resulted in a structured but noisy input file, showing inconsistencies and redundancies. It was shown that BRT can process different data granularities, heterogeneous data and missingness. A comparison with two similar regression models (Random Forests and Least Absolute Shrinkage and Selection Operator, LASSO) showed that BRT outperforms these in this instance. Last but not least, Pittman [53] investigated coral reef ecosystems that are topographically complex environments and possess structural heterogeneity that influences the distribution, abundance and behaviour of marine organisms. They used BRT and LIDAR data that provided high resolution digital bathymetry from which the topographic complexity was quantified at seven spatial resolutions of 4, 15, 25, 50, 100, 200 and 300m [53]. They concluded that the combination of BRT and LIDAR has great utility in the future development of benthic habitat maps and faunal distribution maps to support ecosystem-based management and marine spatial planning.

5. Conclusions

A data reduction scheme on FCover showing only the green vegetation fractions, and using BRT to assess the influence of the data reduction on the predictive power of BRT is proposed in this paper. The first step of the proposed method aims to reduce the heterogeneous green vegetation cover through aggregation based on an evenly spatial grid that served as an overlay for the delineation of the green vegetation fractions. This was performed at four spatial resolutions of 1500m, 3000m, 6000m and 12000m and resulted in 16 input files for the BRT modelling approach. The files were split into training and test set and BRT was then applied to identify the influence of the spatial resolution on prediction accuracy for BRT models. To validate the performance of the proposed method, the RMSE, MAE and MDAE were considered. Further, the predictive performance of the BRT was compared with that of the more common linear regression model and was found to consistently deliver smaller RMSE

values at all four spatial aggregations. The analysis showed that the proposed method can also provide useful visual interpretations by showing, for example, the prediction raster map and the smoothing factor for each spatial resolution. Based on these results, we conclude that boosted regression trees are an appealing method for estimating green vegetation from remotely sensed images and that an appropriate aggregation scale can be identified that balances computational demand with acceptable loss of predictive accuracy.

6. Acknowledgments

This study was supported by the Australian Research Council (Grant No.:FL150100150). The authors wish to thank (i) Department of Environment and Science (DES) for providing the Fractional Cover data, (ii) Brodie Lawson for helpful comments on the predictive performance of BRT on peak observations, (iii) QUT ACEMS for providing office space and infrastructure to achieve this article and (iiii) the Reviewers and the Editors for their constructive comments and helpful suggestions, which have greatly improved this article.

Author Contributions

Brigitte Colin and Kerrie Mengersen conceived and designed this study; Brigitte Colin performed the experiments and wrote the paper; Kerrie Mengersen, Michael Schmidt, Sam Clifford and Alan Woodley analysed the experimental results and edited the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

7. Supplementary Material

7.1. Mathematical explanation of the BRT method

Here, we summarise the method, following Friedman [37]. Consider a response variable y and a vector of predictor variables x that are connected via a joint probability distribution $P(x, y)$. Using a training sample $\{(x_1, y_1), \dots, (x_n, y_n)\}$ of known values of x and corresponding values of y , the goal is to find an approximation $F(x)$ to a function $F^*(x)$ that minimises the expected value of a loss function $\psi(y, F(x))$. Boosting approximates $F^*(x)$ by an additive expansion. The parameters $\{a_m\}_0^M$ and the expansion coefficients are jointly fit to the training data. This is done in a forward stage wise manner. Gradient Boosting [37] approximately solves differentiable loss functions $\psi(y, F(x))$ with a two step procedure. First, the function $h(x; a)$ is fit by least squares to the current pseudo-residuals which represent the residuals from the given stage of the tree building.

Then, given $h(x; a_m)$, the optimal value of the coefficient β_m is calculated via

$$\beta_m = \arg \min_{\beta} \sum_{i=1}^N \psi(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)). \quad (2)$$

Thus, at each iteration m , the tree partitions the feature space into L disjoint regions $\{R_{lm}\}_{l=1}^L$ and predicts a constant value, \bar{y}_{lm} , in each region. Gradient Boosting proceeds in this way until the base learner $h(x; a)$ is an L terminal node regression tree.

The parameters of the estimated tree are the splitting variables and corresponding split points that define the tree, and this defines the corresponding regions $\{R_{lm}\}_1^L$ of the partition at each iteration. These are accomplished in a hierarchical top-down approach using a least squares splitting measure [37]. Equation 2 can be solved individually within each region, R_{lm} defined by the corresponding

terminal node l of the m th tree. Because the tree predicts a constant value \bar{y}_{lm} within each region, R_{lm} , the solution to 2 reduces to a simple location estimate based on the criterion ψ

$$\gamma_{lm} = \arg \min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma). \quad (3)$$

First, we initialize the model. To minimize the square error we initialise $F^*(\mathbf{x})$ with the mean of the training set that is defined through $\{y_i, \mathbf{x}_{i1}\}_i^N$ and the learning rate γ . At the beginning of the algorithm we specify the number of trees/iterations shown as m in the for-loop control structure. Friedman [37] added a stochastic element by proposing to draw a random subsample from the full training data set without replacement. This subsample is then used to fit the base learners and compute the model update for the current iteration. The random subsample of size $\tilde{N} < N$ is given by $\{y_{\pi(i)}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}}$. Adding randomness to the algorithm in this way has been shown to improve the performance of gradient boosting [44]. In the last step of the algorithm the current approximation of F_{m-1} is updated in each corresponding region R_{lm} .

7.2. Limitation

Another limitation is that the absolute error between model prediction and actual observation contains the model error due to the spatial aggregation as demonstrated in Table 5. The error rates are similar and it is difficult to separate those from each other.

7.3. Partial Dependency Plots

[H] Partial dependency plots (PDP) are graphical visualizations of the marginal effect of a given variable (or multiple variables) on an prediction outcome.

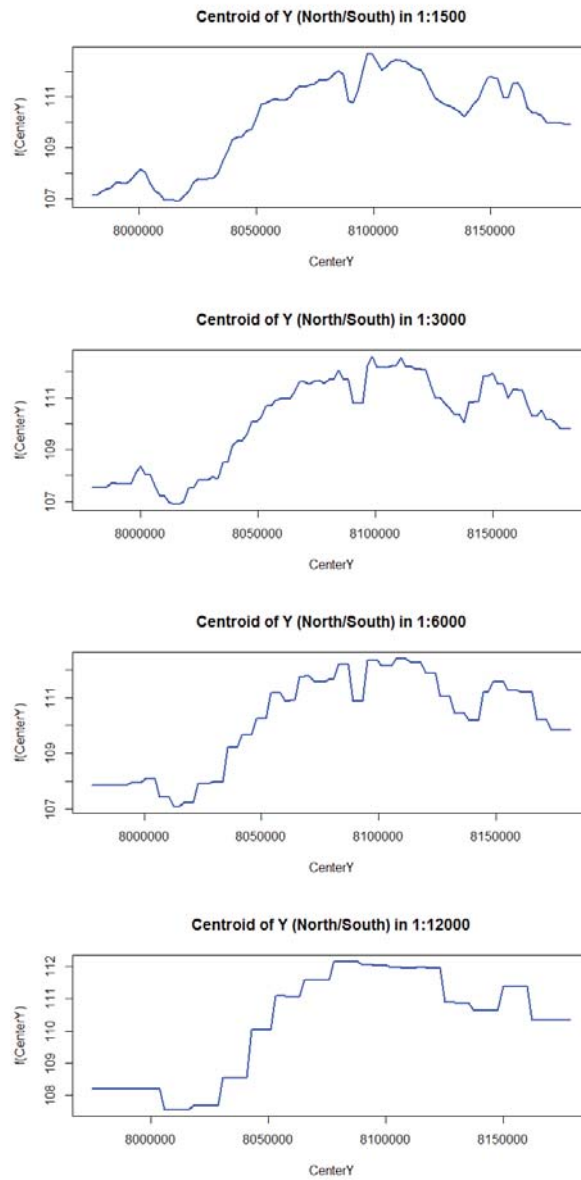


Figure 18. Partial dependency plots in 1500m to 12000m from top to bottom of best year 1990. The effect of latitude (CenterY) show similar patterns for the resolution of 3000m and 6000m, with a steep linear increase in FCover after a threshold latitude.

References

1. Datt, B. A New Reflectance Index for Remote Sensing of Chlorophyll Content in Higher Plants: Tests using Eucalyptus Leaves. *Journal of Plant Physiology* **1999**, *154*, 30–36.
2. Schmidt, M.; Thamm, H.P.; Menz, G.; Bénes, T.E. Long term vegetation change detection in an and environment using LANDSAT data. *Geoinformation for European-Wide Integration*, Millpress, Rotterdam **2003**, p. 145–154.
3. Marsett, R.C.; Qi, J.; Heilman, P.; Biedenbender, S.H.; Watson, M.C.; Amer, S.; Weltz, M.; Goodrich, D.; Marsett, R. Remote Sensing for Grassland Management in the Arid Southwest. *Rangeland Ecology & Management* **2006**, *59*, 530–540.
4. Huete, A.; Ponce-Campos, G.; Zhang, Y.; Restrepo-Coupe, N.; Ma, X.; Susan Moran, M. *Monitoring Photosynthesis From Space*; 2015; pp. 3–22.
5. Jafari, A.; Khademi, H.; Finke, P.A.; Van de Wauw, J.; Ayoubi, S. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. *Geoderma* **2014**, *232–234*, 148–163.
6. Anderson, M.C.; Allen, R.G.; Morse, A.; Kustas, W.P. Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. *Remote Sensing of Environment* **2012**, *122*, 50–65.
7. Washington-Allen, R.; Van Niel, T.; Ramsey, R.; West, N. Remote Sensing-Based Piosphere Analysis. *GIScience & Remote Sensing* **2004**, *41*, 136–154.
8. Stohlgren, T.J.; Ma, P.; Kumar, S.; Rocca, M.; Morisette, J.T.; Jarnevich, C.S.; Benson, N. Ensemble habitat mapping of invasive plant species. *Risk Analysis* **2010**, *30*, 224–235.
9. Lowell, K. A socio-environmental monitoring system for a UNESCO biosphere reserve. *Environmental Monitoring and Assessment* **2017**, *189*, 601.
10. Sarker, C.; Alvarez, L.M.; Woodley, A. Integrating Recursive Bayesian Estimation with Support Vector Machine to Map Probability of Flooding from Multispectral Landsat Data. *International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016*, **2016**.
11. J.Walsh, S.; Crawford, T.W.; Welsh, W.F.; A.Crews-Meyer, K. A multiscale analysis of LULC and NDVI variation in Nang Rong district, northeast Thailand. *Agriculture, Ecosystems & Environment* **2001**, *85*, 47–64.
12. Gallo, K.P.; Easterling, D.R.; Peterson, T.C. The Influence of Land Use/Land Cover on Climatological Values of the Diurnal Temperature Range. *Journal of Climate* **1996**, *9*, 2941–2944.
13. Wulder, M.A.; Masek, J.G.; Cohen, W.B.; Loveland, T.R.; Woodcock, C.E. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment* **2012**, pp. 2–10.
14. Zhang, H.; Li, Q.Z.; Lei, F.; Du, X.; Wei, J.D. Research on rice acreage estimation in fragmented area based on decomposition of mixed pixels. *Remote Sensing and Spatial Information Sciences* **2015**, *40*, 133.
15. Guerschman, J.P.; Scarth, P.F.; McVicar, T.R.; Renzullo, L.J.; Malthus, T.J.; Stewart, J.B.; Rickards, J.E.; Trevithick, R. Assessing the effects of site heterogeneity and soil properties when unmixing photosynthetic vegetation, non-photosynthetic vegetation and bare soil fractions from Landsat and MODIS data. *Remote Sensing of Environment* **2015**, *161*, 12–26.
16. Adams, J.B.; Sabol, D.E.; Kapos, V.; Filho, R.A.; Roberts, D.A.; Smith, M.O.; Gillespie, A.R. Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the Brazilian Amazon. *Remote Sensing of Environment* **1995**, *52*, 137–154.
17. Roberts, D.A.; Smith, M.A.J. Green vegetation, nonphotosynthetic vegetation, and soils in AVIRIS data. *Remote Sensing of Environment* **1993**, *44*, 255–269.
18. Zachary, T.; Dar, R.; Sander, V.; Angeles, C.; Carlos, R.; Susan, U. Evaluating Endmember and Band Selection Techniques for Multiple Endmember Spectral Mixture Analysis using Post-Fire Imaging Spectroscopy. *Remote Sensing* **2018**, *10*.
19. Scarth, P.F.; Röder, A.; Schmidt, M. Tracking Grazing pressure and climate interaction - The Role of Landsat Fractional Cover in time series analysis. *Proceedings of the 15th Australasian Remote Sensing and Photogrammetry Conference* **2010**, p. 13.
20. Scanlon, T.M.; Albertson, J.D.; Caylor, K.K.; Williams, C.A. Determining land surface fractional cover from NDVI and rainfall time series for a savanna ecosystem. *Remote Sensing of Environment* **2002**, *82*, 376–388.

21. Held, A.; Phinn, S.; Soto-Berelov, M.; Jones, S. *AusCover Good Practice Guidelines: A technical handbook supporting calibration and validation activities of remotely sensed data product*; Vol. Version 1.2, TERN AusCover, 2015.
22. Muir, J.; Schmidt, M.; Tindall, D.; Trevithick, R.; Scarth, P.; Stewart, J. *Field measurement of fractional ground cover: a technical handbook supporting ground cover monitoring for Australia*; Queensland Department of Environment and Resource Management for the Australian Bureau of Agricultural and Resource Economics and Sciences, Brisbane, Australia, 2011.
23. Bastin, G.; Scarth, P.; Chewings, V.; Sparrow, A.; Denham, R.; Schmidt, M.; O'Reagain, P.; Shepherd, R.; Abbot, B. Dynamic reference cover method to separate grazing and rainfall effects on rangeland ground cover. *Remote Sensing of Environment* **2012**, *121*, 443–457.
24. Carroll, C.; Waters, D.; Vardy, S.; Silburn, M.; Attard, S.; Thorburn, P.; Davis, A.; Halpin, N.; Schmidt, M.; Wilson, B.; Clark, A. A Paddock to reef monitoring and modelling framework for the Great Barrier Reef: Paddock and catchment component. *Marine Pollution Bulletin* **2012**, *65*, 136–49.
25. Schmidt, M.; Amler, E.; Guerschmann, P.; Scarth, J.; Behn, K.; Thonfeld, F. Fractional Vegetation Cover of East African wetlands observed on ground and from space. May 2016.
26. Trevithick, R.; Soto-Berelov, M.; Jones, S.; Held, A.; Phinn, S.; Armston, J.; Bradford, M.; Broomhall, M.; Cabello, A.; Chisholm, L.; Clarke, K.; Davies, K.; Farmer, E.; Flood, N.; Gill, T.; Guerschman, J.; Hacker, J.; E. Howorth, J.; Hueni, A.; Youngentob, K. *AusCover Good Practice Guidelines: A technical handbook supporting calibration and validation activities of remotely sensed data products*; 2015.
27. A. C. Cressie, N. Change of Support and The Modifiable Areal Unit Problem **1996**. *3*, 159–180.
28. Ershadi, A.; McCabe, M.; Evans, J.; Walker, J. Effects of spatial aggregation on the multi-scale estimation of evapotranspiration. *Remote Sensing of Environment* **2013**, *131*, 51 – 62.
29. Schucknecht, A.; Meroni, M.; Kayitakire, F.; Boureima, A. Phenology-Based Biomass Estimation to Support Rangeland Management in Semi-Arid Environments. *Remote Sensing* **2017**, *9*, 463.
30. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *Journal of Animal Ecology* **2008**, *77*, 802–813.
31. Paruelo, J.M.; Lauenroth, W.K. Relative Abundance of Plant Functional Types in Grasslands and Shrublands of North America. *Ecological Applications* **1996**, *6*, 1212–1224.
32. McNab, H.W.; Lloyd, T.F.d. Testing Ecoregions in Kentucky and Tennessee with Satellite Imagery and Forest Inventory Data. 2009.
33. Chen, H. *Köppen climate classification*. Retrieved from <http://hanschen.org/koppen/>, 2017.
34. Bureau of Meteorology. Climate Classification of Australia, 2016.
35. Australia, G. Fractional Cover (FC25) Product Description. Technical report, Australian Government, 2015.
36. Scarth, P.; Byrne, M.; Danaher, T.; Henry, B.; Hassett, R.; Carter, J.; Timmers, P. State of the paddock: monitoring condition and trend in groundcover across Queensland. *13th Australasian Remote Sensing and Photogrammetry Conference (ARSPC)* **2006**, p. 11.
37. Friedman, J. Recent Advances in Predictive (Machine) Learning. *Journal of Classification* **2006**, p. 175.
38. Ridgeway, G. Generalized Boosted Models: A guide to the gbm package. *Compute* **2005**, pp. 1–12.
39. Robinzonov, N. Advances in boosting of temporal and spatial models. PhD thesis, Ludwig-Maximilians-Universität München, 2013.
40. Tarling, R. *Statistical Modelling for Social Researchers: Principles and Practice*; Taylor & Francis Group: London and New York, 2009.
41. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Vol. 103, 2013; pp. 856–875.
42. Breiman, L. Arcing classifiers. *Annals of Statistics* **1998**, *26*, 801–849.
43. Freund, Y.; Schapire, Robert, E. Experiments with a New Boosting Algorithm. *International Conference on Machine Learning* **1996**, pp. 148–156.
44. Friedman, J.H. Stochastic gradient boosting. *Computational Statistics and Data Analysis* **2002**, *38*, 367–378.
45. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, second ed.; Springer Series in Statistics: New York, 2009; pp. 337–387.
46. Matteson, A. *Boosting the accuracy of your Machine Learning models*. Retrieved from <https://www.datasciencecentral.com/profiles/blogs/boosting-the-accuracy-of-your-machine-learning-models>, 2013.

