KING'S COLLEGE LONDON

University of London

M.Sc. Environmental Monitoring, Modelling and Management

Mapping deforestation stages and spatial patterns in the Amazon rainforest using fractal analysis and data mining techniques

Alejandro Coca Castro

(Student: 1415431)

Year of Submission: 2015

This dissertation is submitted as part of an MSc degree in Environmental Monitoring, Modelling and Management at King's College London.

Declarations and Dedications

KING'S COLLEGE LONDON

University of London

DEPARTMENT OF GEOGRAPHY

MA/MSc DISSERTATION

I, Alejandro Coca Castro hereby declare (a) that this Dissertation is my own original work and that all source material used is acknowledged therein; (b) that it has been specially prepared for a degree of the University of London; and (c) that it does not contain any material that has been or will be submitted to the Examiners of this or any other university, or any material that has been or will be submitted for any other examination.

This Dissertation is 11,800 words.

Signed: ....

Date: .28/08/2015

Abstract

Several analyses can be derived from the land-change information provided by remote sensing image datasets. This study examines deforestation stages and spatial patterns information provided by two methodological frameworks (fractal analysis and data mining) in the Amazon rainforest and how they may be affected by two deforestation dataset sources (Terra-i and Global Forest Change) and extent of analysis (four increasing fixed grids sizes). Analysis shows that for the fractal analysis framework, the highest values of the output information (fractal dimension) was obtained using the Terra-i dataset and at the finest grid sizes. The data mining results suggest the feasibility of artificial neural networks for mapping spatial patterns. The use of this algorithm in combination with a grid size of 30720 m provided the best true model performance using either the GFC or Terra-i datasets (Kappa values of 0.73 and 0.70, respectively).

Key works: deforestation, stages, spatial pattern, fractal analysis, data mining

Contents

List of Tables

List of Figures

List of Abbreviations

ANN(s) – Artificial Neural Network(s)

AOI – Area of Interest

C4.5 – Decision tree algorithm type

CF – Confidence level for C4.5 algorithms

DM – Data mining

DTs – Decision trees

FD or D – Fractal dimension

FNN - Feed-forward network

GFC – Global Forest Change

IGH - Interrupted Goode Homolosine projection

INPE - Brazilian National Institute for Space Research

IQR – Boxplot interquartile range

k – Number of equal subsets that split a full training set in cross validation technique

KDD - Knowledge Discovery in Databases

km – Kilometres

m – Metres

MLP - Multilayer Perceptron

MODIS - The Moderate-resolution Imaging Spectroradiometer

RAISG – The Amazon Geographic Socio-Environmental Information Network

RI_x – Relative variable importance in ANNs models

sq. metres – Square metres

TRMM –Tropical Rainfall Measuring Mission

κ – Kappa metric

Acknowledgements

Firstly, my thanks to Dr Mark Mulligan for providing his inputs, support and invaluable feedback as supervisor of this research. I am also indebted to Dr. Jin Sung, from University of Michigan, who kindly shared the codes for implementing the fractal analysis performed in this research.

I proudly admit the irreplaceable experience with the staff and multicultural atmosphere from my former institution (The International Center of Tropical Agriculture - CIAT), in special with the Terra-i project. The research skills acquired in this experience combined with the knowledge obtained during the master course at King's were fundamental to perform robust methodologies for analysing a large area such as the Amazon rainforest. As part of the same institution, I am personally glad with Andy Jarvis, Glenn Hyman, Louis Reymondin and Simon Cook for their unconditional support, which has been relevant to my formation as human, and my young career as a scientist.

Finally, I do not have enough words to express my gratitude to relatives and friends who were supportive to conclude this work and to pursue a new experience out of my home country Colombia. Special thanks to my dear parents and brothers with whom I have celebrated common achievements from the very beginning and throughout. Another special thank to my friend Caity who provided a significant English revision of this manuscript and of course I am glad for her unique friendship until nowadays.

1
Introduction

The production of geospatial data related to land use and land cover changes by governments and civil society organizations has vastly increased during the last decade (Coca, 2015). Going beyond the valuable information (location, rates and absolute values) provided by these datasets, it is important to have a better understanding of the spatial configurations and composition of the detected change areas at multiple spatial resolutions and time periods (Li and Reynolds, 1994; Turner *et al.*, 2001).

In the case of forested areas, zooming in to a particular region like the Amazon permits the visualisation of differences in the shapes (composition) and distribution (configuration) of deforested areas (Figure 1) (Coca, 2015). These structural aspects or stages of deforestation, usually denoted as "deforestation spatial patterns", may be linked with known land-use processes and agents (e.g small-scale farmers, large plantations, cattle ranchers) that promote forest disturbances on the ground (Lambin *et al.* 2003). Moreover, Pritz *et al.* (2012) determined that impacts in terms of biodiversity (species richness) differ among types of deforestation spatial patterns.

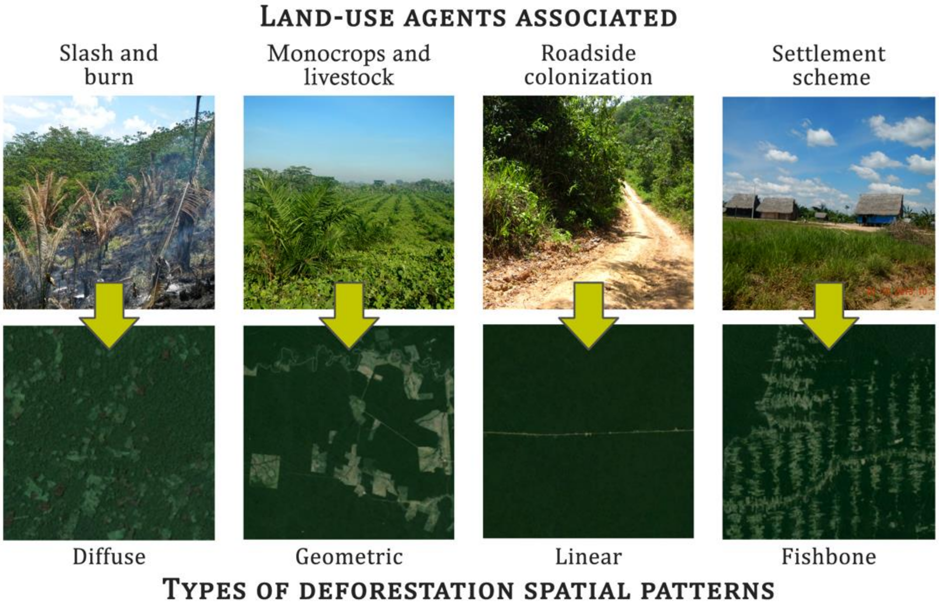


Figure 1. Relationships between four types of deforestation spatial patterns, visually differentiated using Google Earth Imagery, and related land-use agents. Photos provided by the Terra-i project (2015).

As a result, the identification and mapping of spatial patterns and stages in deforested areas is key for providing multi-level, flexible information on land use to forest conservation groups, land managers and related practitioners. Moreover, knowledge on the spatial configuration and composition of deforestation patterns assists in the modelling of future land-use scenarios (Coca, 2015).

The main aim of this research was to explore techniques to map types of deforestation spatial patterns and stages in the Amazon rainforest from existing remote-sensing image databases in order to contribute to knowledge on deforested landscapes. The specific objectives were:

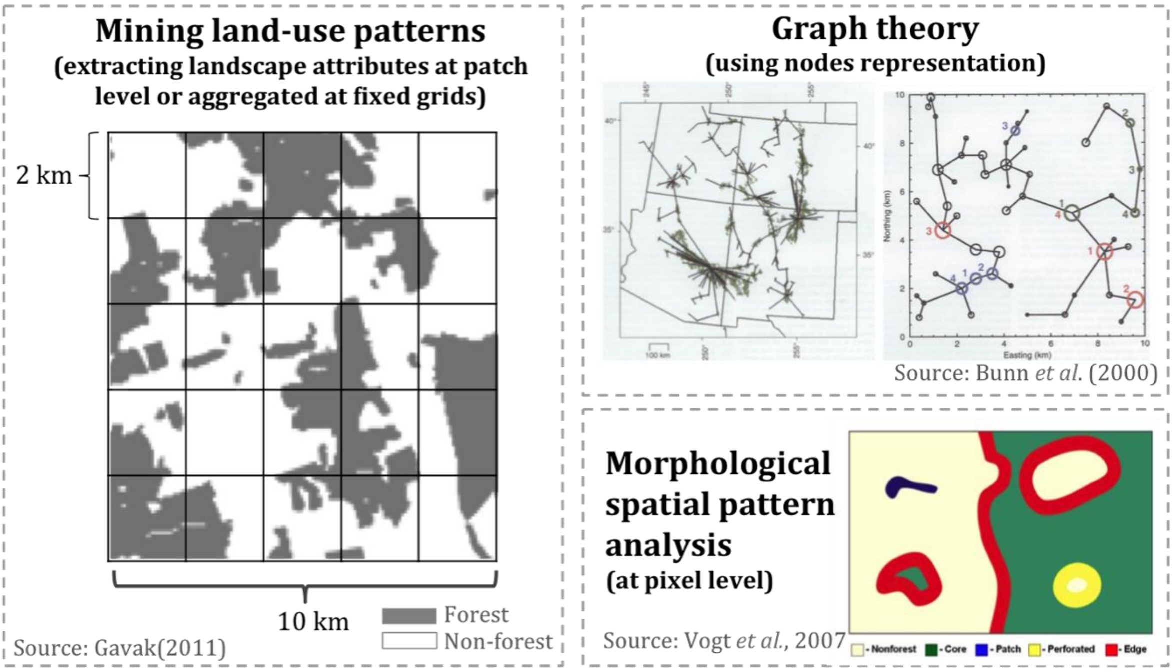
- Create cumulative maps of recent deforestation (2004-2013) from two existing remote-sensing image datasets (Terra-i and Global Forest Change);
- Characterize and map types of deforestation spatial patterns by dataset using fractal dimension, landscape fragmentation metrics and data mining techniques;
- Compare and discuss the results of deforestation spatial patterns and stages mapping techniques between and within deforestation datasets.

This manuscript begins with a brief literature review about the main developments and concepts involved in the study of deforestation spatial patterns and associated techniques. Next, the study area and the main pre-processing features performed for the deforestation datasets are described. Two potential approaches are presented to discriminate deforestation patterns types in the target area. On the one hand, the five patterns types given by Sun *et al.* (2014) are mapped using fractal analysis techniques. On the other hand, building mostly on earlier works by the Brazilian National Institute for Space Research (Silva *et al.* 2011), four patterns types are mapped under a "Knowledge Discovery in Databases" (KDD) approach. The latter involves first the use of data mining algorithms for classification with landscape-ecology metrics. Then the modelling and mapping results are exposed and confronted analysing the effects of the grain size (spatial resolution) of the datasets, the extent (unit of analysis) and the approaches assessed. Finally, the main findings and gaps in the patterns mapping methodologies implemented are summarized, including a list of implications and recommendations for future research. Main programming scripts and other supporting information can be found in the Appendices section.

2
Literature review

2.1 General perspectives on deforestation spatial patterns mapping

Several analyses can be derived from the land-change information provided by remote sensing image datasets (Silva *et al*., 2011). Among these, a collection of techniques and methodologies facilitate the understanding of deforested landscapes and their dynamics (Sun *et al.,* 2014) (Figure 2). These have involved the use of the mining land-use patterns method (Silva *et al*., 2011), moving window analysis (Riitters *et al*., 2002; Zurlini *et al*., 2007), graph theory (Baggio *et al*., 2011; Bunn *et al*., 2000; Minor and Urban, 2008; Saura and Pascual-Hortal, 2007; Maciel, 2012), the normalised spectral entropy index (Sun and Southworth, 2013a; Zaccarelli *et al*., 2012) and morphological spatial pattern analysis (Vogt *et al*., 2007).


Mining land-use patterns
(extracting landscape attributes at patch level or aggregated at fixed grids)

Source: Gavak(2011)

Graph theory
(using nodes representation)

Source: Bunn *et al.* (2000)

Morphological spatial pattern analysis
(at pixel level)

Source: Vogt *et al*., 2007

Figure 2. Some examples of methodologies for analysing spatial patterns and elements of deforested areas. Refer to the sources provided for further information on these methodologies.