47. Kuhn, M. The caret Package. *Journal of Statistical Software* **2008**, *5*, 1–10.
48. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
49. Adams, M. *Generalized Boosted Models: A guide to the gbm package*. Retrieved from <https://cran.r-project.org/web/packages/lm.br/index.html>, 1991.
50. Colin, B.; Clifford, S.; Wu, P.; Rathmanner, S.; Mengersen, K. Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making. *Open Journal of Statistics* **2017**, *07*, 859–875.
51. Tsangaratos, P.; Ilia, I. Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. *Landslides* **2016**, *13*, 305–320.
52. Kotta, J.; Kutser, T.; Teeveer, K.; Vahtmäe, E.; Pärnoja, M. Predicting Species Cover of Marine Macrophyte and Invertebrate Species Combining Hyperspectral Remote Sensing, Machine Learning and Regression Techniques. *PLoS ONE* **2013**, *8*, 1–11.
53. Pittman, S.J.; Costa, B.M.; Battista, T.A. Using Lidar Bathymetry and Boosted Regression Trees to Predict the Diversity and Abundance of Fish and Corals. *Journal of Coastal Research* **2009**, *10053*, 27–38.

© 2019 by the authors. Submitted to MDPI for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

5 Analysis of spatial smoothing effects on green vegetation to improve prediction accuracy using Boosted Regression Trees

Preamble

This chapter addresses the fourth research objective and we investigate when spatial smoothing is incorporated into a BRT what effect this has on the model fit, estimation of green vegetation fractions and prediction accuracy. Spatial smoothing accounts for spatial autocorrelation by smoothing the estimates of a spatial random field towards the arithmetic mean of the adjacent pixels around a centre point of the neighbourhood matrix. The neighbourhood is defined in the form of matrix of equal weights. The statistical model developed in this chapter can be viewed as an extension of the model developed in chapter 2.

Statement for Authorship

This chapter has been written as a journal article. The authors listed below have certified that:

- (a) They meet the criteria for authorship as they have participated in the conception, execution or interpretation of at least the part of the publication in their field of expertise;
- (b) They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (c) There are no other authors of the publication according to these criteria;
- (d) Potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of the journals of other publications and the head of the responsible academic unit; and
- (e) They agree to the use of the publication in the student's thesis and its publication on the Australian Digital Thesis database consistent with any limitations set by publisher requirements.

The reference for the publication associated with this chapter is; **Student Brigitte Colin**, Other supervisors (2018). Analysis of focal neighbourhood effects on green vegetation pixels to improve prediction accuracy using Boosted Regression Trees. Draft version.

Contributor	Statement of contribution
Student Brigitte Colin	Conduct the research, develop code for the statistical approach, write the manuscript
Signature and date:	
Kerrie Mengersen	Propose and supervise research, comments on manuscript.

Principal Supervisor Confirmation: I have sighted email or other correspondence for all co-authors confirming their authorship.

Name: K Mengersen Signature: QUT Verified Signature Date: _____

5.1 Introduction

Spatial data are becoming increasingly prevalent in many domains. Many current methods for analysing spatial data use smoothing kernels such that the number of parameters are fixed a-priori and hyperparameters, such as the width of the neighbourhood specifications are pre-defined. In this paper a non-parametric Gaussian Process (GP) is proposed. GP are used for mathematical and statistical modelling such as kriging (Karnieli et al., 2008) and in supervised machine learning approaches (Ak, Ergönül, Şencan, Torunoğlu, & Gönen, 2018). In order to improve the signal-to-noise ratio in fractional cover (FCover) data, spatial techniques can be used.

In many spatial models, of image data represented as pixels, a neighbourhood matrix (also known as spatial filter) is often defined in terms of cells gathered around a midpoint that share the same edges as the centre pixel (also known as the centroid) or are within a specified boundary of the centroid (Owen, 1984). A moving window smoothing kernel applies an aggregation function to all cell values within the specified neighbourhood matrix, calculates the mean and creates a new smoothed output value for the centroid, and moves on to the next cell. The simplest smoothing technique uses unweighted averaging over a fixed neighbourhood size such as the Box filter (Lopes, Touzi, & Nezry, 1990). Smoothing is sometimes referred to as filtering, because smoothing has the effect of suppressing high frequency signal (fast variations) and enhancing low frequency signal (slow variations). The non-linear Gaussian Process provides an alternative method to smoothing, whereby the smoothness of the (non-linear) regression is defined by a covariance function that ensures that values that are close together will produce a function that matches our data. This covariance function, along with a mean function defines a Gaussian Process.

Boosted Regression Trees (BRT) are a popular non-parametric decision tree method in statistical and machine learning approaches (Freund & Schapire, 1996; J. Friedman, 2006; J. H. Friedman, 2002). BRT works very well with large datasets, allowing for inconsistencies, missing data, and different data granularities allowing a variety of distributions (Elith & Leathwick, 2017). Moreover, BRT can deal with skewed and multi-model distributions as well as categorical data. However, past applications of the method for spatial data have generally ignored spatial autocorrelation, i.e., that locations closer to each other exhibit similar characteristics than those further apart and are related through their geographic location (Dubin, 1988; F. Dormann et al., 2007; Miller, 2004).

In this paper we focus on a challenging example of spatial data analysis, namely prediction of green vegetation using remotely sensed data showing heterogeneous land cover characteristics that violate the key assumptions of standard statistical analyses, that each random variable is independent and identically distributed (i.i.d) (F. Dormann et al.,

2007; Jacobs, 1992).

To address those challenges we perform a Gaussian smoothing kernel approach to cope with green vegetation cover which is changing in space, and is also corrupted by random noise. By using smoothing kernels we capitalise on the assumption that nearby points measure nearly the same underlying value and averaging can reduce the level of noise without unduly biasing the value obtained (Miller, 2004). The drawback is a loss of spatial variation due to homogenising local characteristics. By applying a Gaussian smoothing kernel on our Fraction Cover (FCover) data a new raster data set will be calculated showing the smoothed values of green vegetation cover. In our case study we investigate four different sizes of the Gaussian smoothing kernel and the effect of this choice on prediction accuracy when fed into the BRT as the new response variable .

5.2 Material

5.2.1 Case Study

Our case study is located in the Northern Territory, Australia and shows a heterogeneous topology of native grass types. The prediction and quantitative estimation of biomass is of primary interest. We use FCover scene of the Landsat footprint of path 102 row 72 at the Worldwide Reference System-2 (WRS-2) that cover the extent of 185 x 185 km of commercial grazing land and contain over 50 million pixels. Our study area is defined as “dry” with variations of “desert, hot arid” and “dry summer, hot arid” (BWh and Bsh) based on the Koeppen-Geiger scheme and is very vulnerable (Bureau of Meteorology, 2016; Chen, 2017). The highest point is 255 m and the lowest is located at 23 m (232 m) above mean sea level (MSL) referenced to the Australian Height Datum (AHD). There is a constant increase of about 15 % of the elevation between the upper right corner (min 23 m) to the lower left corner (max 255 m).

5.2.2 Fractional Cover Data

Landsat satellite data offer great benefits for monitoring vegetation (Gallo, Easterling, & Peterson, 1996; J.Walsh et al., 2001) since one Landsat footprint covers an area of 185 x 185 km^2 , is freely available (Wulder et al., 2012) and it avoids expensive, extensive and often impractical in-situ measurement. One Landsat pixel covers an area of 30 x 30 m^2 on the ground and represents the reflected or emitted radiation from several different objects on the Earth’s surface. The combined individual spectra of objects represented in one pixel cannot discriminated anymore and results in mixed pixel or Mixel (Kamal & Phinn, 2011; Malenovský et al., 2007; Zhang, Li, Lei, Du, & Wei, 2015). A spectral unmixing approach separates the spectral reflectance of one pixel into its single ground

cover components to determine the proportions of each of the three endmembers as graphically demonstrated in Figure 5.1. In our case the endmembers are photosynthetic/green vegetation (PV), non-photosynthetic vegetation (nPV) and bare soil (BS). Further information about how to derive spectral endmembers using a spectral unmixing approach is provided in (Kamal & Phinn, 2011; Scarth et al., 2006).

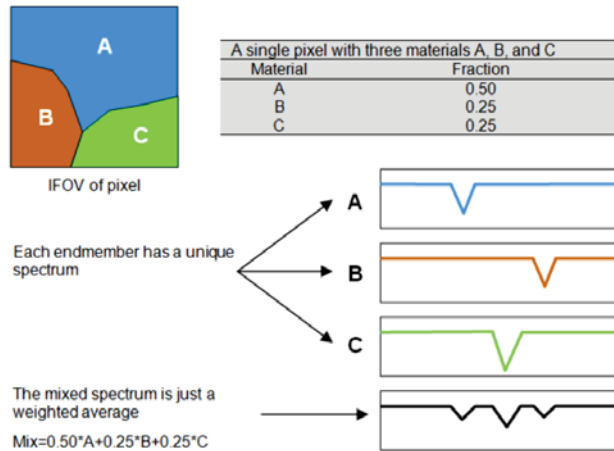


Figure 5.1: Spectral unmixing approach explained graphically using the Instantaneous Field of View (IFOV) as the geometric resolution of one FCover pixel where the spectral information and the fractions of three objects on the ground are combined together. The spectral unmixing approach aims to separate the unique reflected or emitted radiations and to derive three map layers for each endmember, here photosynthetic/green vegetation (PV), non-photosynthetic vegetation (nPV) and bare soil (BS) (Kamal & Phinn, 2011).

Our FCover data set is geo-referenced to the Worldwide Reference System 2 (WRS-2) and the pixels are described using geographic coordinates in latitude and longitude. The two-dimensional feature space is represented through a x-value and a y-value, where x describes the East-West direction (latitude) and y describes the North-South direction (longitude). The pair of latitude and longitude are numeric values, representing a unique location within our study area showing green vegetation fraction present in each individual pixel.

5.3 Research design

We used two approaches to developing a non-linear BRT algorithm to predict green vegetation fractions. First, we investigated how the prediction accuracy is affected using the non-smoothed and smoothed green vegetation fractions as alternating response variables on latitude and longitude. Second, we added the smoothed green vegetation fractions as additional covariates to the latitude and longitude and extended the list of input variable in the BRT model. To smooth the green vegetation fractions, a non-linear covariate-driven

Gaussian Process was applied as a smoothing kernel to the FCover data. We used four different kernel sizes with varying widths of the neighbourhood matrix around the centroid pixel to define those pixels that were to be incorporated in the Gaussian smoothing process. Then we compared the prediction results and goodness of model fit using the RMSE.

5.3.1 Neighbourhood analysis

The application of Gaussian smoothing kernels on the green vegetation fractions provides an output raster where the value for each output cell is a function of the values of all the input cells that are in the specified neighbourhood. It calculates the average of all green vegetation fractions encountered in the defined kernel size via a neighbourhood matrix. The centroid for which the average is being calculated is referred to as the processing cell. The value of the processing cell, as well as all the cell values in the identified neighbourhood, is included in the neighbourhood statistics calculation. The neighbourhoods can overlap so that cells in one neighbourhood may also be included in the neighbourhood of another processing cell. Missing values are excluded in the calculation. In our case we use a low-pass or smoothing filter that removes extreme values. By contrast, high-pass filters accentuate features (Kumar, 2013).

5.3.1.1 Gaussian smoothing kernels

The extent of the Gaussian smoothing kernel is defined by a parameter sigma (σ). The smoothing parameter σ is defined by a bandwidth and the extent of the bandwidth controls the smoothing effect. Smoothing is also called low pass filtering since it removes high spatial frequency noise and fine details. The Gaussian smoothing kernel operates like a moving window that affects each individual fractional cover value at a time. For our case study we defined four different kernel sizes and applied those on the observed green vegetation fractions: $\sigma = 0.2$ performs the most smoothing, $\sigma = 1$ and $\sigma = 10$ both will apply moderate smoothing and $\sigma = 20$ will perform the least smoothing. When using Gaussian smoothing kernels the value of σ defines the range of a pixel's ability to move, therefore a larger σ captures more variance and patterns in the data and the magnitude of the smoothing approximate the original green vegetation fractions. Thus larger the value of sigma, the lower the smoothing effect (Kumar, 2013).

5.3.2 Gaussian Processes

Gaussian processes are stochastic process and aim to fit a multivariate normal distribution to the (continuous) FCover data. The Gaussian Process is often denoted as $f \sim GP(\mu, k)$, where the function f is distributed as a Gaussian Process with a mean function $\mu(x)$ and

covariance function $k(x, x)$ (Kumar, 2013; Wilson & Adams, 2013).

Using the Gaussian Process, one can find a distribution over the possible smooth functions $f(x)$ that are consistent with the observed data. We choose the covariance function such that values that are close together in input space will produce output values that are close together.

In a simple linear regression setting, we have a dependent variable y and we assume that it can be modelled as a function of an independent variable x , $y = f(x) + \epsilon$, where ϵ is the error rate. We assume further that the function f defines a linear relationship so the aim is to find the parameters θ_0 and θ_1 which define the intercept and slope of the regression line as $y = \theta_0 + \theta_1 x + \epsilon$.

In contrast, a Gaussian Process has relaxed assumption about the form of $f(x)$, allowing it to be distributed: $f(x) \sim GP(\mu(x), k(x, x'))$ where $\mu(x)$ is represented as $N(x) = 0$ and $k(x, x')$ is the kernel and $x_1, x_2, \dots, x_n \in \mathbb{R}^n$, $n \in \{1, 2, \dots\}$.

$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} \sim N \left(\underbrace{\begin{bmatrix} \mu(x_1) \\ \mu(x_1) \\ \vdots \\ \mu(x_1) \end{bmatrix}}_{\mu \in \mathbb{R}^n}, \sigma^2 \underbrace{\begin{bmatrix} k(x, x_1) & k(x, x_1) \\ k(x, x_2) & k(x, x_2) \\ \vdots & \vdots \\ k(x, x_n) & k(x, x_n) \end{bmatrix}}_{K \in \mathbb{R}^{n \times n}} \right) \quad (5.1)$$

The kernel $k(x_1, x_2) = \sigma^2 e^{-\frac{(x_1 - x_2)^2}{2l^2}}$, where l defines the standard deviation of the Gaussian kernel. If $l \sim 0$ we get a very complex function and risk overfitting, if $l \gg 0$ we get a very smoothed function and risk underfitting. An appropriate l can be determined by cross validation as part of hyperparameter tuning (Banerjee et al., 2008; Gelfand, Schmidt, Banerjee, & Sirmans, 2004; Wilson & Adams, 2013)

Further, the range of the values in the kernel is $0 \leq K_{(ij)} \leq 1$. Alternatively, we can select hyperparameters such that they maximise the likelihood that a certain choice of function produced the data. Due to the Gaussian nature of the GP, this likelihood has an explicit form,

$$\log \mathcal{L} = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2} \log(\det(K)) - \frac{1}{2} \log 2\pi \quad (5.2)$$

where the first term is the fit of the data, the second is the complexity and so maximising the likelihood balances between over- and underfitting (C. K. Williams & Rasmussen, 2006).

We use the Gaussian kernel as our covariance function, with the geographic coordinates as our covariates \mathbf{x} in a 2-dimensional case, defined by latitude and longitude. The Gaussian kernel function $k_{(G)}(.,.)$ between two green vegetation fraction pixels \mathbf{x}_i and \mathbf{x}_j can be defined as follows (Ak et al., 2018):

$$k_{(G)}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / s^2), \quad (5.3)$$

$$k_s(\mathbf{s}_l - \mathbf{s}_m) = k_{(G)}(\mathbf{s}_l - \mathbf{s}_m) \quad (5.4)$$

and the kernel width parameter is chosen as the mean of the Euclidean distances (Ak et al., 2018).

5.3.3 Boosted Regression Tree

Boosted Regression Tree (BRT) models are a combination of two techniques: decision tree algorithms and boosting methods. BRT repeatedly fit many basic decision trees to improve the accuracy of the model by taking a random subset of a training data set allowing for replacement. By building the subsequent trees BRT uses a boosting algorithm in which the drawn samples are weighted, where a higher weight will be given to poorly modelled decision trees. The higher weight is used in the next tree to prioritise the mis-specified observations. The boosting part of BRT takes into account the previous fits of the decision trees, and sequentially builds new models focusing on the errors to improve the overall prediction accuracy of BRT.

BRT have four hyperparameters that significantly impact the performance, namely (1) the shrinkage (how quickly the algorithm adapts); (2) tree complexity, the total number of trees in the final model (number of iterations); (3) interaction depth, the interaction between different nodes along the branch and minimum observations in node, and (4) minimum number of training set samples in a node to commence splitting (Kuhn, 2008, 2015; Kuhn & Johnson, 2013). For the BRT algorithm, we use the `gbm` R package (Ridgeway, 2005) with 10-fold cross validation, introducing randomness by setting the `bag.fraction` to 0.75, and using a Gaussian distribution. The following tuned hyperparameters as used by Kuhn (Kuhn, 2008, 2015) were:

- bagging fraction; 0.75
- shrinkage; 0.01

- tree complexity; 2500
- interaction depth; 5
- minimum number of training set samples in a node; 10
- all other values were held at their default values.