Of these methods, the mining land-use patterns-based analysis has been widely used not only to characterise the spatial configurations and composition of deforested areas, but also to determine the associated agents of change (small-scale farmers, large plantations, cattle ranchers and road construction, among others) (Batistella *et al.*, 2003; Ewers and Lawrence, 2006; Geist and Lambim, 2001; Husson *et al.*, 1995; Mertens and Lambin, 1997). This method has primarily been supported by the extraction of landscape-ecology metrics from deforestation objects that are subsequently analysed by data mining techniques. Although these metrics are mainly extracted using specialised software for assessing landscape fragmentation such as the FRAGSTATS analysis package (McGarigal and Marks, 1995), most of them fail to properly describe irregular spatial patterns due to their Euclidean geometry basis (Sun *et al.*, 2014). To address this issue, it is important to mention the recent advances and potential use of fractal analysis to characterise spatial stages of deforested areas or also denoted by the authors (Sun and Southworth, 2013b; Sun *et al.*, 2013; Sun *et al.*, 2014) as developed areas.

2.1.1 The mining land-use patterns-based analysis

The mining land-use patterns approach consists of the extraction of object properties (landscape-ecology metrics) that are then used to discriminate, through data mining techniques, known deforestation spatial pattern types in a target area. In accordance with Silva (2015), this approach can be further divided into two sub-approaches: the landscape object (patch) approach or the fixed grid or window aggregation approach. The former analyses the geometric structure by individual patches. The latter aggregates a set of landscape objects representing a distinct occupation pattern in a fixed grid or window size. Although both methods have strengths and limitations, the latter seems to be more feasible in computational terms than the former for application to larger areas (a maximum reported area of 700,000 sq. metres as opposed to 150,000 sq. metres; see Silva *et al.*, 2011).

According to Silva *et al.* (2011), the data mining technique involves the study of *land-change objects*, i.e. individual or aggregated closed area(s) detected in a remote sensing image and associated with a change in land cover. The authors' method consists of two parts, both illustrated in Figure 3. The first part is the training procedure, in which the analyst defines the spatial pattern typologies according to an application domain (patch or grid-aggregated patches), linking them to established knowledge about the agents that cause land change. The expert selects a training set of land-change objects and labels each one according to the different types of spatial patterns that they represent. Each object has a set of geometrical properties that can be described using landscape-ecology metrics. The output is a training set of objects, where each member has a label and a set of numerical features. The training set is used to build a data mining algorithm for classification (decision-tree classifier), whereupon each type of land-change object is discriminated based on its patch metrics.