5.4 Results

5.4.1 Gaussian Kernel smoothing

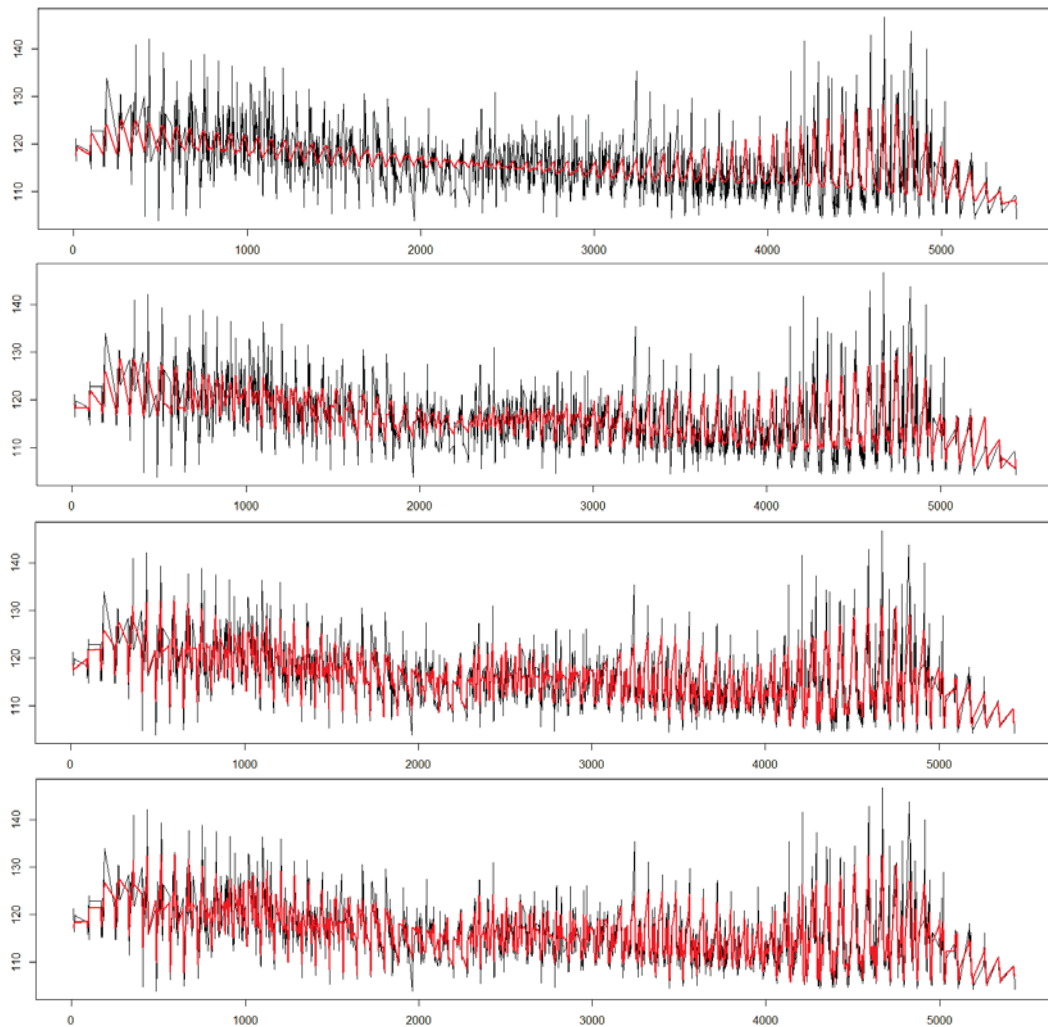


Figure 5.2: The magnitude of the smoothing is depicted in red on the original data in black using four kernel sizes from ranging from $\sigma = 0.2$ (most smoothing) at the top to $\sigma = 20$ (least smoothing) on the bottom. The y-axis shows the green vegetation fractions and the x-axis demonstrates their unique location ID's plotted as a 1-dimensional vector from top left to bottom right.

Figure 5.2 shows four smoothing lines as a 1-dimensional plot where the smoothed values are shown in red and the original green vegetation fractions are shown in black. We can see that the smaller the value of σ , the stronger the approximation of the kernel towards

the observed data. When σ is set to 0.2 the red smoothed line fails to capture the peak values and hence loses most of the local characteristics. The corresponding 2-dimensional smoothed outcomes depicted as raster maps are shown in Figure 5.4 where the kernel sizes with the defined values of σ show the effect or magnitude of smoothing. Figure 5.3 shows the original raster image prior to the smoothing. The data gap, shown in white, at the upper right side of the imagery results from data refinement procedures such as masking out obscuring elements like clouds and cloud shadows.

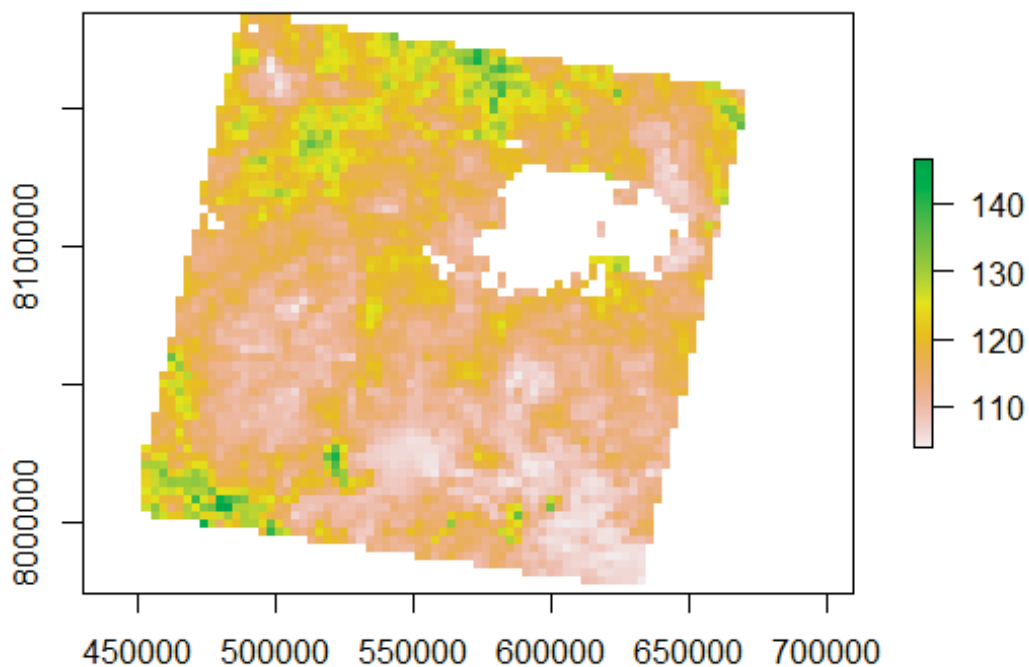


Figure 5.3: Original Raster image (not smoothed) showing the green vegetation fractions where higher values represent a higher fraction of green vegetation represented in the individual pixel.

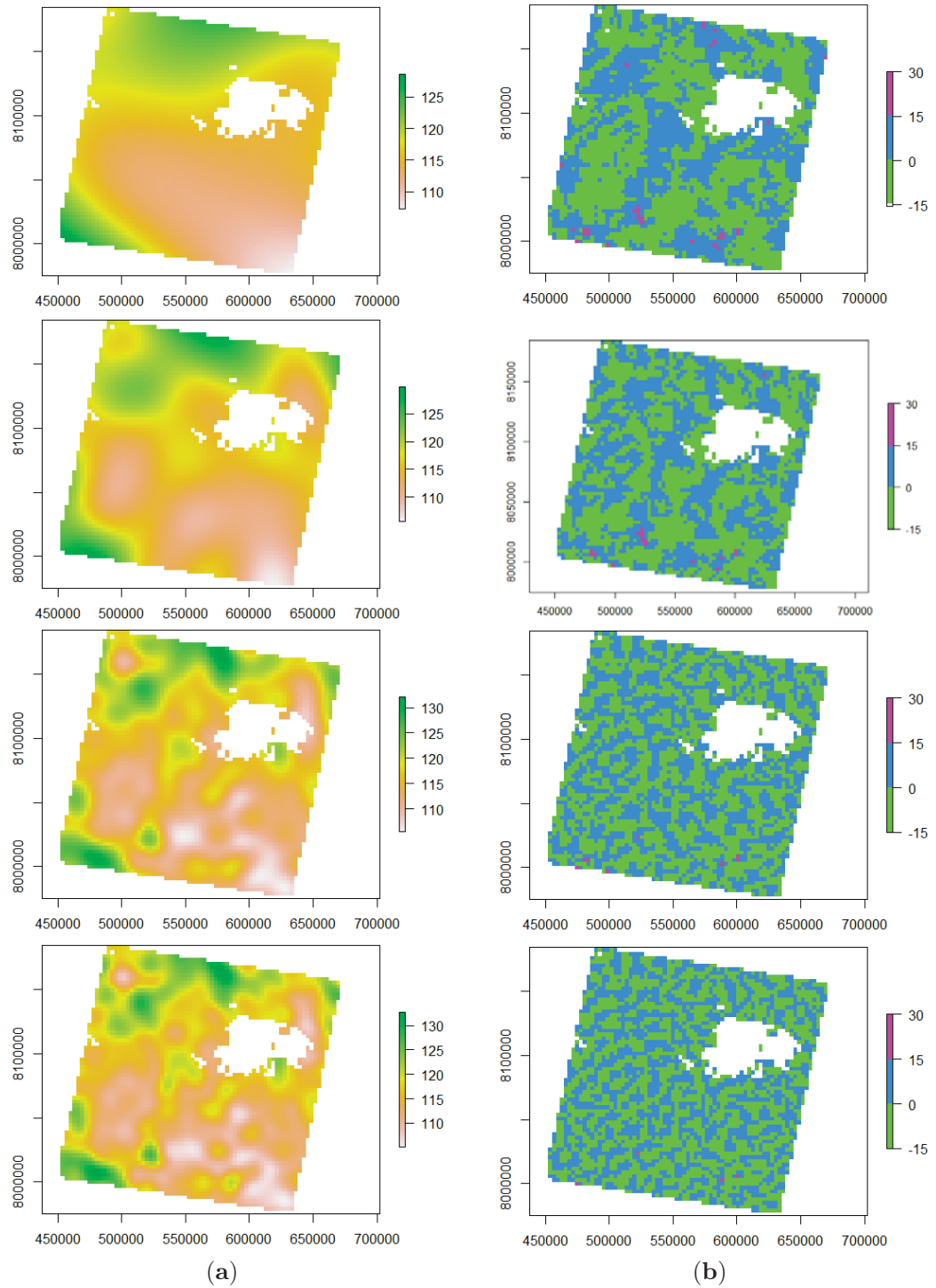


Figure 5.4: Plots of the Gaussian smoothing showing (a) smoothed raster maps on the left and (b) raster maps of the distribution of the residuals on the right. The top panel shows the kernel size $\sigma = 0.2$ (maximum smoothing), followed by $\sigma = 1$, the third shows the smoothing using $\sigma = 10$ and the last panel was performed using $\sigma = 20$ (the least smoothing).

By comparing the four smoothed raster maps, we can see that in the case of $\sigma = 0.2$ all local features are smoothed out whereas the raster map corresponding to a kernel size of

$\sigma = 20$ shows smoothed values where local features were maintained. The corresponding residual map shows that most noise has been smoothed or filtered out whereas in case of $\sigma = 0.2$ a pattern of distinctive and strong residuals is clearly visible. Further, local characteristics are maintained with $\sigma = 10$ and $\sigma = 20$ but are lost with $\sigma = 1$. In all residual plots we can see very distinctive features at the bottom of the imagery in pink demonstrating a high residual error around +20.

Figure 5.5 presents boxplots that describe the distribution of the smoothed green vegetation fraction in comparison with the original values on the very right. A boxplot describes the dispersion of the data via the upper and lower whiskers, the inter quartile range (IQR) and the median. The IQR, indicated as the grey box around the black median line in the middle is very similar throughout all boxplots regardless of the size of the Gaussian smoothing kernel. It can be clearly seen that the smoothing eliminated the outliers.

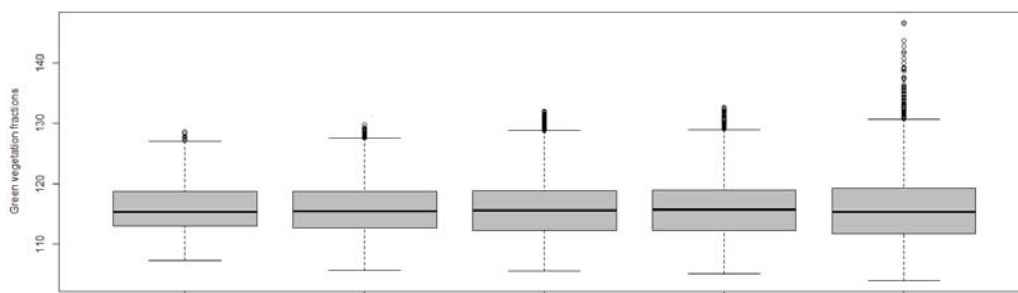


Figure 5.5: Boxplots of the four different Gaussian smoothing kernels and the original values. From left to right: $\sigma = 0.2$, $\sigma = 1$, $\sigma = 10$, $\sigma = 20$ and the original values of not smoothed green vegetation fractions.

5.4.2 Effect of smoothing on BRT prediction results

After performing the Gaussian smoothing using four different kernels sizes the smoothed values were fed into the BRT and we investigated two scenarios as described in 6.3; The two scenarios resulted in altogether 13 prediction results. Table 5.1 presents the prediction results and the RMSE as a goodness of fit for the BRT model outcome along with the contribution of the covariates listed in terms of their relative influence on the response. In order to enable a better comparison between the assessment of the model fit we added the RMSE for scenario 1, and to the smoothed data (second column), and to the original green vegetation fractions (third column), both Scenario 1, and the original green vegetation fractions using the smoothed values as covariates (sixth column) for Scenario 2.

	Scenario 1				Scenario 2			
	latitude + longitude		rel. influence in %		covariate + latitude + longitude		rel. influence in %	
kernel	smoothed	original	lat	long	original	covariate	lat	long
original	3.1836	-	56.37	43.63	-	-	-	-
$\sigma = 0.2$	0.1679	4.452	71.22	28.78	3.2359	66.14	18.85	15.01
$\sigma = 1$	0.5637	3.791	61.76	38.24	3.2092	79.94	11.31	8.74
$\sigma = 10$	1.4436	3.414	58.53	41.47	2.7231	93.78	3.51	2.71
$\sigma = 20$	1.6872	2.698	58.94	41.06	2.5251	95.16	2.46	2.38

Table 5.1: BRT prediction results on two scenarios. Scenario 1: only latitude and longitude have been used as covariates. Scenario 2: latitude, longitude and the smoothed values have been used to predict the observed green vegetation fractions (original values). The best RMSE is printed in bold numbers and the significance of the covariate in predicting the response and their importance in the splitting process of the BRT is depicted as their relative influence in %.

Analysis of the residual plots (Figure 5.4 b) and the RMSE (Table 5.1) reveals that the RMSE corresponding to $\sigma = 10$ in Scenario 1 is better than that for $\sigma = 20$ in Scenario 2. Both residual maps demonstrate that noise was smoothed out in the imagery adequately without showing a pattern in the residuals. Also shown in Figure 5.2 we see that $\sigma = 10$ has smoothed the data and the magnitude of the smoothing is sufficiently, whereas $\sigma = 20$ approximates very close to the real data. Therefore the preference is on $\sigma = 10$. The RMSE of Scenario 1 shows a strong improvement using the smoothed values in comparison with the original green vegetation fractions.

We also quantitatively assessed the utility of Gaussian smoothing as a data pre-processing step. By analysing Figure 5.4 and comparing the RMSE of the original values of the two scenarios we can demonstrate that smoothing as an applied pre-processing step helps to obtain better prediction accuracy and a better model fit as described in column 3 (original) and 6 (original) of Table 5.1 and compared with column 2 (smoothed). This can be explained as outliers and extreme values are smoothed out and the data follow a more symmetrical distribution than the original green vegetation fractions (shown in Figure 5.5) by comparing the IQR of all four kernels with the original, not smoothed or pre-processed data on the very right.

Marginal plots were also produced to facilitate comparison of the predicted and the observed values and provide diagnostic information about the fitted model. Figure 5.6 shows the marginal plots for all four Gaussian smoothing kernels. All plots indicate that the BRT model under-predicts high observed values as shown in the long tail of the corresponding right-skewed histogram. This is as expected. In general, all plots exhibit a positive and relatively strong relationship, with a tendency towards a stronger correlation using

a higher value of σ . The predicted values show a light tendency of a bi-modal distribution.

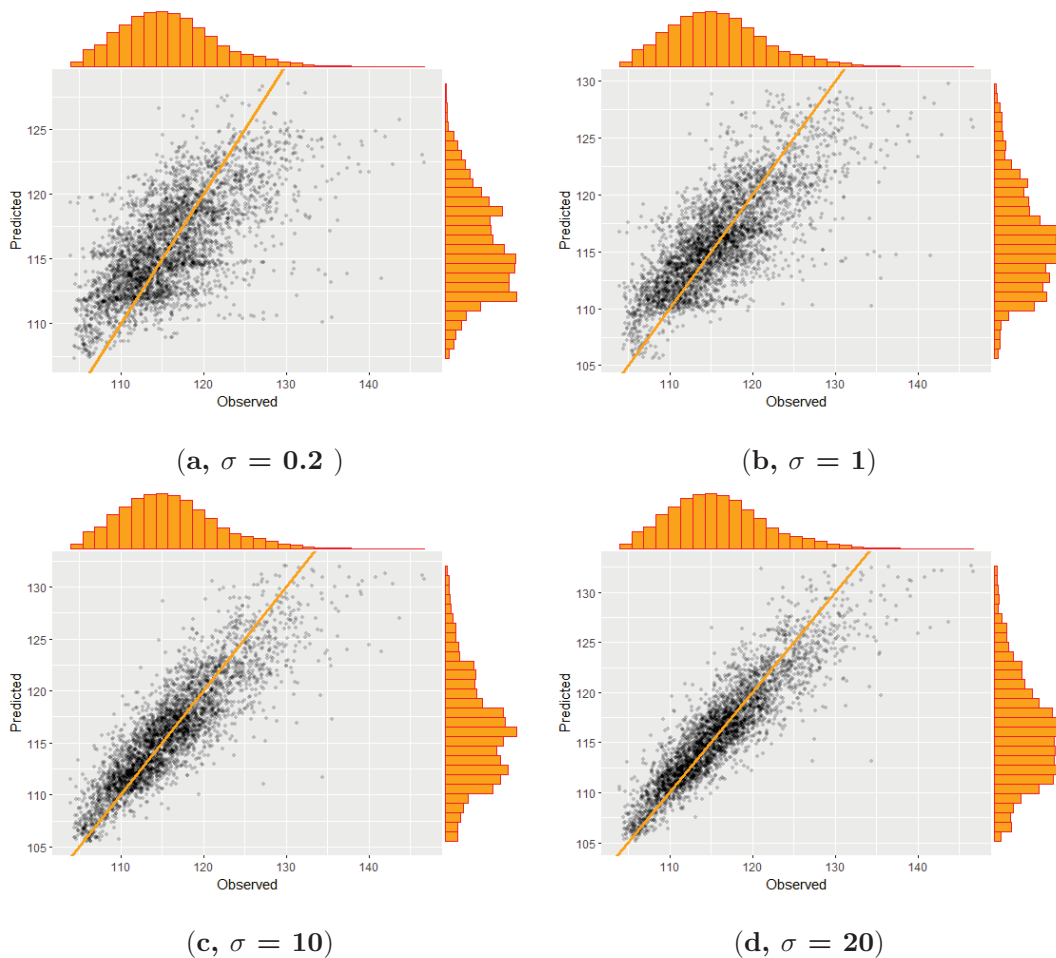


Figure 5.6: Marginal plots of the four different Gaussian kernels show that BRT under-predicts the green vegetation fractions shown on the y-axis in comparison to the observed values on the x-axis in (a) $\sigma = 0.2$ (maximum smoothing); (b) $\sigma = 1$; (c) $\sigma = 10$ (d) $\sigma = 20$ (the least smoothing). The distribution of the observed values are positively skewed and show a long tail starting at about 130 up to 140 and higher whereas the predicted values are less skewed and reach the maximum of 130 in histogram of the margin of the plot.