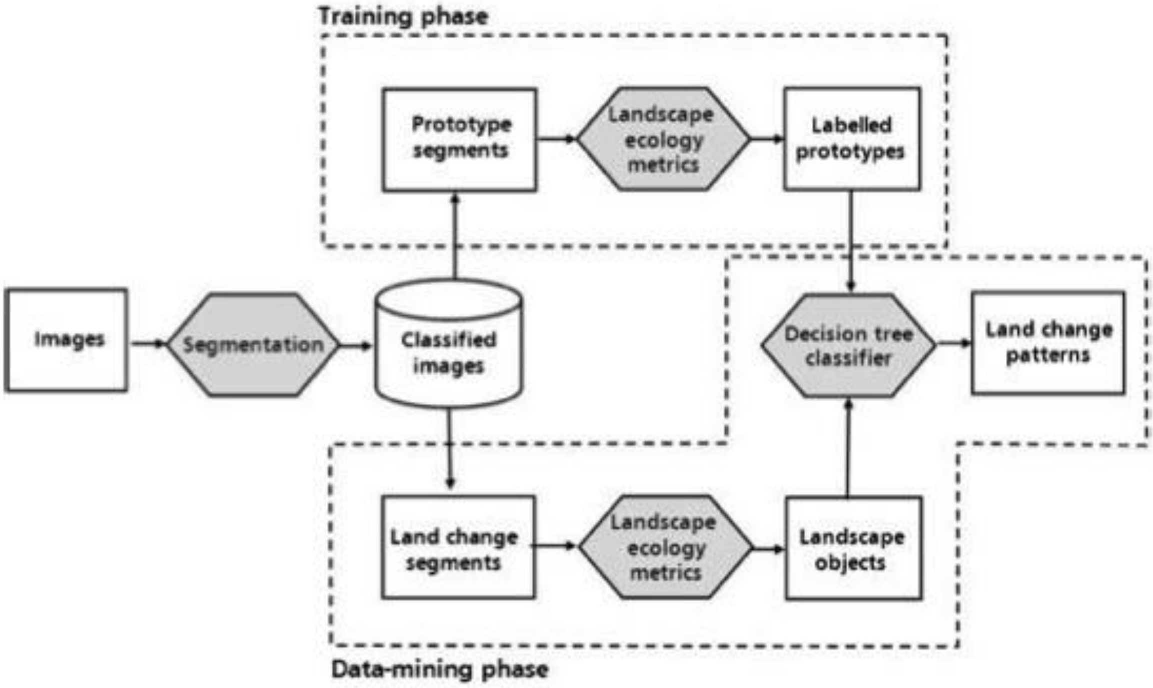


Figure 3. Workflow of the mining land-use patterns-based method. Source: Silva *et al.* (2011).

The second step of the method is the data-mining procedure. The analyst computes a set of numerical attributes for all land-change objects using landscape-ecology metrics. The data mining algorithm for classification (built in the training procedure) maps each land-change object to one of the user-defined spatial patterns. Finally, the user performs an analysis of the spatial and/or temporal trends of the resulting land-change patterns.

Figure 4 illustrates a case study in a 190,000 sq. km area in the Brazilian Amazon where the above methodology was implemented. For this case, using the patch grid-aggregation approach at a fixed grid of 10 km x 10 km, six spatial patterns of deforestation were initially identified. All unlabelled land-change objects were automatically classified using a built decision tree classifier created by a training set.

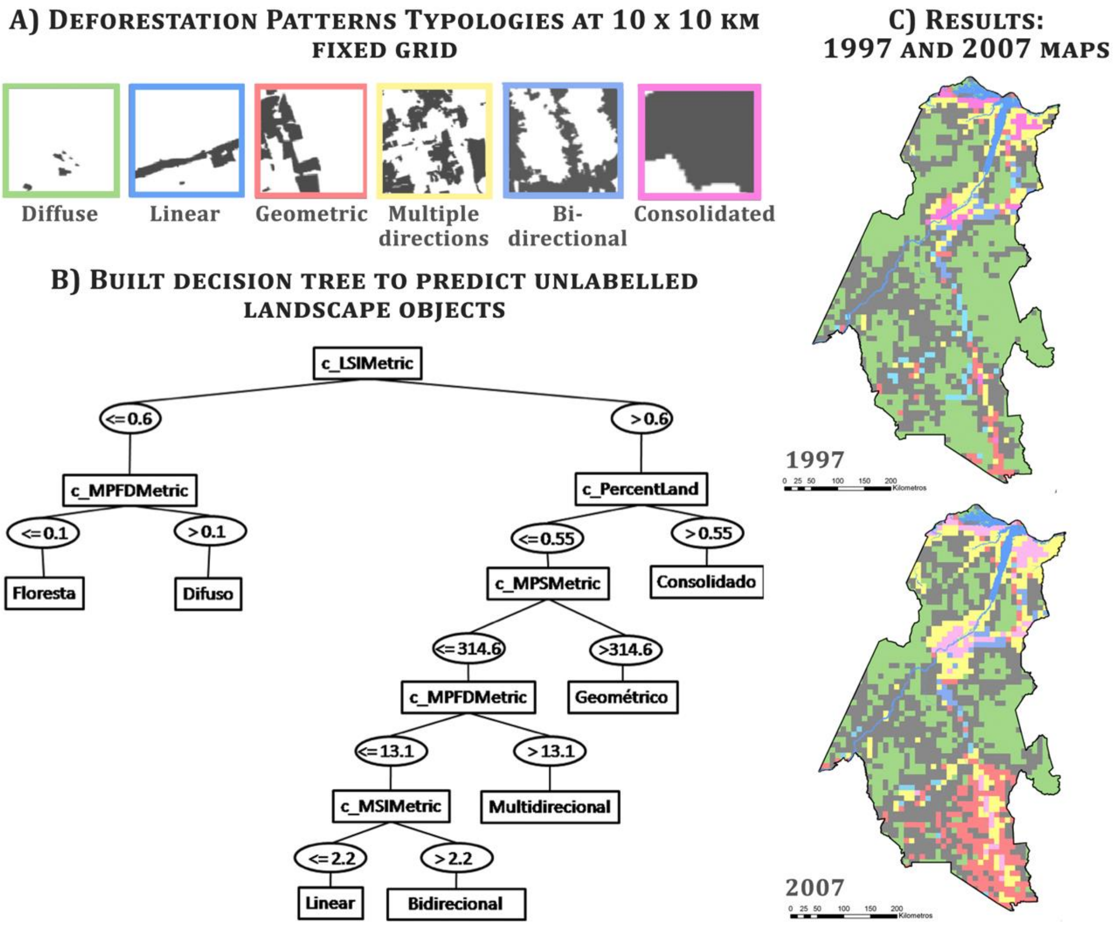


Figure 4. Example of the application of the mining land-use patterns-based analysis. Six spatial patterns (A) were identified at the unit of analysis (10 x 10 km) from PRODES deforestation dataset (30 m). Using a built decision tree (B) these patterns were discriminated for multiple periods (C). Colours by pattern box in (A) are related with grid colors in (C). Adapted from Gavak (2011).

2.1.2 The fractal analysis approach

Fractal geometry can provide a mathematical description for many natural forms such as coastlines, mountains, and clouds. Originally introduced by Mandelbrot (1975), fractals are defined as objects that display self-similarity at various scales and can be usually characterized by computing the fractal dimension (FD).

Although there are different approaches proposed to estimate the FD, they can be grouped into three major categories: the box-counting methods, the variance methods, and the spectral methods (Li *et al.*, 2009). Of these approaches, the box-counting approach, which is widely used due to its simplicity and automatic computability (Peitgen *et al.*, 1992 cited by Li *et al.*, 2009), was recently proposed for analysing spatiotemporal dynamics of forest clearings from remote-sensing image datasets (Sun and Southworth, 2013b; Sun *et al*., 2013; Sun *et al*., 2014).

These studies on spatio-temporal dynamics have computed the FD from multi-temporal Landsat (30m) forest/non-forest maps in the Western Amazon. After assessing several techniques to determine an optimal and consistent method to cartographically represent the fractal structures using AOIs (areas of interest) in heterogeneous landscapes, the authors claimed the high feasibility of the bottom-up method using a fixed-grid scans strategy. According to this method, a series of regular grids of decreasing box sizes are recursively superimposed over a target object (deforested area). The counting, or the number of boxes occupied by the target object, is recorded for each box (Sun *et al*., 2013). Five deforestation stages (Figure 5) were characterised using this approach. From a practical point of view, the fractal dimensions extracted can be used to indicate the spatial fill capacity, or the extent to which deforested areas can occupy the entire box area. The more clearings and conversion in a grid area, the closer the fractal dimension approaches the value of 2.