We also investigated the influence of the covariates on the prediction of green vegetation fractions. In Table 5.1 we list the relative influence of our smoothed covariates, measured as a percentage in both scenarios. The relative influence measures the contribution of the covariates in the splitting process of the BRT modelling process. Figure 5.7 shows graphically the very strong influence of the smoothed values in predicting the original green vegetation fraction as our response variable (Scenario 2). We can see the larger the value of σ , the smaller the contribution of the geographic coordinates on the response and the stronger the influence of the smoothed green vegetation fractions.

This is not surprising, since $\sigma = 20$ approximates the real data very well and its inclusion as a covariate also comes closer to simply providing the BRT with the actual data set as a covariate; this explains the contribution of 95.16 % as listed in Table 5.1. Surprisingly, even in these low-smoothing cases the BRT prediction accuracy suffered compared to the use of only latitude and longitude in Scenario 1.

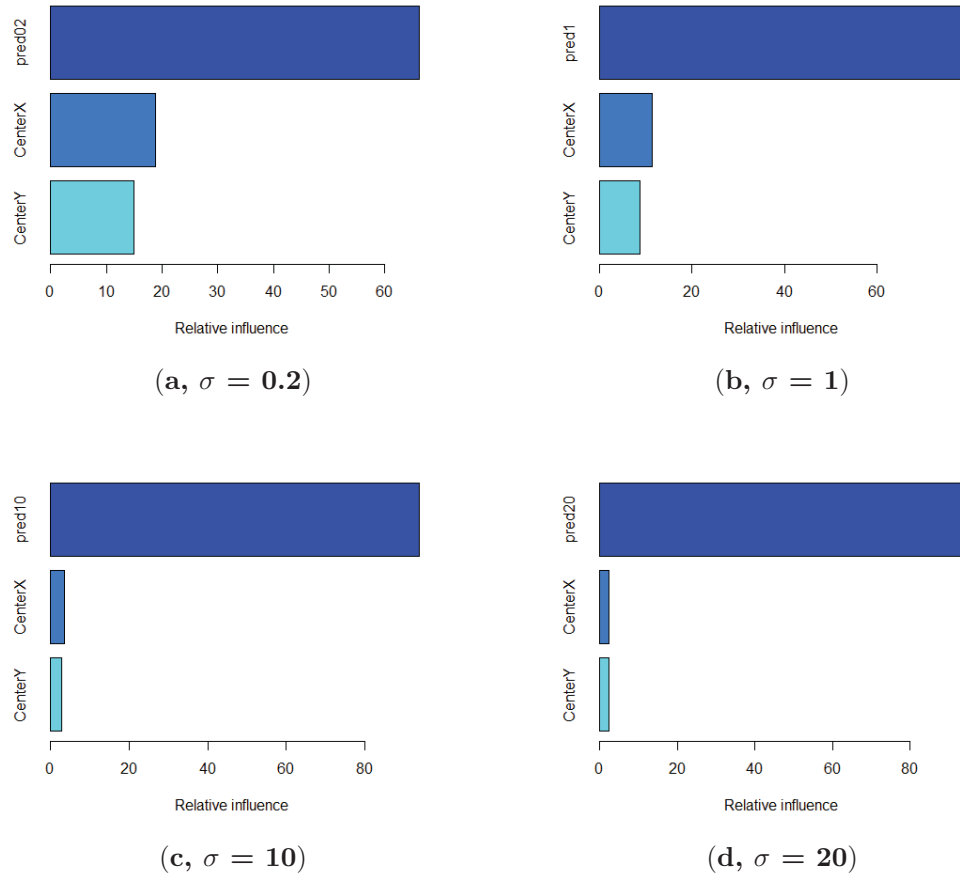


Figure 5.7: Relative Influence plots of the four Gaussian smoothing kernels showing the influence of the three covariates in predicting the green vegetation fractions (response variable) on the x-axis where (a) used $\sigma = 0.2$ (maximum smoothing); (b) $\sigma = 1$; (c) $\sigma = 10$ (d) $\sigma = 20$ (the least smoothing).

5.5 Discussion

The goal of this paper was to investigate two scenarios of smoothing kernels applied to green vegetation fraction data as a pre-processing step or as calculating a new response variable, and its effect on prediction accuracy using BRT. Spatial filtering blurs the edges of the pixels and the overall image quality deteriorates but the signal-to-noise-ratio will be improved by suppressing change in high frequency signals and enhancing low frequency

signals (Kumar, 2013). We used four Gaussian smoothing kernels, but many other statistical operations are also possible as described by Simonoff (2012).

We considered the effects of different values of the parameter σ that controls the level of smoothing applied. Our recommendation for the best suitable sigma is $\sigma = 10$ for our data set obtained in our study area. It would be interesting to find a suitable value of σ when smoothing on a regional level where several FCover scenes are combined and cover a large area. Also interesting to analyse further is the relative influence of the covariates on the response when smoothing larger areas. A drawback of our study is that for each scale and new FCover scene the value of σ needs to be tuned. We focused on one FCover scene only since it contained sufficient heterogeneity of geographic features to serve our inferential purpose. Our goal was not to find the most appropriate Gaussian smoothing kernel and its bandwidth for FCover data, but to investigate in neighbourhood dependencies of the existing topography present in this FCover image.

The focus of our study was to investigate the prediction capabilities of BRT and the effect of incorporating neighbourhood information in the spatial prediction on green vegetation fractions. In Table 5.1 we can see that by adding an additional covariate of green vegetation fractions, the prediction accuracy was positively affected and resulted in a smaller RMSE as this is an interesting outcome and provides reassurance in a better model fit and yields a higher prediction accuracy. Also very significant is the relative influence of the additional covariate in explaining the response variable as shown in Table 5.1 and Figure 5.7. We can see that in using $\sigma = 20$ the influence of the added smoothed green vegetation fractions is strongest, with a relative contribution of 95.16 %. The contribution of the added covariate significantly increases as σ increases, whereas the contribution of the latitude and longitude becomes insignificant and negligible.

In the marginal plots it was apparent that the predicted values showed a light tendency of a bi-modal distribution. In Chapter 4 of the thesis we showed marginal plots of BRT predictions of the same data set and the bi-modal distribution was stronger and indicated clustering in the data. Here, after smoothing the green vegetation fractions, the predicted values showed no clustering and a more symmetric, unimodal distribution in predicted values. We can conclude that in using a Gaussian smoothing kernel our prediction results follow a nearly unimodal distribution and clusters within the data have been resolved. Furthermore, a significant under or over prediction of peak values in the observed data is clearly visible when using smaller values of σ , and a stronger relationship between observed and predicted values is obtained when using larger values of σ . The BRT did not predict peak values very well, matching the results seen in Chapter 4 of this thesis, where we investigated spatial aggregation and where four different spatial aggregation schemes were applied on the same data set used here in Chapter 4. We can therefore conclude that applying Gaussian smoothing kernels as a pre-processing step prior BRT modelling,

should be favoured over spatial aggregation.

5.6 Future Work

A qualitative and quantitative assessment of the suitable size of the smoothing parameter σ in the kernel of the Gaussian has been used in this case study by comparing raster maps that demonstrate the magnitude of the smoothing and the corresponding residual maps. We can see that the value of σ is scale dependent and larger areas would profit from a smaller σ , whereas using one FCover scene the $\sigma = 0.2$ smoothed out too many local characteristics. Since our focus was on the investigation of prediction accuracy obtained using BRT and different smoothed green vegetation values, we refrained from performing cross validation to find the most suitable value of σ , although this could be undertaken in an obvious manner. CV tries to choose the value of σ to fit the smoothed data to the original, which is a separate issue to the question of using four different extents of smoothing that either assist or hinder the predictive capabilities using BRT.

Another possibility would be to use our approach on remotely sensed data showing data gaps for predictions and infilling purposes. Data gaps could be filled with prediction of the smoothed surface to ensure a smooth transition of raster values over time and/or space.

6 Estimating Spatial and Temporal Trends in Environmental Indices Based on Satellite Data: A Two-Step Approach

Preamble

This paper focuses on the applied aim 4 in which we investigate in different regression slope trends on green vegetation on the best identified scale of chapter 4. For this we used the scale 1:3000 and performed a linear regression in each grid cell to delineate the slope coefficient out of the model summary. Each grid cell with the corresponding slope coefficient is visualised as a raster tile showing negative trends, neutral trend on positive trends on the FCover on a colour range. Monitoring long term trends of green vegetation in a semi-arid region gives valuable insight into dependencies and changing quantities influenced by climate variability. For our study we investigated in the spatio-temporal trends of green vegetation using 3 scenarios, namely a data set covering 30 years of data, three consecutive decades analysed individually, and 30 years of data exploring four segments of the divided FCover scene.

This paper provides insight into spatio-temporal trends of green vegetation based on three scenarios and resulting predictive accuracy.

Statement for Authorship

This chapter has been written as a journal article. The authors listed below have certified that:

- (a) They meet the criteria for authorship as they have participated in the conception, execution or interpretation of at least the part of the publication in their field of expertise;
- (b) They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
- (c) There are no other authors of the publication according to these criteria;
- (d) Potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of the journals of other publications and the head of the responsible academic unit; and
- (e) They agree to the use of the publication in the student's thesis and its publication on the Australian Digital Thesis database consistent with any limitations set by publisher requirements.

The reference for the publication associated with this chapter is; **Student Brigitte Colin**, Other supervisor (2018). Estimating Spatial and Temporal Trends in Environmental Indices Based on Satellite Data: A Two-Step Approach. Sensors, Special Issue "Computational Intelligence in Remote Sensing". This chapter has been prepared as a paper and has been published.

Contributor	Statement of contribution
Student Brigitte Colin	Conduct the research, develop code for the statistical approach, write the manuscript, make modifications to the manuscript as suggested by coauthor and reviewers.
Signature and date:	
Kerrie Mengersen	Propose and supervise research, comments on manuscript.

Principal Supervisor Confirmation: I have sighted email or other correspondence for all co-authors confirming their authorship.

Name: K Mengersen Signature:  Date: _____

QUT Verified
Signature

Article

Estimating spatial and temporal trends in environmental indices based on satellite data: a two-step approach

Brigitte Colin ¹, and Kerrie Mengersen ¹

¹ Affiliation 1; School of Mathematical Sciences, Queensland University of Technology, Brisbane QLD 4000, Australia, b.colin@qut.edu.au

* Correspondence: b.colin@qut.edu.au; Tel.: +61-3138-2019

Version July 25, 2019 submitted to *Sensors*

Abstract: This paper presents a method for employing satellite data to evaluate spatial and temporal patterns in environmental indices of interest. In the first step, a linear regression coefficients are extracted for each area in the image. These coefficients are then employed as a response variable in a boosted regression tree with geographic coordinates as explanatory variables. Here, a two-step approach is described in the context of a substantive case study comprising 30 years of satellite derived fractional green vegetation cover for a large region in Queensland, Australia. In addition to analysis of the entire image and timeframe, separate analyses are undertaken over decades and over sub-regions of the study region. The results demonstrate both the utility of the approach and insights into spatio-temporal trends in green vegetation for this site. These findings support the feasibility of using the proposed two-step approach and geographic coordinates in the analysis of satellite derived indices over space and time.

Keywords: Boosted Regression Tree; spatio-temporal analysis; fractional cover data; prediction of location-based vegetation trends

1. Introduction

Remotely sensed data are available from a wide range of sources, ranging from satellites to drones, and have been used for a very wide range of environmental applications and analysis of spatial and temporal trends. For example, Landsat data are freely available [1]; the imagery covers a wide geographical area, and it avoids expensive, extensive and often impractical in-situ measurement. There is a strong advantage in using remotely sensed Landsat imagery for land use and land cover (LULC) analyses in detecting and estimating the magnitude of spatio-temporal trends in measures of the quantity of green vegetation [2,3]. Monitoring long term trends of green vegetation in a semi-arid region gives valuable insight into dependencies and changing quantities influenced by climate variability. For monitoring and analysis of green vegetation, the infrared (IR) and near infrared (NIR) spectral channels are best suited since they discriminate between green and active vegetation versus woody vegetation or organic litter [4]. Fractional cover (FCover) data is a derived product out of Landsat imagery and shows the fractions of existing land cover in one pixel as percentages that are contained within the pixel. A satellite pixel combines the reflected radiation from different objects on the earth surface, and this spectral mixing effect results in a so called mixed pixel, or Mixel [5]. In a spectral unmixing approach the Landsat pixel is divided into assigned biophysical variables [6–9]. For example, in our study described below, we used only the band that shows the fractions for green vegetation out of a three layer composite containing two additional layers for bare soil and for non-green vegetation. The derivation of FCover is described in [6,10,11].

Environmental Modelling is important when we want to understand and monitor the local variability and spatial trends over time of green vegetation. Instead of applying our method to the full resolution we used aggregated FCover pixels showing a much coarser resolution to examine green vegetation trends. The aggregation scheme, the best resolution for LULC studies and its suitability are described in our paper [12]. The goal is to detect spatial and temporal trends based on 30 years of data. More specifically, we want to understand how the linear trend in FCover changes over the spatial region, and how well this can be described by geographic coordinates. For this, we specify qualitative factor levels showing six categories of green vegetation trends that are listed and further described in Table 2. Then we model these trends using latitude and longitude coordinates that serve as a North-South gradient (latitude) and East-West gradient (longitude).

The use of latitude and longitude as surrogate covariates is not uncommon. For example, in a study of [13] the authors used latitude and longitude coordinates as surrogate variables for North-South and East-West gradients to account for the variation in deciduous forested ecoregions. The response was an aggregated Normalized Difference Vegetation Index (NDVI) variable used as an on-site quantification of vegetation in North America. Similarly, in a study of the geographic distribution of plant functional types [14] the authors examined the relationship of precipitation and temperature on C3 (cool-season grasses) and C4 (more drought resistant warm-season grasses) grass types and shrubs using latitude and longitude coordinates. Along a given longitude, C3 grasses increased with latitude and as one moved westward, C4 grasses were replaced by shrubs. They concluded that latitude and longitude can be used as surrogate variables for the main climatic dimensions of the area. The latitude and longitude explained a substantial portion of the variability of the distribution of the relative abundance of shrubs, C3 grasses, and C4 grasses.

In general, there are many methods using machine learning approaches for predicting temporal and spatio-temporal trends that are not limited to green vegetation. Examples include long term seasonal changes of the Danube River eco-chemical status [15], epidemiology studies and analysis of disease processes in public health [16], spatial and temporal trends of birds over France [17], long term trends in dryland vegetation variability in Ethiopia [18], and identification of environmental controls in fire-prone biome and spatial patterns at several spatial scales in the Canadian boreal forest [19]. In a previous paper, we evaluated the performance of a popular machine learning technique, namely boosted regression trees, and concluded that BRT can perform well in high-dimensional and complex problems, deal with missing data by default without the need for interpolation/infilling, describe complex non-linearity and interactions between variables, deal with spatial and non-spatial data and different data granularities, and reduce data complexity without negatively affecting prediction performance [20]. However, the focus of that paper was on spatial estimation of environmental indices at a single point in time. In this paper, we focus on estimation over both space and time. We do this by proposing a two-step approach comprising the extraction of slope coefficients out of the model summary and the predictions of the extracted slope coefficients using BRT. The combination of a linear regression and a non-linear BRT model defines our two step approach.

The detection of trends in change of green vegetation over time is essential for the assessment of the impacts of climate variability on the LULCC (Land Use Land Cover Change) of a region. The study described in this paper aims to determine the annual trends of slope coefficients over a semi-arid region. Long term (1987–2017) gridded aggregated FCover fractions of green vegetation data are used to spatially divide the FCover scene. Historical trends are examined using a linear model to regress the aggregated green fraction over time for each grid cell. The extracted grid-specific slope coefficients are then used as a response variable, with the corresponding latitude and longitude as covariates, in the hierarchical supervised machine learning BRT model. The BRT results thus provide an evaluation of

the spatial nature of the overall temporal trend in green vegetation over time.

The paper is structured as follows. Section 1 provides background information, places our study in the context to other studies and demonstrates why there is a gap we need to fill. Section 2 introduces the study area and presents the context of the other linear model approach to extracting the slope coefficient. In Section 3 we introduce the BRT modelling approach and describe the hyperparameter tuning steps and the model goodness of fit. Section 4 presents the results of the two stages of the analysis. The implications of the data and the output of the prediction of the BRT, as well as strengths and limitation measures of BRT are discussed in Section 5.

2. Data description

2.1. Case Study

The FCover scene of our study area is located the Northern Territory, Australia, in the Landsat footprint of path 102 row 72 according to the Worldwide Reference System-2 (WRS-2). The scene covers an area of 185 x 185 km and the elevation is ranging from 50 m to 213 m. The location of our study area is classified as “Dry” with variations of “desert, hot arid” and “dry Summer, hot arid” (BWh and Bsh) based on the Köppen-Geiger scheme and presents heterogeneous and complex topography of native grass types. Arid and semi-arid areas cover a large part of the earth’s surface and are located around the tropics at 23° north and south of the equator. According to the commonly used Köppen-Geiger climate classification these areas are defined by limited precipitation and high potential evaporation rates. Further, our study area is used for commercial grazing purposes and is highly dependent on grazing practices that ensure future sustainable land use [21].

2.2. Fractional cover data

There is a strong advantage in using Landsat satellite data for monitoring vegetation trend in land use and land cover (LULC) studies [2,3]. The imagery covers a wide geographical area; it avoids expensive, extensive and often impractical in situ measurement and it is freely available [1]. The spatial resolution of a Landsat pixel combines the reflected or emitted radiation from different objects on the Earth’s surface and, as described in the Introduction above, this spectral mixing effect results in a so-called mixed pixel or Mixel where individual spectra of objects cannot be separated [5]. Fractional cover (FCover) data is a derived product based on Landsat 5 Thematic Mapper (TM) imagery. An extensive ground cover sampling study [10] was used to inform a spectral unmixing algorithm [6,10,11].