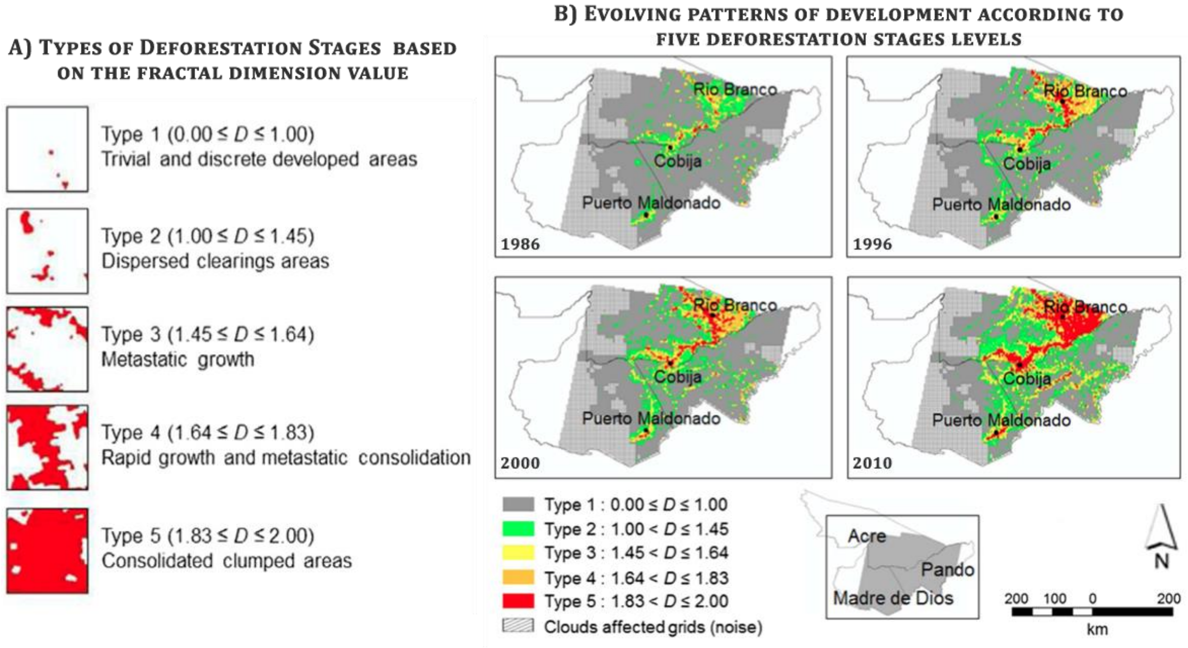

A) Types of Deforestation Stages based on the fractal dimension value
Type 1 (0.00 ≤ D ≤ 1.00) Trivial and discrete developed areas
Type 2 (1.00 ≤ D ≤ 1.45) Dispersed clearings areas
Type 3 (1.45 ≤ D ≤ 1.64) Metastatic growth
Type 4 (1.64 ≤ D ≤ 1.83) Rapid growth and metastatic consolidation
Type 5 (1.83 ≤ D ≤ 2.00) Consolidated clumped areas
B) Evolving patterns of development according to five deforestation stages levels
Rio Branco
Cobija
Puerto Maldonado
1986
1996
2000
2010
Type 1 : 0.00 ≤ D ≤ 1.00
Type 2 : 1.00 < D ≤ 1.45
Type 3 : 1.45 < D ≤ 1.64
Type 4 : 1.64 < D ≤ 1.83
Type 5 : 1.83 < D ≤ 2.00
Clouds affected grids (noise)
Acre
Pando
Madre de Dios
N
200 100 0 200
km

Figure 5. Types of deforestation stages (A) according to the fractal dimension *D* computed using the box-counting approach and bottom-up method. These stages were mapped on multi-temporal Landsat forest/non-forest maps in the Western Amazon (B). Adapted from Sun *et al*. (2014).

2.2 Data mining

As the amount of data available has increased considerably in last decades, the need is imminent to transform these into useful information and knowledge (Han and Kamber, 2006). A common technique that addresses this need is data mining (DM). Data mining is the process of using one or more computer learning techniques (algorithms) to automatically analyse and extract knowledge from data contained within a database (Roiger and Geatz, 2003).

The scientific method behind the data mining process, denoted as Knowledge Discovery in Databases (KDD), includes a methodology for extracting and preparing data as well as making decisions about actions to be taken once data mining has taken place (Roiger and Geatz, 2003) (Figure 6). KDD includes multidisciplinary activities that aim to develop new information from existing databases, and its major application areas are in marketing, fraud detection, telecommunication and manufacture (Fayyad *et al.*, 1996).