2.3. Data pre-processing

In Colin et al. [12] we investigated spatial aggregation schemes that are best suited for this study site with regard to up-scaling FCover data and maintaining sufficient local characteristics for accurate prediction of green vegetation. Instead of dealing with the original amount of 54 million pixels we thus reduced the data volume to 5530 individual spatial grid cells containing 100 x 100 pixels in each as demonstrated in Figure 1.

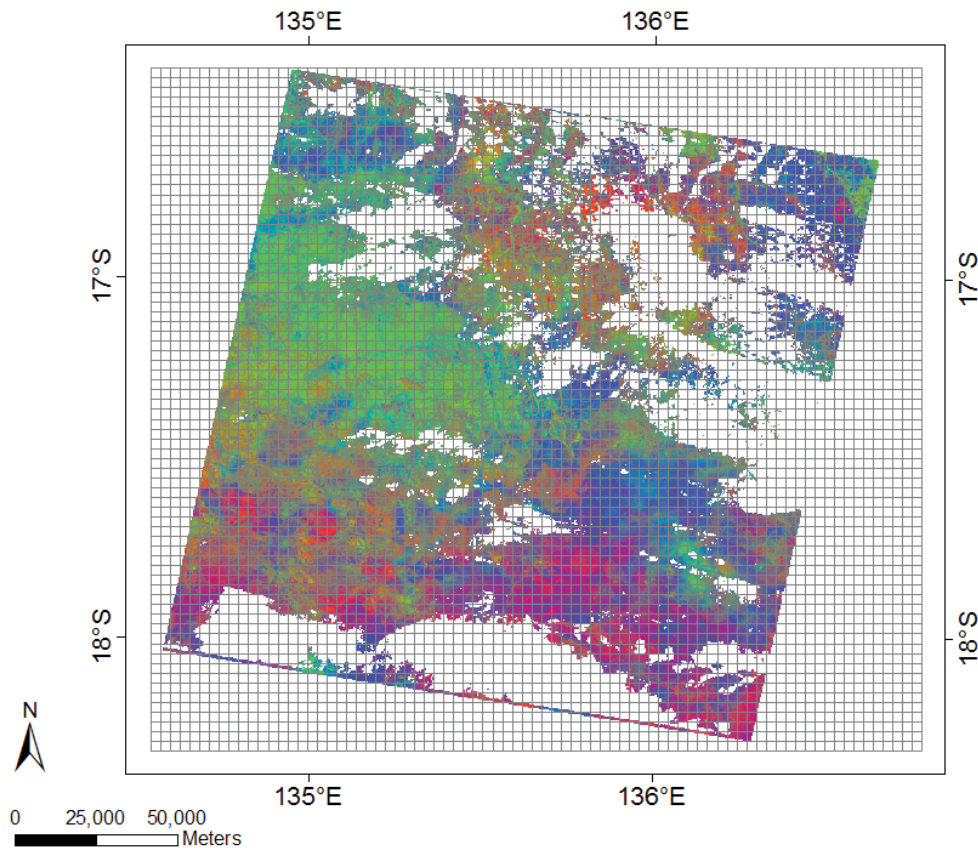


Figure 1. The FCover data are overlaid with an evenly spaced grid where each grid cell contains 100×100 pixels and covers an area of 3000×3000 m. Each of the total 5530 grid cells will be used to delineate the slope coefficients showing green vegetation trend on unique locations of the spatial grid. The FCover scene shows the relationship of the three ground cover classes of green vegetation (green), non-photosynthetic vegetation (blue) and bare soil (red) referenced on the Worldwide Reference System-2 [12].

Missing data are common in remotely sensed imagery. As part of enhancing data quality obscuring elements such as clouds and cloud shadows are filtered out, resulting in data gaps. The amount of missing data can be substantial and often imagery can not be used at all due to too much data gaps as demonstrated in Figure 2a in the year 1992 and 2000. Further, Figure 1 shows a FCover scene with an overlaid evenly spaced grid where we can see that we have empty grid cells and data gaps where no information of green vegetation fractions is present.

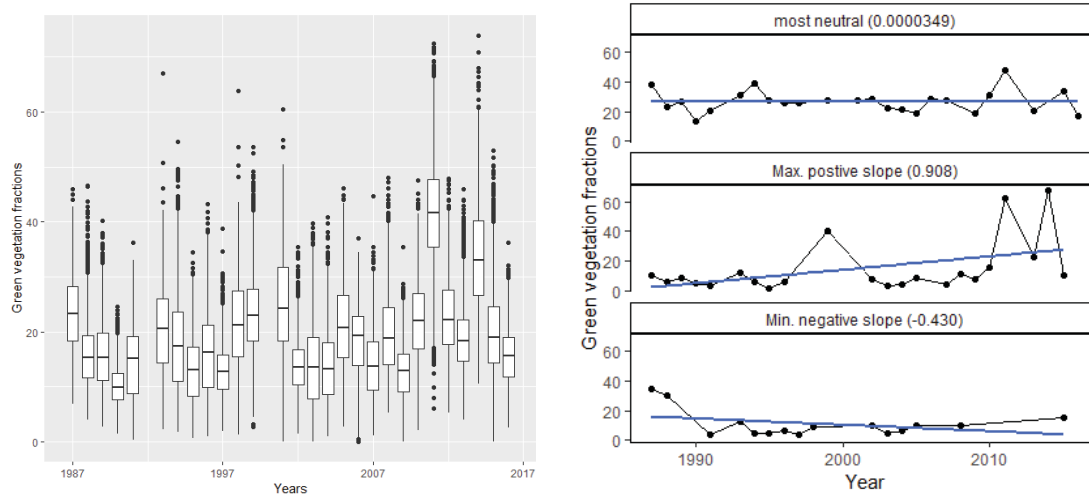
For studies on marine and vegetation monitoring using remotely sensed imagery from earth observation satellites a geographic scale of 1km and finer is mostly used, for example MODIS, Landsat and ENVISAT MERIS. For climate related studies a coarser spatial resolution is preferred. The Meteosat Second Generation (MSG) deliver data recorded in 12 channels with a spatial resolution of 3km. MSG data are primarily designed for meteorological observations of the atmosphere, but there are land use applications based on MSG data as well [22,23]. In a particularly interesting investigation [24], the authors used a spatial scale of 3km to present the first results of NDVI for the whole of the African continent.

2.4. Data Exploration

Table 1. Descriptive statistics of the green vegetation fractions for the whole data set covering 30 years

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	11.64	17.03	18.37	23.16	73.92	0

Summaries of the data extracted from the FCover scenes covering a 30 year time frame for this case study are presented in Table 1 and Figure 2. Table 1 presents overall summary statistics for the observed grid-level values of green vegetation fraction. Figure 2 summarises the spatio-temporal nature of the data, through boxplots of the annual distribution of the grid-level values as well as three trends over time.



(a) Boxplots showing a strong variation of green vegetation for each year over the 30 year timeframe, 1987-2016. For consistency over time, and because the FCover in the study area is dominated by wet and dry seasons, only December scenes have been used for this case study.

(b) Trend of green vegetation in a 30 years time frame overlaid with a blue linear regression line showing the direction of trends. Top) most neutral, middle) maximum positive slope and bottom) minimum negative slope. Only sites with at least 15 observations over the time period were considered for this plot.

Figure 2. Development of green vegetation fractions and their trends over time shown as boxplots in Figure 2a or as the three most distinctive trends in Figure 2b.

3. Methods

Our goal for this study is to evaluate the spatial, temporal and spatio-temporal nature of vegetation trends in FCover data. For this we followed a two-step approach comprising a linear model for the extraction of the slope coefficients (serving as trends in vegetation) followed by the prediction of those extracted slope coefficients using geographic coordinates through a BRT model.

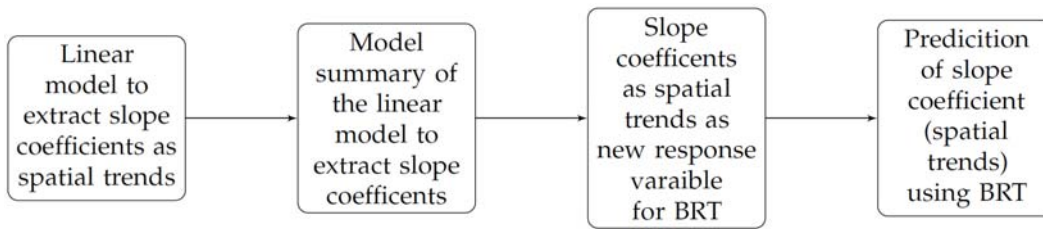


Figure 3. Two-step modelling approach to predict extracted slope coefficients (as spatial trends) using a BRT model.

3.1. Linear Model

3.1.1. Extraction of slope coefficients

In the first step, for each grid cell, we extracted linear slope estimates from least squares linear regression models for which we used a continuous response variable showing the aggregated fraction of green vegetation and a discrete predictor variable indicating the years from 1987 to 2017. (This was performed using the `lm` package in R, with the formula: green vegetation fractions \sim years). A positive (negative) slope coefficient indicates an increasing (decreasing) trend and increasing (decreasing) quantity in green vegetation fractions over time. The R software [25] and its basic linear model function was used to fit the model. Each individual grid cell is indexed as i and our data set comprise 5530 individual grid cells per year and FCover scene as demonstrated in Figure 1. We further assume that green vegetation, denoted as Y_i , is linearly related to the covariates year, denoted as X , and that the residuals ϵ_i are distributed $N(0, \sigma^2)$. A linear regression model with one predictor variable can be expressed with the following equation: $Y_i = \beta_0 + \beta * X_i + \epsilon_i$. The parameters in the model are β_0 as the Y-intercept and β as the regression coefficient (the slope coefficient representing the linear trend over time) which we extract from the model summary.

Table 2 shows the different slope coefficients that were calculated. It can be seen that there are three dominating slope coefficient classes ranging from a negative trend of -0.5 up to a positive trend of 1. Further, we can see that the most extreme values are exclusively found on the outer rim of the rastermap shown in Figure 4 a and their overall representation/contribution is marginal (0.02 % and 0.03 %) as demonstrated in Table 2. We suspect this is based on the natural shift of the Landsat footprint recording which therefore results in extreme data gaps that adversely influence the linear regression analyses and hence the accuracy of the estimates of the corresponding slope coefficients. By visualising all six categories of our extracted slope coefficients, we can see that the three strongest categories ranging from negative -0.5 up to a positive trend of 1, together comprise 99.93 % of the slope coefficients as demonstrated in Figure 4a.

Table 2. Six categories of slope coefficients with corresponding numbers of observations in the data set, and their overall representation/contribution as percentages in the case study.

Slope coefficient categories	Observations	Percentages %
slope coefficient >1	14	0.02 %
slope coefficient ≥ 0.5 and slope coefficient <1	5088	5.44 %
slope coefficient ≥ 0 and slope coefficient <0.5	79032	84.48 %
slope coefficient ≥ -0.5 and slope coefficient <0	9364	10.01 %
slope coefficient ≥ -0.5 and slope coefficient <-1	30	0.03 %
slope coefficient <-1	19	0.02 %

The extracted and categorised slope coefficients were plotted using their geographic centroid coordinates from the overlaid spatial grid. This resulted in the Figure 4a showing six categories and

their location-based slope coefficients based on a 30 year time frame. Further, we can see that the most extreme values are exclusively found on the outer rim of the plot. We suspect this is based on a natural shift of the Landsat footprint and therefore extreme data gaps and in general a lower availability of FCover fractions to sufficiently estimate reliable slope coefficients. Statistically significant trends were assumed to exist if the p-values of the slope coefficient were different from zero at a level of 5% ($p < 0.05$) or smaller. The levels of statistical significance for each grid cells are shown in Figure 4b where the p-values >0.05 are most common.

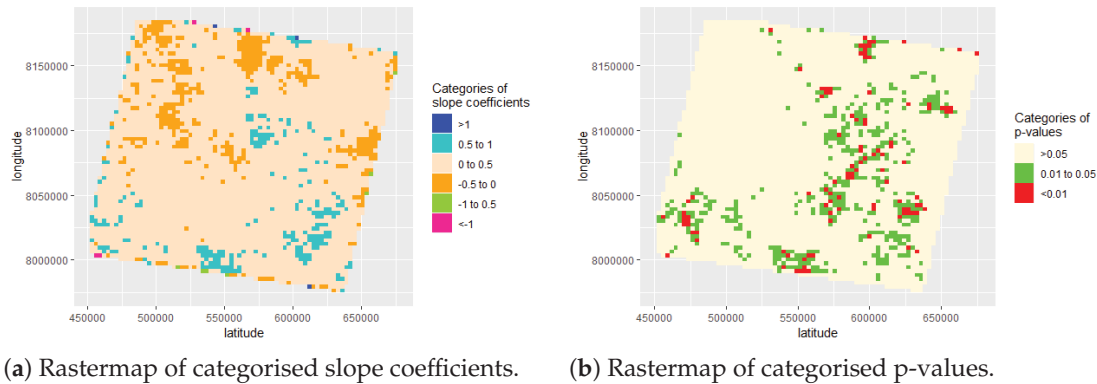


Figure 4. Location of p-values and of slope coefficients categories.

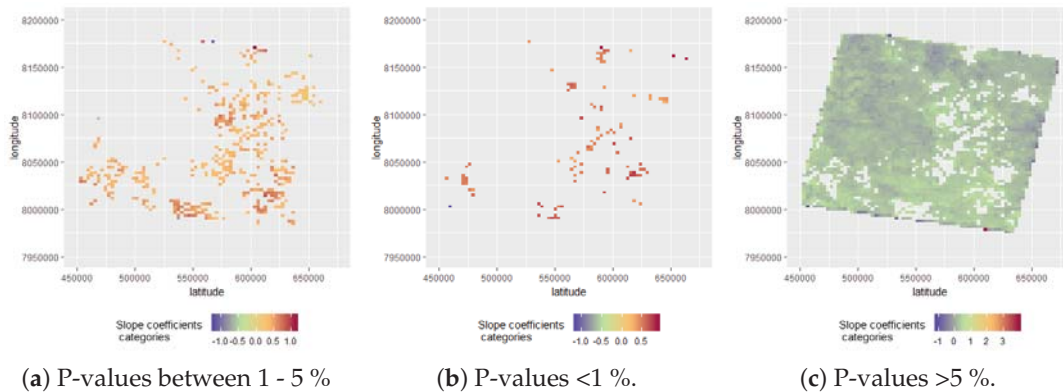


Figure 5. Comparison of the three p-value categories with their associated slope coefficients.

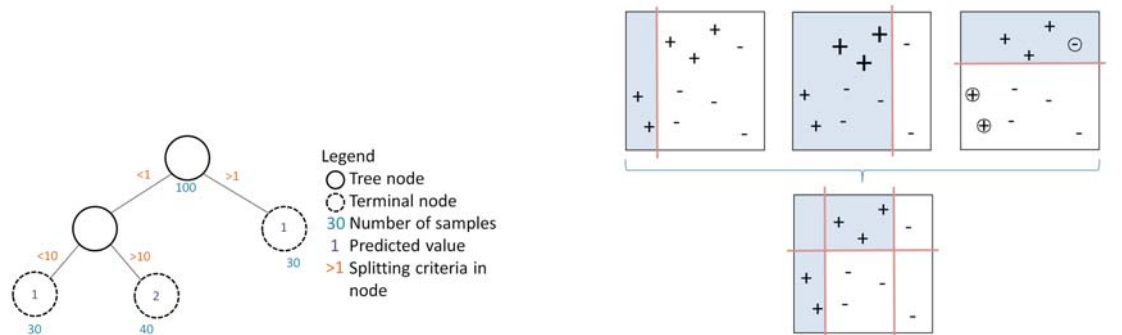
Figure 5 shows the extracted three categories of p-values, their slope coefficients and associated statistically significant localised effect. We can clearly see that the p-values $>5\%$ are most common. In the second step of our two-step approach, we used the estimated slope coefficients for each grid cell as our new response variable in a BRT model to predict green vegetation trends using geographic coordinates. The model is described in detail below.

3.2. Boosted Regression Tree

Our research aim was to investigate in spatial, temporal and spatio-temporal trends in green vegetation. The methodological aim was the extension of the created BRT model to predict quantitative long-term trends of green vegetation cover in a semi-arid region that is vulnerable and sensitive to climate variability. For this we used 30 years of data and a 2-step approach comprising the extraction of the slope coefficients (our trends) using a linear model, and use those slope coefficients as our new

response variable for the BRT modelling.

A BRT is a flexible supervised machine learning method that consists of two algorithms: first a regression tree approach and second boosting. In a regression tree, the feature space is divided in binary trees as shown in Figure 6a, whereas Boosting combines several single binary trees to create a more flexible partition of the feature space Figure 6b. Boosting proceeds by fitting another regression tree to the residuals, until a stopping criteria is reached, or when an acceptable goodness of fit for predictive accuracy is achieved. Details of BRTs are provided in [26].



(a) Regression Trees: Hierarchical regression and binary splitting process showing observations in the nodes, predicted values in the terminal nodes and splitting criteria along the tree branches [12].

(b) Boosting: BRT as an ensemble approach combines several binary splits to create complex prediction rules that offer more flexibility in dividing the feature space than a single regression tree. Boosting additively fits binary trees and gradually prioritise poorly modelled data to produce a set of binary splits that maximally reduce the BRT loss function. Adapted from [27], [12].

Figure 6. Combination of two algorithms, namely Boosting and linear regression tree within a BRT.

There is a number of statistical machine learning methods for fitting complex regressions of the type considered here. For example, generalised additive models (GAM) [28] provide a flexible extension of well known generalised linear models and have been widely used in ecological applications, for example to predict tree species in Utah [29]. This author also compared GAM with stochastic gradient boosting (SGB) and found that both had merits with respect to predictive fit. Using another ecological case study, [30] compared GAM and BRT and found that BRT showed substantially superior predictive performance. Further, in our paper we compared three different regression methods (Random Forests and Least Absolute Shrinkage and Selection Operator, LASSO) with each other and evaluated their predictive performance on heterogeneous spatial data and concluded that BRT outperformed these in this instance [20].

3.2.1. Hyperparameter tuning and goodness of fit evaluation

The R package caret [31] was used to determine the hyper-parameters for the BRT model and split the dataset into training data to train the model and a test set to validate the predictive performance of the created model on unseen data. The goodness of fit of the model was evaluated using the root mean square error (RMSE). The results are summarised in Table 3. The final suggestions of the hyper-parameter tuning process were number of trees ($n.trees = 2500$), interaction depths between nodes in the tree branches ($interaction.depth = 5$), learning rate or shrinkage ($shrinkage = 0.01$) and a set number of minimum observations in the splitting node ($n.minobsinnode = 10$).

The BRT model was fit using the `gbm` package [32]. We set the formula parameter to "slope coefficients ~ latitude + longitude" to describe the model, set the hyperparameters as described above and held all other parameters at their default values. The computational environment was the R statistical modelling software version 3.3.3 [33] running inside Windows 7 SP1 (64-bit) on a 2.60 GHz Intel i7 CPU with 16GB of RAM. All of the plots were generated in the R programming language [33].

4. Results

4.1. BRT predictions

4.1.1. Overall results of the whole data set

The marginal plot in Figure 7a shows a strong linear relationship between the observed values and the corresponding predicted values obtained under the BRT model. It also shows the marginal distributions of the two sets of values, reflecting a relatively normal distribution for the observed values and a bimodal distribution for the predicted values. This is also shown in the histograms in Figures 7b and 7c. We can see somewhat unexpected slightly positive trends in the quantity of green vegetation cover in our study area throughout the years. This phenomenon is not unusual for tree-based models.

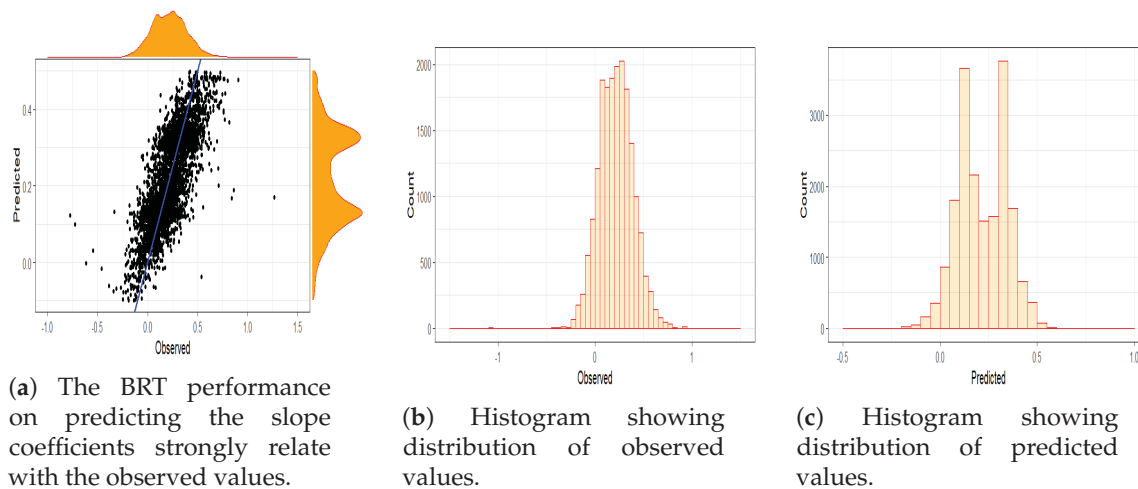


Figure 7. Data set showing 30 years in a) marginal effects, b) histogram of observed values and c) of predicted values.

4.1.2. Decadal analyses

The linear model captures only an overall monotonic rate of change over all 30 years and can not detect changing trends over time. To overcome this, we subdivided the 30 years into three 10-year time frames to investigate if changing trends within the decades could be detected individually. Figure 8 shows the comparison of our results and demonstrate, that there is no significant difference. The analysis over three 10 year time frames showed a very similar temporal trend and a slightly increasing vegetation cover as in using the whole data set covering 30 years demonstrated in Figure 7. The scatterplot shows a strong positive correlation between the predicted values and the original measurements. You can see that BRT has the tendency to under- and over- predict extreme values demonstrated here on the predicted values following along the blue line of equality.

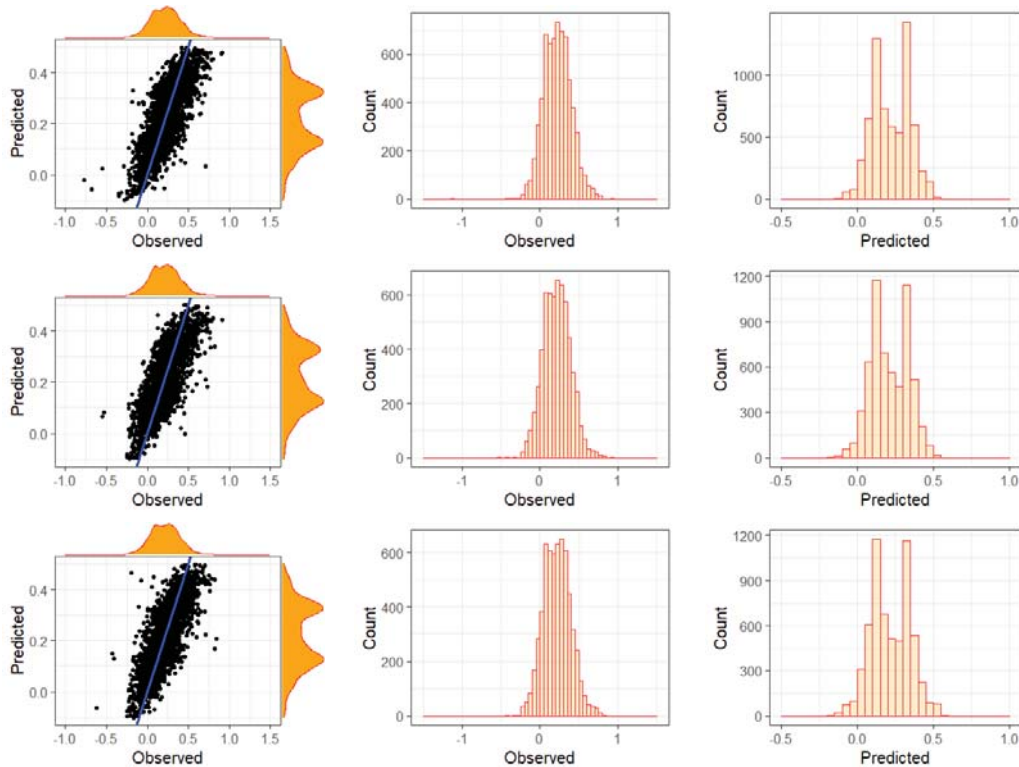


Figure 8. Plots of the decades showing a) marginal effects, b) histogram of observed values and c) of predicted values. The top panel shows the decade starting at 1987, the middle panel starting at 1997, and the last panel starts with 2007.

The decade-specific models displayed a similar fit to the overall model, with respect to RMSE (Table 3).

4.1.3. Segmented areas

In addition to splitting the data into decades we used all 30 years and divided our study area into four even segments, namely upper left corner, lower left corner, upper right corner, and lower right corner. Altogether the splits resulted in eight scenarios comprising 1) the whole data set described in 4.1.1, 2) three decades described in 4.1.2 and 3) four segmented areas of the whole data set described here. The overall model fit of all eight scenarios is shown in Table 3.

Partial dependency plots (PDP) are graphical visualizations of the marginal effect of our latitude and longitude covariates on the predicted response, here the extracted slope coefficients from the linear model described in 3.1.1. The Figure 9 shows all eight scenarios. On the left hand side we can see PDP showing the longitude. The general pattern shows an increase starting at the lower left corner at about the geographic coordinate of 8000000° reaching its peak and then decreasing slowly again. The only exception can be seen on segment upper right corner. On the right hand side we see the latitude starting high at 450000° and falling and increasing again at 520000° with the exception of the lower left corner. Please see Figure 4 for further details on the categories of slope coefficients and their associated p-values with geographic coordinates.

To investigate the model fit we used the Root Mean Square Error (RMSE) as shown in Table 3 All eight scenarios were created using 80 % of data for training the model and the remaining 20 % of data

for testing the model performance on unseen data. The Table shows the RMSE for the predicted values on the test data.

Table 3. RMSE on the test data using all 30 years, first 10 years, middle 10 years and last 10 years and in four segmented areas comprising a 30 year time frame.

Scenario	RMSE
All 30 years	0.1150
First 10 years	0.1112
Middle 10 years	0.1214
Last 10 years	0.1063
Four segments	
1 - Upper left	0.1076%
2 - Upper right	0.0915%
3 - Lower left	0.1112%
4 - Lower right	0.1265%

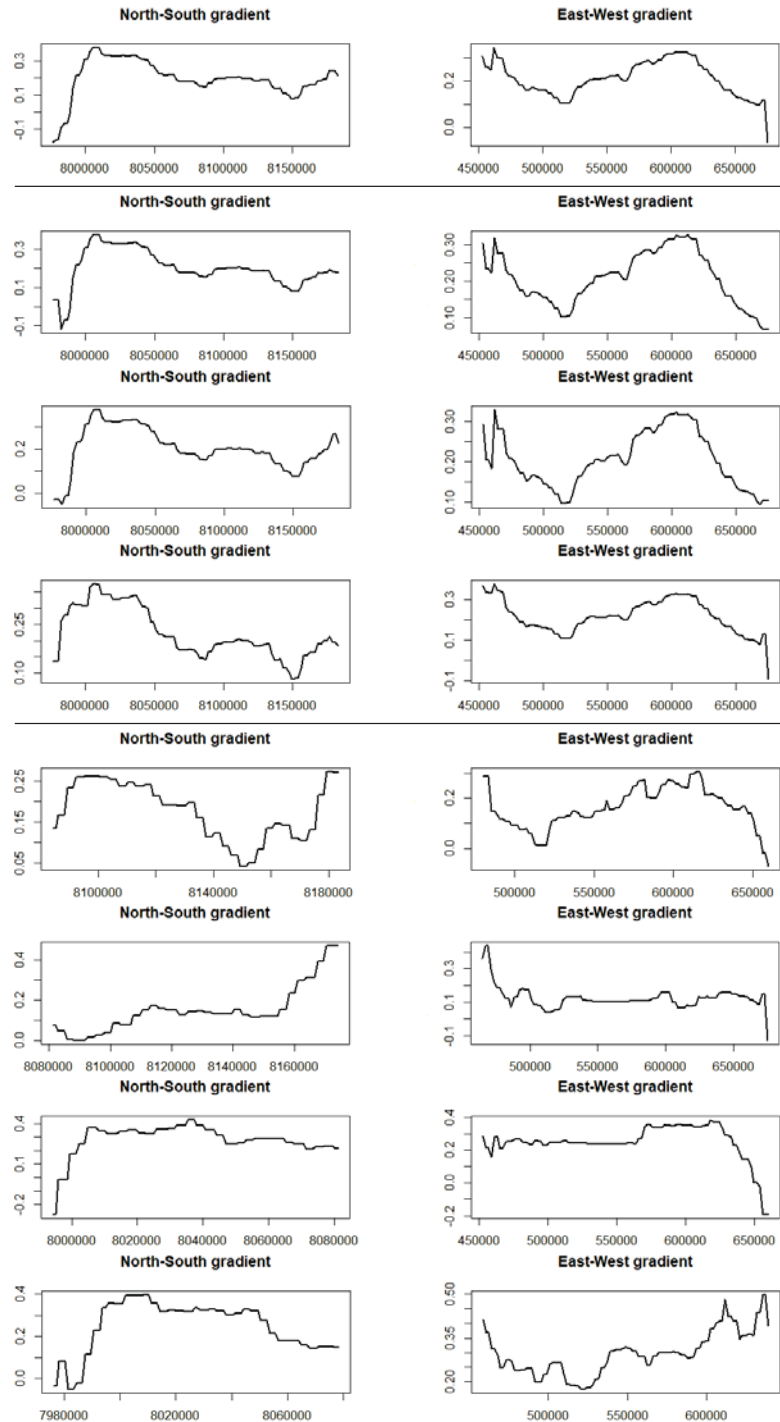


Figure 9. Partial dependency plots (PDP) of all eight scenarios in the order from top to bottom. All 30 years, first 10 years, middle 10 years, last 10 years, segment 1 (upper left), segment 2 (upper right), segment 3 (lower left), segment 4 (lower right). On the left there is the PDP of the latitude (North-South gradient) and the right shows the longitude (East-West gradient).

4.2. Relative influence

The spatial pattern of the trend values over the case study image can be further assessed by evaluating the relative influence of each of the covariates, latitude and longitude, in the BRT analysis.

The North-South gradient was slightly more influential than the East-West gradient. This pattern persisted in the decade analyses as well (Table 4).

Table 4. Relative influence of the longitude in explaining the response using all 30 years, first 10 years, middle 10 years, last 10 years, and in all four segments of the FCover scene.

Scenario	North-South gradient
All 30 years	56.77%
First 10 years	57.04%
Middle 10 years	55.68%
Last 10 years	57.67%
Four segments	
1 - Upper left	34.63%
2 - Upper right	47.71%
3 - Lower left	40.79%
4 - Lower right	43.24%

5. Discussion

In this paper we have proposed a two-step method for evaluating the spatial patterns of linear trends across a landscape, based on the geographic covariates latitude and longitude. The relative importance of these covariates, combined with the trend estimates themselves, can provide a deeper understanding of environmental impacts on the target response. For instance, in the case study considered here, the analyses allow insight into whether climate variability appears to have little to no impact on the existing green vegetation our study area. In Figure 2a we show the distribution of green vegetation fractions in boxplots covering 30 years and visualising the inter quartile range, minimum and maximum values. We can see an increase in the median especially in the year 2011 and 2014 the highest fraction of green vegetation of 70% and higher. This is surprising because in many studies a general trend of desertification in semi-arid regions around the world could be found. However, our results demonstrate that 84.48 % of all extracted slope coefficient show a neutral to a slightly positive trend in green vegetation as shown in Table 2.

We conducted a temporal and spatio-temporal investigation on one overall data set or on three data sets covering one decade each to get a better understanding if there are seasonal patterns that will not be captured by the overall 30 year time frame. Our findings are demonstrated in Figure 7, and 8. The RMSE error listed in Table 3 indicate that there is no significant influence in dividing the data set to improve prediction accuracy. In addition, it can be seen that BRT under-predicts the slope coefficient when using geographic coordinates as spatial gradients.

By plotting the p-values using their geographic coordinates we can demonstrate a spatial trend of significant strong p-values associated with the extracted slope coefficients as demonstrated in Figure 5.

Further, we demonstrate a stronger influence of the longitude coordinates in explaining our response variable as demonstrated in Table 4. To get insight in temporal and spatio-temporal trends, we split up the FCover scene and investigate several scenarios, namely the whole data set, the three decades and 30 years in four even segments of the FCover scene. We investigated if there are spatial trends in the slope coefficients and trends of green vegetation in each scenario. Figure 9 shows the influence of latitude and longitude in all 8 scenarios as partial dependency plots. We can clearly see that each segment and each decade differ from each other and affirm our approach in using consecutive time intervals to investigate in spatial green vegetation trends individually to get spatio-temporal insight in the amount of green vegetation fractions and how the greenness developed over space and time.

Using a linear model to extract slope coefficients allows formal, statistical investigation of the vegetation trends and associating p-values. However, it only considers trends as a linear monotonic trends of green vegetation. We tried to overcome this by dividing the data set into three decades, but it has been demonstrated that it did not improve prediction accuracy substantially. Further, no turning points or extreme events were taken into consideration that would have described changes in green vegetation fractions since the linear approach could not have detected them. As seen in Figure 8 we used this to split the data in three decade shows that there were no other significant trends captured. We only used geographic surrogate gradients without adding any other environmental covariates to the BRT model. In addition, no further testing using other FCover scene of a different Landsat footprint was taken to determine if the results are restricted to our location only.

There are many generalisations of the approach presented here. For example, while this paper has intentionally focused on a single analytic method for each of the two steps in the proposed approach, it is clear that these methods could be replaced by any one – or indeed a number – of a wide variety of statistical machine learning methods designed for estimation and/or prediction. For example, instead of the linear regression and BRT approaches illustrated here, one could consider other regression models that capture temporal and spatial correlation (e.g., exponentially weighted moving average models, Markov random field models, respectively) or other non-linear models such as neural networks or support vector machines. It is also valuable to look to the literature in other domains that evaluate and compare these and other methods for spatial and temporal estimation and prediction, such as [34–36]. Generalising in another direction, although this paper has focused on a single output from the first step (the estimated regression coefficient) and used this as a univariate response for the analysis in the second step, a multivariate approach could be adopted whereby the outputs from the first step (and inputs for the second step) are the regression estimates and their associated standard errors and/or RMSE, or estimates of multiple coefficients in a multiple regression, or parameter estimates and associated goodness of fit estimates from an alternative supervised learning method such as a neural network.

6. Conclusions

In this study, we demonstrated that a localised and quantitative distribution of temporal and spatio-temporal trends of green vegetation cover can be predicted using BRT. All together eight scenarios have been investigated, namely the whole data set covering 30 years, then three data sets covering a decade each, then the four quadrants of the image over all years. We showed that the prediction of location-based trends of green vegetation achieved good results by using the RMSE as goodness of model fit by combining a linear model and BRT. The extracted slope coefficient and p-values were categorised and further analysed by their direction of their increase of the quantity in green vegetation and their associated statistical significance through the p-values. A limitation can be found that 84% of the slope coefficients were positive but most were associated with non-significant p-values. In our paper [12] we concluded that a North-South gradient is dominating over the East-West gradient in predicting the quantity of green vegetation fractions used in a spatial context. Here, we are using the same data and we can see that the North-South gradient does not contribute on the rate of change in green vegetation and its influence for temporal trends based on the three decades. Our results confirm the results of the author [13] where they concluded that latitude and longitude can be used to explain the spatial variability in the distribution of C3 and C4 grass along North-South and East-West gradients. In analysing 30 years as a spatio-temporal aspect in the four segments demonstrated in Table 4 we show an decrease of the influence of the North-South gradient and an increase of the East-West gradient as the relative influence of predicting vegetation trends. By analysing the data using either the whole data set of 30 years, three decades or four segments that show 30 years in each quadrant, we can conclude that in the shorter time frames no temporal trends

were observed and the overall linear trend of 30 years seems sufficient.

7. Appendix

It is also apparent that very few locations have a strictly non-zero gradient. In most of the cases the overall trend of green vegetation is not strong. One proposal is to transform the green vegetation fractions or use p-values as an additional covariate to weight observations for the BRT.

Table 5. Comparison of RMSE on the test data using all 30 years, first 10 years, middle 10 years and last 10 years and in four segmented areas comprising a 30 year time frame.

Scenario	RMSE of extracted slope coefficients	RMSE of extracted t-values
All 30 years	0.1150	2.8308
First 10 years	0.1112	0.9805
Middle 10 years	0.1214	0.7128
Last 10 years	0.1063	0.4465
Four segments		
1 - Upper left	0.1076	0.9719
2 - Upper right	0.0915	1.1589
3 - Lower left	0.1112	0.3884
4 - Lower right	0.1265	0.3884

Author Contributions: Brigitte Colin and Kerrie Mengersen conceived and designed this study; Brigitte Colin performed the experiments and wrote the paper; Kerrie Mengersen analysed the experimental results and edited the manuscript.

Funding: “This research received no external funding.”

Acknowledgments: This study was supported by the Australian Research Council (Grant No.:FL150100150). The authors wish to thank (i) Department of Environment and Science (DES) for providing the Fractional Cover data, (ii) Brodie Lawson for helpful comments, (iii) QUT ACEMS for providing office space and infrastructure to achieve this article and (iiii) the Reviewers and the Editors for their constructive comments and helpful suggestions, which have greatly improved this article.

Conflicts of Interest: “The authors declare no conflict of interest.”

References

1. Wulder, M.A.; Masek, J.G.; Cohen, W.B.; Loveland, T.R.; Woodcock, C.E. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment* **2012**, pp. 2–10. doi:10.1016/j.rse.2012.01.010.
2. J.Walsh, S.; Crawford, T.W.; Welsh, W.F.; A.Crews-Meyer, K. A multiscale analysis of LULC and NDVI variation in Nang Rong district, northeast Thailand. *Agriculture, Ecosystems & Environment* **2001**, *85*, 47 – 64. doi:10.1016/S0167-8809(01)00202-X.
3. Gallo, K.P.; Easterling, D.R.; Peterson, T.C. The Influence of Land Use/Land Cover on Climatological Values of the Diurnal Temperature Range. *Journal of Climate* **1996**, *9*, 2941–2944. doi:10.1175/1520-0442(1996)009<2941:TIO LUC>2.0.CO;2.
4. Datt, B. A New Reflectance Index for Remote Sensing of Chlorophyll Content in Higher Plants: Tests using Eucalyptus Leaves. *Journal of Plant Physiology* **1999**, *154*, 30 – 36. doi:10.1016/S0176-1617(99)80314-9.
5. Zhang, H.; Li, Q.Z.; Lei, F.; Du, X.; Wei, J.D. Research on rice acreage estimation in fragmented area based on decomposition of mixed pixels. *Remote Sensing and Spatial Information Sciences* **2015**, *40*, 133.
6. Guerschman, J.P.; Scarth, P.F.; McVicar, T.R.; Renzullo, L.J.; Malthus, T.J.; Stewart, J.B.; Rickards, J.E.; Trevithick, R. Assessing the effects of site heterogeneity and soil properties when unmixing photosynthetic vegetation, non-photosynthetic vegetation and bare soil fractions from Landsat and MODIS data. *Remote Sensing of Environment* **2015**, *161*, 12–26. doi:10.1016/j.rse.2015.01.021.

7. Adams, J.B.; Sabol, D.E.; Kapos, V.; Filho, R.A.; Roberts, D.A.; Smith, M.O.; Gillespie, A.R. Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the Brazilian Amazon. *Remote Sensing of Environment* **1995**, *52*, 137–154. doi:doi.org/10.1016/0034-4257(94)00098-8.
8. Roberts, D.A.; Smith, M.A.J. Green vegetation, nonphotosynthetic vegetation, and soils in AVIRIS data. *Remote Sensing of Environment* **1993**, *44*, 255–269. doi:10.1016/0034-4257(93)90020-X.
9. Zachary, T.; Dar, R.; Sander, V.; Angeles, C.; Carlos, R.; Susan, U. Evaluating Endmember and Band Selection Techniques for Multiple Endmember Spectral Mixture Analysis using Post-Fire Imaging Spectroscopy. *Remote Sensing* **2018**, *10*. doi:10.3390/rs10030389.
10. Scarth, P.F.; Röder, A.; Schmidt, M. Tracking Grazing pressure and climate interaction - The Role of Landsat Fractional Cover in time series analysis. *Proceedings of the 15th Australasian Remote Sensing and Photogrammetry Conference* **2010**, p. 13. doi:10.6084/m9.figshare.94250.
11. Scanlon, T.M.; Albertson, J.D.; Caylor, K.K.; Williams, C.A. Determining land surface fractional cover from NDVI and rainfall time series for a savanna ecosystem. *Remote Sensing of Environment* **2002**, *82*, 376–388. doi:10.1016/S0034-4257(02)00054-8.
12. Colin, B.; Schmidt, M.; Clifford, S.; Woodley, A.; Mengersen, K. Influence of Spatial Aggregation on Prediction Accuracy of Green Vegetation Using Boosted Regression Trees. *Remote Sensing* **2018**, *10*, 1260. doi:10.3390/rs10081260.
13. McNab, H.W.; Lloyd, T.F. Testing Ecoregions in Kentucky and Tennessee with Satellite Imagery and Forest Inventory Data: USDA Forest Service Proceedings – RMRS-P-56. 2009, number 10, pp. 1–19.
14. Paruelo, J.M.; Lauenroth, W.K. Relative Abundance of Plant Functional Types in Grasslands and Shrublands of North America. *Ecological Applications* **1996**, *6*, 1212–1224.
15. Živadinović, I.; Ilijević, K.; Gržetić, I.; Popović, A. Long-term changes in the eco-chemical status of the Danube River in the region of Serbia. *Journal of the Serbian Chemical Society* **2010**, *75*, 1125–1148. doi:10.2298/JSC091102075Z.
16. Waller, L.A.; Gotway, C.A. *Applied Spatial Statistics for Public Health Data*; Wiley Series in Probability and Statistics: New York, 2004.
17. Gaüzère, P.; Jiguet, F.; Devictor, V. Rapid adjustment of bird community compositions to local climatic variations and its functional consequences. *Global change biology* **2015**, *21*, 3367–3378. doi:10.1111/gcb.12917.
18. Zewdie, W.; Csaplovics, E.; Inostroza, L. Monitoring ecosystem dynamics in northwestern Ethiopia using NDVI and climate variables to assess long term trends in dryland vegetation variability. *Applied Geography* **2017**, *79*, 167–178. doi:https://doi.org/10.1016/j.apgeog.2016.12.019.
19. Parisien, M.A.; Parks, S.A.; Krawchuk, M.A.; Flannigan, M.D.; Bowman, L.M.; Moritz, M.A. Scale-dependent controls on the area burned in the boreal forest of Canada, 1980–2005. *Ecological Applications* **2011**, *21*, 789–805.
20. Colin, B.; Clifford, S.; Wu, P.; Rathmanner, S.; Mengersen, K. Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making. *Open Journal of Statistics* **2017**, *07*, 859–875.
21. Bureau of Meteorology. Climate Classification of Australia, 2016.
22. Schmetz, J.; Pili, P.; Tjemkes, S.; Just, D.; Kerkmann, J.; Rota, S.; Ratier, A. An introduction to Meteosat second generation (MSG). *Bulletin of the American Meteorological Society* **2002**, *83*, 977–992.
23. Geiger, B.; Carrer, D.; Franchisteguy, L.; Roujean, J.L.; Meurey, C. Land surface albedo derived on a daily basis from Meteosat Second Generation observations. *IEEE Transactions on Geoscience and Remote Sensing* **2008**, *46*, 3841–3856.
24. Fensholt, R.; Sandholt, I.; Stisen, S.; Tucker, C. Analysing NDVI for the African continent using the geostationary meteosat second generation SEVIRI sensor. *Remote Sensing of Environment* **2006**, *101*, 212–229. doi:10.1016/j.rse.2005.11.013.
25. R Development Core Team, 2008.
26. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *Journal of Animal Ecology* **2008**, *77*, 802–813. doi:10.1111/j.1365-2656.2008.01390.x.
27. Matteson, A. *Boosting the accuracy of your Machine Learning models*. Retrieved from <https://www.datasciencecentral.com/profiles/blogs/boosting-the-accuracy-of-your-machine-learning-models>, 2013.
28. Hastie, T.J.; Tibshirani, R. Generalized additive models. *Statistical Science* **1986**, *1*, 297–318. doi:10.1016/j.jcsda.2010.05.004.

29. Moisen, G.G.; Freeman, E.A.; Blackard, J.A.; Frescino, T.S.; Zimmermann, N.E.; Edwards, T.C. Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling* **2006**, *199*, 176–187. doi:10.1016/j.ecolmodel.2006.05.021.
30. Leathwick, J.R.; Elith, J.; Francis, M.P.; Hastie, T.; Taylor, P. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology-Progress Series* **2006**, *321*, 267–281. doi:10.3354/meps321267.
31. Kuhn, M. The caret Package. *Journal of Statistical Software* **2008**, *5*, 1–10.
32. Ridgeway, G. Generalized Boosted Models: A guide to the gbm package. *Compute* **2005**, pp. 1–12.
33. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
34. Liu, K.; Subbarayan, S.; Shoults, R.; Manry, M.; Kwan, C.; Lewis, F.; Naccarino, J. Comparison of very short-term load forecasting techniques. *IEEE Transactions on power systems* **1996**, *11*, 877–882.
35. Shi, J.; Lee, W.J.; Liu, Y.; Yang, Y.; Wang, P. Forecasting power output of photovoltaic systems based on weather classification and support vector machines. *IEEE Transactions on Industry Applications* **2012**, *48*, 1064–1069.
36. Methaprayoon, K.; Yingvivanapong, C.; Lee, W.J.; Liao, J.R. An integration of ANN wind power estimation into unit commitment considering the forecasting uncertainty. *IEEE Transactions on Industry Applications* **2007**, *43*, 1441–1448.

© 2019 by the authors. Submitted to *Sensors* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

7 Discussion

The overall aims of this thesis were to develop spatio-temporal decision tree models using big spatial data and geographic coordinates as surrogate variables, with application to the estimation of green vegetation cover using remotely sensed imagery. In this chapter, we review the major findings and contributions of the thesis and discuss their significance in the context of the current state of the art. These contributions will be mapped back to our research questions and aims outlined in Chapter 1. The chapters build upon each other and results from previous chapter will be incorporated in the following work. In summary, in Chapter 3 we assessed the general suitability of BRT for addressing our research aims. In Chapter 4 we aggregated green vegetation fractions to find a suitable spatial resolution for BRT to predict with satisfying accuracy and without loss of local characteristics of the study area. The identified best resolution was then used for Chapter 5 where we proposed a two-step approach to evaluate trends in green vegetation using extracted slope coefficients output from a linear model. Finally we presented in Chapter 6 an alternative aggregation, using spatial smoothing to increase the signal-to-noise ratio in the underlying green vegetation fractions. The chapter concludes with a discussion of strengths and limitations of the work presented in the thesis and areas of possible future research.

Chapter 3 presented an evaluation of BRT as a statistical machine learning algorithm that can address big data challenges to enable data-driven decisions and applications. Comparison with other regression methods such as Random Forest and Least Absolute Shrinkage and Selection Operator show that BRT outperforms these with respect to goodness of fit. Since BRT is often considered as a black box method, where little knowledge is known about its internal process and details of the implementation, it is important to stress that the BRT model offers a wide range of powerful methods to interpret the results. One example is a list of all covariates ranked in hierarchical order along with their contribution to explaining the response variable used in the splitting process. By identifying the individual relative influence of each covariate we can filter out less influential variables that can significantly decrease complexity and computational processing time while simultaneously maintaining a good model fit and low error rates.

Aim 1 was successfully completed by the research presented in this chapter. The encouraging results regarding the utility of BRTs for the problems considered in this thesis agree with previous studies also demonstrating the successful application of BRTs in ecology (Jafari et al., 2014; Leathwick et al., 2006; Stohlgren et al., 2010). This motivated the use of BRT for the following research aims addressed in the subsequent chapters.

Chapter 4 presented a data aggregation scheme to enable effective data processing of data-rich studies and we investigated how spatial aggregation affects prediction accuracy. A case study showed that the spatial resolution of 3000m achieved better prediction accuracy over the finer scale of 1500m. However, it was apparent that peak values were consistently under-predicted in all four spatial resolutions. BRT, as an ensemble approach, allows for a flexible partition of the feature space and therefore fine-scale characteristics in green vegetation cover could be maintained and predicted well, while the aggregation scheme significantly reduced the data volume. The computational cost of the delineation of green vegetation fractions was significantly reduced from 4 hours in the spatial aggregation resolution of 1500m to the next smaller resolution of 3000m to only 45 minutes. This is a significant decrease of computational cost and especially beneficial when dealing with a larger number of FCover scenes without the necessity of using high performance capabilities. In addition, we demonstrated that there is a stronger influence of the North-South gradient dominating the influence of the East-West direction. The proposed data reduction scheme is applicable more generally for studies where a spatio-temporal approach is desired because this requires processing of the scenes for each point in time, significantly increasing computational cost. In summary, Aim 2 was successfully completed and the research presented in this chapter demonstrates that it is not necessary to compute FCover imagery at full (30m) spatial resolution to achieve high predictive accuracy. The R code of the data delineation is published on my GitHub page: <https://github.com/BrigitteColin>

Chapter 5 presents an approach to using a Gaussian smoothing kernel on green fractional cover data to improve the signal-to-noise ratio in our green vegetation FCover data. Our aim is to gain better estimates and prediction results using smoothed data and we investigated in two scenarios; quantitatively by assessing the RMSE and qualitatively by comparing raster maps showing the smoothed imagery. Scenario 1 assess the RMSE on replacing the response variable with the smoothed values and concludes that the magnitude of the smoothing influences the goodness of fit. The RMSE starts to worsen when using a high value of sigma after reaching a minimum. In scenario 2 the smoothed green vegetation fractions were added as additional covariates to latitude and longitude and the response variable remained the same. We demonstrate that by adding an extra covariate to the BRT the prediction accuracy increases.

In the 2-dimensional raster maps we can see the magnitude of smoothing as either a very homogenised imagery showing less local features or as a smoothed imagery where local characteristics are maintained and where the signal-to-noise ratio has been improved and

extreme outliers were smoothed out. We can also prove that smoothing as a pre-processing step resolved any clusters in the data and the distribution of the smoothed values follow a now a unimodal and more symmetric distribution. With the application of a Gaussian smoothing kernel we demonstrates the successful completion of Aim 4.

Chapter 6 presented the idea of evaluating spatio-temporal patterns by first fitting a linear regression to each area to identify the temporal trend, and then assessing the spatial nature of these trends by using a BRT with the extracted slope coefficients as the response and geographic coordinates as predictors. At first the whole data set was used covering 30 years. However, a linear model captures only an overall monotonic rate of change over all 30 years and can not detect changing trends over time. To overcome this, we subdivided the 30 years to investigate if there are changing trends within the three decades that could be detected individually. A comparison of our results demonstrates, that there is no significant difference. However, we could see unexpected results showing slightly positive trends in the quantity of green vegetation cover in our study area. The estimated slope coefficients extracted from the linear model were then fed into the BRT to operate as our new response variable. Only geographic coordinates were used to predict spatio-temporal trends. In the next step, the FCover scene was divided in four even segments for which we investigated the spatio-temporal trajectories in the segments. By plotting the p-values gained from the linear model we discovered a gradient of statistically strong p-values as a North-South gradient across our FCover scene. We then inferred that there is variation of vegetation trends in the individual segments, which could be confirmed by plotting them separately as individual trajectories for each decade. With the completion of Aim 4, the method extends the modelling approach of Chapter 4 by a two step approach that resulted in a better understanding of the quantity of green vegetation over time and space, including localised spatial trends and their trajectories in our study area.

The combination of the four aims provide a methodology for a spatio-temporal analysis of FCover data in predicting green vegetation cover using geographic coordinates as co-variates. By achieving Aim 1 we demonstrated the suitability of BRT as a strong and flexible statistical machine learning method addressing various spatial data sets and data characteristics, achieving satisfying prediction accuracy and allowing insight in complex spatial relationships of data-driven analyses.

The methodology developed in Aim 2 demonstrates a successful data reduction scheme that achieved satisfying prediction results despite a massive decrease of data volume. Most importantly, one aggregation scheme was most suitable for BRT prediction and achieved the highest prediction accuracy by maintaining local characteristics in our heterogeneous region. We demonstrate that information delineation in data-rich studies is feasible and variability in our data set can be explained by the BRT model. Further, we gained location-based prediction densities shown as prediction raster maps that confirm

that our data reduction scheme enables an effective data processing in an decreased computational time. Aim 3 addresses spatial smoothing using Gaussian Processes and the investigation of the impact on prediction accuracy using smoothed versus non-smoothed data in two scenarios. Applying Gaussian smoothing kernels homogenise the imagery and the overall quality deteriorates due to suppressing change in high frequency signals and enhancing low frequency signals (Kumar, 2013). Extreme outliers were smoothed out and this resulted in a more symmetric and near Gaussian distribution in the smoothed data sets. We demonstrate that in replacing the response variable with the smoothed values (scenario 1) the prediction accuracy first increase, but when the value of σ is to high and only minimum smoothing is be applied, the RMSE starts to worsen. In scenario 2 we demonstrate that smoothing applied as a pre-processing step has a positive effect on the BRT predictions as supported in Figure 5.1. Aim 4 demonstrates that location based green vegetation trends will certainly be useful in other LULC analyses where remotely sensed imagery is used to detect spatio-temporal trends and their trajectories.

With the four aims it is possible to comprehensively describe the influence of spatial gradients such as the dominate North-South gradient on green vegetation, the influence of spatial aggregation and spatial smoothing on prediction accuracy and lastly the detection of spatio-temporal local trend in green vegetation using a two step approach. By satisfying all the aims we enable a better informed decision process with implications to the future by identifying areas that are sensitive to environmental impacts such as climate variability in semi-arid land. prone to change.

By meeting these research aims, several gaps in the literature were addressed and the overall aim of this thesis was accomplished. The research presented in this thesis also identified new directions for future research. BRT could be applied to other types of remotely sensed data. Examples include active sensors such as TerrSAR-S Radar, LIDAR, Laser altimeter, Sentinel 1, or passive sensors like Landsat, Spot, or MODIS. We have already mentioned other statistical models that could be used on remotely sensed data, such as other regression methods like GAM, LASSO, GLM, CART, NB, MARS and RF.

There are however some limitations to this research which must be outlined and discussed, and which may present future directions for research. Firstly, we only used geographic coordinates as covariates and no other environmental information has been added to enhance/support prediction accuracy. It is expected that with the addition of further covariates such as rainfall, soil type, vegetation indices et cetera. the prediction accuracy will be improved. A second important limitation is that we did not test our approaches on different locations using other FCover scenes in other climate zones, topology on different green vegetation characteristics. Our results and conclusions are restricted to our study area. Thirdly, the methods presented here have a substantial computational cost. Due do interaction effects between the tree nodes, BRT cannot be run in parallel. However,

it was possible to run our code in two individual R sessions by using each available core separately. Specification of the computational utility are described in paper 3 in Section 1.5.1 Hyperparameter tuning. Lastly, BRT offers a set of different loss functions. For our continuous response and its distributions we used the squared error loss, but there are alternatives to extend and improve the basic framework. For example, the absolute deviation and Huber loss function could be used to investigate if a different loss function minimizes the expectation further (Ridgeway, 2007).

Bibliography

- Ak, Ç., Ergönül, Ö., Şencan, İ., Torunoğlu, M. A., & Gönen, M. (2018). Spatiotemporal prediction of infectious diseases using structured gaussian processes with application to crimean–congo hemorrhagic fever. *PLoS neglected tropical diseases*, *12*(8), e0006737.
- Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 825–848.
- Bastin, G., Denham, R., Scarth, P., Sparrow, A., & Chewings, V. (2014). Remotely-sensed analysis of ground-cover change in Queensland’s rangelands, 1988-2005. *Rangeland Journal*, *36*(2), 191 – 204. doi: 10.1071/RJ13127
- Box, G. E. P. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, 201 – 236.
- Breiman, L. (1996). *Bias, Variance, and Arching Classifiers* (Tech. Rep.).
- Breiman, L. (1997). ARCING THE EDGE Leo Breiman Technical Report 486 , Statistics Department University of California, Berkeley CA. 94720. , *4*, 1-14.
- Breiman, L. (1998). Arcing classifiers. *Annals of Statistics*, *26*(3), 801-849. doi: 10.1214/aos/1024691079
- Brown, G., & Raymond, C. (2007). The relationship between place attachment and landscape values: Toward mapping place attachment. *Applied Geography*, *27*(2), 89 - 111. doi: <https://doi.org/10.1016/j.apgeog.2006.11.002>
- Bureau of Meteorology. (2016). *Climate Classification of Australia*. Retrieved 2016-04-12, from <http://www.australiaforeveryone.com.au/climate.htm>
- Carter, J. O., & Bruget, D. (2015). *AussieGRASS Environmental Calculator* (Tech. Rep. No. May). Brisbane: Department of Science, Information Technology and Innovation.
- Chen, H. (2017). Köppen climate classification [Computer software manual].
- Cruse, B., Liedloff, A. C., & Wintle, B. a. (2012). A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, *35*(10), 879-888. doi: 10.1111/j.1600-0587.2011.07138.x
- Cressie, N., Shi, T., & Lang, E. L. (2010). Fixed Rank Filtering for Spatio-Temporal Data. *Journal of Computational and Graphical Statistics*, *19*(3), 724-745. doi: 10.1198/jcgs.2010.09051

- De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, *88*(1), 243-251. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17489472> doi: 10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2
- De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, *81*(11), 3178-3192. doi: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2
- DeFries, R., & Chan, J. C.-W. (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, *74*(3), 503 – 515. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0034425700001425> doi: [https://doi.org/10.1016/S0034-4257\(00\)00142-5](https://doi.org/10.1016/S0034-4257(00)00142-5)
- Dube, T., Mutanga, O., Elhadi, A., & Ismail, R. (2014). Intra-and-inter species biomass prediction in a plantation forest: testing the utility of high spatial resolution spaceborne multispectral rapideye sensor and advanced machine learning algorithms. *Sensors*, *14*(8), 15348–15370.
- Dubin, R. A. (1988, August). Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *The Review of Economics and Statistics*, *70*(3), 466-474.
- Elith, J., & Leathwick, J. (2017). Boosted Regression Trees for ecological modeling [Computer software manual]. Retrieved from <http://cran.r-project.org/web/packages/dismo/vignettes/brt.pdf>
- Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802-813. doi: 10.1111/j.1365-2656.2008.01390.x
- Emelyanova, I. V., McVicar, T. R., Van Niel, T. G., Tao Li, L., & Van Dijk, A. I. J. M. (2013). Remote Sensing of Environment Assessing the accuracy of blending Landsat – MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics : A framework for algorithm selection. *Remote Sensing of Environment*, *133*, 193-209. doi: 10.1016/j.rse.2013.02.007
- F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., ... Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, *30*(5), 609-628. doi: 10.1111/j.2007.0906-7590.05171.x
- Forward, G. R. (2009). *Assessment of ground cover monitoring sites in south australia*. Department of Water, Land and Biodiversity Conservation.
- Freund, Y., & Schapire, E., Robert. (1996). Experiments with a New Boosting Algorithm. *International Conference on Machine Learning*, 148-156. doi: 10.1.1.133.1040
- Friedman, J. (2006). Recent Advances in Predictive (Machine) Learning. *Journal of Classification*(23), 175. doi: 10.1007/s00357-006-0012-4
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367-378. doi: 10.1016/S0167-9473(01)00065-2

- Gallo, K. P., Easterling, D. R., & Peterson, T. C. (1996). The influence of land use/land cover on climatological values of the diurnal temperature range. *Journal of Climate*, *9*(11), 2941-2944. doi: 10.1175/1520-0442(1996)009<2941:TIOLOC>2.0.CO;2
- Gelfand, A. E., Schmidt, A. M., Banerjee, S., & Sirmans, C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, *13*(2), 263-312.
- George, E. I. (2000). The Variable Selection Problem. *Journal of the American Statistical Association*, *95*(452), 1304-1308. Retrieved from <http://www.jstor.org/stable/2669776?origin=crossref> doi: 10.2307/2669776
- Guerschman, J. P., Scarth, P. F., McVicar, T. R., Renzullo, L. J., Malthus, T. J., Stewart, J. B., ... Trevithick, R. (2015). Assessing the effects of site heterogeneity and soil properties when unmixing photosynthetic vegetation, non-photosynthetic vegetation and bare soil fractions from Landsat and MODIS data. *Remote Sensing of Environment*, *161*(March), 12-26. doi: 10.1016/j.rse.2015.01.021
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Second ed.). New York: Springer Series in Statistics. doi: 10.1007/b94608
- Hastie, T. J., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297-318. doi: 10.1016/j.csda.2010.05.004
- Humphries, G. R. W. (2015). Estimating regions of oceanographic importance for seabirds using a-spatial data. *PLoS ONE*, *10*(9), 1-15. doi: 10.1371/journal.pone.0137241
- Irons, J. R. (2018, December 10). Landsat Science, Practical Uses [Computer software manual].
- Jacobs, K. (1992). Independent identically distributed (iid) random variables. In *Discrete stochastics* (pp. 65-101). Springer.
- Jafari, A., Khademi, H., Finke, P. A., Van de Wauw, J., & Ayoubi, S. (2014). Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. *Geoderma*, *232-234*, 148-163. doi: 10.1016/j.geoderma.2014.04.029
- Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, *31*(8), 651-666.
- Jin, Y., Xu, B., Yang, X., Qin, Z., Li, J., Zhao, F., ... Wu, Q. (2014). Grassland aboveground biomass retrieval from remote sensing data by using artificial neural network in temperate grassland, northern china. In (pp. 1-6).
- J. Walsh, S., Crawford, T. W., Welsh, W. F., & A. Crews-Meyer, K. (2001). A multiscale analysis of lulc and ndvi variation in nang rong district, northeast thailand. *Agriculture, Ecosystems & Environment*, *85*(1), 47 - 64. doi: 10.1016/S0167-8809(01)00202-X
- Kamal, M., & Phinn, S. (2011). Hyperspectral data for mangrove species mapping: A comparison of pixel-based and object-based approach. *Remote Sensing*, *3*(10), 2222-2242.
- Karnieli, A., Gilad, U., Ponzet, M., Svoray, T., Mirzadinov, R., & Fedorina, O. (2008). Assessing land-cover change and degradation in the central asian deserts using satellite image processing and geostatistical methods. *Journal of Arid Environments*,

- 72(11), 2093–2105.
- Kuhn, M. (2008). The caret Package. *Journal of Statistical Software*, 5(28), 1-10.
- Kuhn, M. (2015). A Short Introduction to the caret Package [Computer software manual]. Retrieved from Retrievedfromcran.r-project.org/web/packages/caret/vignettes/caret.pdf
cran.r-project.org/web/packages/caret/vignettes/caret.pdf
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling [Hardcover]*. Retrieved from http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/ref=pdf_b_img_z doi: 10.1007/978-1-4614-6849-3
- Kumar, B. S. (2013). Image denoising based on non-local means filter and its method noise thresholding. *Signal, image and video processing*, 7(6), 1211–1227.
- Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., & Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology-Progress Series*, 321, 267-281. doi: 10.3354/meps321267
- Li, M., Im, J., & Beier, C. (2013). Machine learning approaches for forest classification and change analysis using multi-temporal landsat tm images over huntington wildlife forest. *GIScience & Remote Sensing*, 50(4), 361–384.
- Lindquist, E., & D’Annunzio, R. (2016, August). Assessing Global Forest Land-Use Change by Object-Based Image Analysis. *Remote Sensing*, 8(8), 678. Retrieved from <http://www.mdpi.com/2072-4292/8/8/678> doi: 10.3390/rs8080678
- Lopes, A., Touzi, R., & Nezry, E. (1990). Adaptive speckle filters and scene heterogeneity. *IEEE transactions on Geoscience and Remote Sensing*, 28(6), 992–1000.
- Malenovský, Z., Bartholomeus, H. M., Acerbi-Junior, F. W., Schopfer, J. T., Painter, T. H., Epema, G. F., & Bregt, A. K. (2007). Scaling dimensions in spectroscopy of soil and vegetation. *International Journal of Applied Earth Observation and Geoinformation*, 9(2), 137–164.
- Matteson, A. (2013). Boosting the accuracy of your machine learning models [Computer software manual].
- Mazzotti, F. J., Hughes, N., & Harvey, R. G. (2007). *Why do we need environmental monitoring for Everglades restoration?* (Tech. Rep. No. November). Fort Lauderdale, University of Florida: Institute of Food and Agricultural Sciences (IFAS).
- McCarthy, M. A., & Possingham, H. P. (2007). Active adaptive management for conservation. *Conservation Biology*, 21(4), 956-963. doi: 10.1111/j.1523-1739.2007.00677.x
- Miller, H. J. (2004). Tobler ’ s First Law and Spatial Analysis. *Annals of the Association of American Geographers*, 94(2), 284-289. doi: 10.1111/j.1467-8306.2004.09402005.x
- Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E., & Edwards, T. C. (2006). Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling*, 199(2), 176 – 187. doi: 10.1016/j.ecolmodel.2006.05.021

- Muir, J., Schmidt, M., Tindall, D., Trevithick, R., Scarth, P., & Stewart, J. (2011). *Field measurement of fractional ground cover: a technical handbook supporting ground cover monitoring for australia*. Queensland Department of Environment and Resource Management for the Australian Bureau of Agricultural and Resource Economics and Sciences, Brisbane, Australia.
- Müller, D., Leitão, P. J., & Sikor, T. (2013). Comparing the determinants of cropland abandonment in Albania and Romania using boosted regression trees. *Agricultural Systems*, *117*, 66 – 77. doi: 10.1016/j.agsy.2012.12.010
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(DEC). doi: 10.3389/fnbot.2013.00021
- Owen, A. (1984). A neighbourhood-based classifier for landsat data. *Canadian Journal of Statistics*, *12*(3), 191 – 200. doi: 10.2307/3314747
- Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57–65.
- Park, S., Im, J., Jang, E., & Rhee, J. (2016). Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. *Agricultural and forest meteorology*, *216*, 157–169.
- Pourghasemi, H. R., & Rahmati, O. (2018). Prediction of the landslide susceptibility: Which algorithm, which precision? *CATENA*, *162*, 177 - 192. doi: <https://doi.org/10.1016/j.catena.2017.11.022>
- Pugh, M., & Waxman, A. (2006). Classification of spectrally-similar land cover using multi-spectral neural image fusion and the fuzzy artmap neural classifier. In *Geoscience and remote sensing symposium, 2006. igarss 2006. ieee international conference on* (pp. 1808–1811).
- R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- R Development Core Team. (2008). Vienna, Austria. Retrieved from <http://www.r-project.org/le>
- Reiche, J., de Bruin, S., Hoekman, D., Verbesselt, J., & Herold, M. (2015). A Bayesian Approach to Combine Landsat and ALOS PALSAR Time Series for Near Real - Time Deforestation Detection. *Remote Sensing*, *7*(5), 4973-4996. Retrieved from <http://www.mdpi.com/2072-4292/7/5/4973/> doi: 10.3390/rs70504973
- Ridgeway, G. (2005). Generalized Boosted Models: A guide to the gbm package. *Compute*(1), 1-12.
- Ridgeway, G. (2007). Generalized Boosted Models : A guide to the gbm package. *Compute*, *1*(4), 1-12. Retrieved from <http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf> doi: 10.1111/j.1467-9752.1996.tb00390.x
- Rigge, M., Smart, A., Wylie, B., & Kamp, K. V. (2014). Detecting the Influence of Best Management Practices on Vegetation Near Ephemeral Streams With Landsat Data. *Rangeland ecology & management*, *67*(1), 1 – 8. doi: 10.2111/REM-D-12-00185.1

- Rizzo, D., Martin, L., & Wohlfahrt, J. (2014). Miscanthus spatial location as seen by farmers: A machine learning approach to model real criteria. *Biomass and Bioenergy*, 66(0), 348 – 363. doi: 10.1016/j.biombioe.2014.02.035
- Robinzonov, N. (2013). *Advances in boosting of temporal and spatial models* (Doctoral dissertation, Ludwig-Maximilians-Universität München). Retrieved from <http://edoc.ub.uni-muenchen.de/15338/>
- Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93 – 104. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0924271611001304> doi: <https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, 24 (February 2016), 4-15. doi: 10.7152/acro.v24i1.14671
- Sancetta, A., et al. (2016). Greedy algorithms for prediction. *Bernoulli*, 22(2), 1227–1277.
- Sarker, C., Alvarez, L. M., & Woodley, A. (2016). Integrating Recursive Bayesian Estimation with Support Vector Machine to Map Probability of Flooding from Multispectral Landsat Data. *International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016*. doi: 10.1109/DICTA.2016.7797054
- Scarth, P. (2012, March). *Fractional cover - landsat, joint remote sensing research program algorithm, australia coverage* (Tech. Rep.). TERN AusCover. Retrieved from <http://data.auscover.org.au/xwiki/bin/view/Product+pages/Landsat+Fractional+Cover>
- Scarth, P., Byrne, M., Danaher, T., Henry, B., Hassett, R., Carter, J., & Timmers, P. (2006). State of the paddock: monitoring condition and trend in groundcover across Queensland. *13th Australasian Remote Sensing and Photogrammetry Conference (ARSPC)*, 11.
- Scarth, P., Röder, A., & Schmidt, M. (2010). Tracking Grazing pressure and climate interaction - The Role of Landsat Fractional Cover in time series analysis. *Proceedings of the 15th Australasian Remote Sensing and Photogrammetry Conference*, 13. doi: 10.6084/m9.figshare.94250
- Schmidt, M., Carter, J., Stone, G., & O'Reagain, P. (2016). Integration of Optical and X-Band Radar Data for Pasture Biomass Estimation in an Open Savannah Woodland. *Remote Sensing*, 8(12), 989. doi: 10.3390/rs8120989
- Schmidt, M., Thamm, H.-P., Menz, G., & Bénes, T. E. (2003). Long term vegetation change detection in an and environment using LANDSAT data. *Geoinformation for European-Wide Integration, Millpress, Rotterdam*, 145–154.
- Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.
- Stohlgren, T. J., Ma, P., Kumar, S., Rocca, M., Morissette, J. T., Jarnevich, C. S., & Benson, N. (2010). Ensemble habitat mapping of invasive plant species. *Risk Analysis*, 30(2), 224-235. doi: 10.1111/j.1539-6924.2009.01343.x

- Tarling, R. (2009). *Statistical Modelling for Social Researchers: Principles and Practice*. London and New York: Taylor & Francis Group. doi: 10.1111/j.1467-985X.2009.00624_14.x/abstract
- Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Liu, D. L., ... Sides, T. (2018). Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern australia. *Ecological Indicators*, 88, 425 – 438. doi: <https://doi.org/10.1016/j.ecolind.2018.01.049>
- Wang, S. E. (2013). *Analysis of Time Series Models with Iterated Boosting*. University of California, Los Angeles.
- Wikipedia, t. f. e. (2018, December 26). R (programming language) [Computer software manual].
- Williams, B. K. (2011). Adaptive management of natural resources-framework and issues. *Journal of Environmental Management*, 92(5), 1346 – 1353. doi: 10.1016/j.jenvman.2010.10.041
- Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning. *the MIT Press*, 2(3), 4.
- Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning* (pp. 1067–1075).
- Wolfe, R. (2018, December 10). National Aeronautics and Space Administration, Goddard Space Flight Center, MODIS Land [Computer software manual].
- Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*.
- Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., & Woodcock, C. E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of landsat. *Remote Sensing of Environment*(122), 2-10. doi: 10.1016/j.rse.2012.01.010
- Zhang, H., Li, Q. Z., Lei, F., Du, X., & Wei, J. D. (2015). Research on rice acreage estimation in fragmented area based on decomposition of mixed pixels. *Remote Sensing and Spatial Information Sciences*, 40(7), 133.