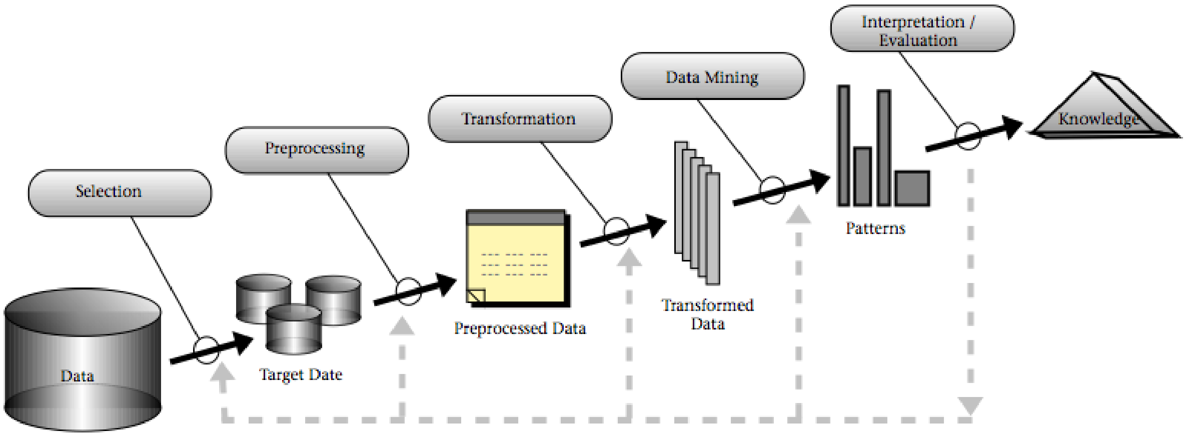


Figure 6. An overview of the steps in the KDD Process. Source: Fayyad *et al.* (1996).

2.2.1 Data mining algorithms

A data mining model can be either predictive or descriptive in nature (Dunham, 2002). Figure 7 shows some of the common data mining tasks for each model type.

A *predictive model* consists of predictions about data values using known results found from existing data. A *descriptive model,* on the other hand, involves patterns or relationships in data. Unlike predictive models, descriptive models allow to explore the properties of the data examined rather than predicting new properties (Dunham, 2002).

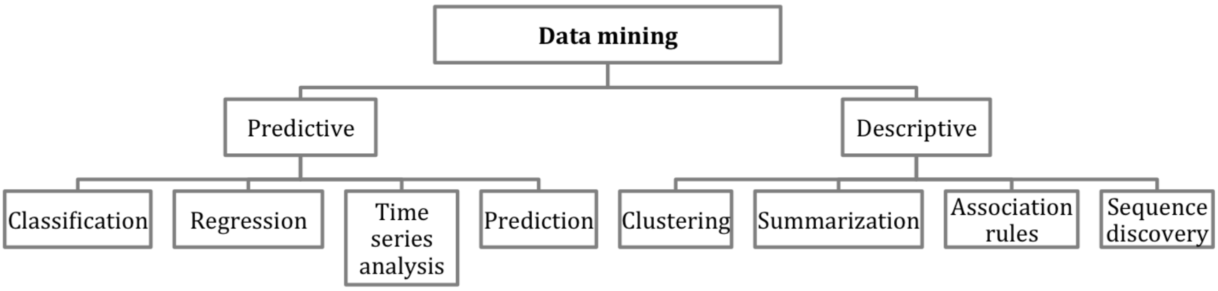


Figure 7. Data mining models and tasks. Adapted from Dunham (2002).

In the predictive models group, classification algorithms (also designated as classifiers) are perhaps the most familiar and most popular data mining technique (Dunham, 2002). This technique assigns a known "class" to unlabelled instances. No one classification technique is always superior to the other in terms of classification accuracy. However, there are advantages and disadvantages to the use of each (Dunham, 2002).

In the case of research on spatial pattern type mapping in deforested areas, the classification task has been limited to decision tree (DT) classifiers (Silva *et al.*, 2011). The wide adoption of DTs can be attributed to their easy interpretation, low data preprocessing requirements (features do not require normalisation or scaling), and computational efficiency. However, there are others robust algorithms that should also be tested due to their high performance when associating complicated information with target attributes without any constraints on the sample distribution. Among these algorithms, Artificial Neural Networks (ANNs) are powerful tools that improve their performance by 'learning,' a process that may continue even after the training set has been analysed. Some disadvantages of ANNs are difficulty in explaining their workings to end users (unlike decision trees), overfitting and failing to converge in the learning phase (they do not guarantee convergence or optimality) (Dunham, 2002).

2.2.1.1 Decision trees

DTs, initially designated as "discrimination nets" by Feigenbaum and Simon (1963), are defined as hierarchical models composed of decision rules that recursively split independent features into homogenous zones (Myles *et al.*, 2004). DTs combine features in a hierarchical manner, the most important being the one located at the root of the tree. Each node in the tree refers to one of the features. Each leaf is assigned to one target class representing the most frequent class value. Additionally, the leaf holds a probability vector that indicates the probability of a target class occurring (Lee and Park, 2013) (Figure 8).

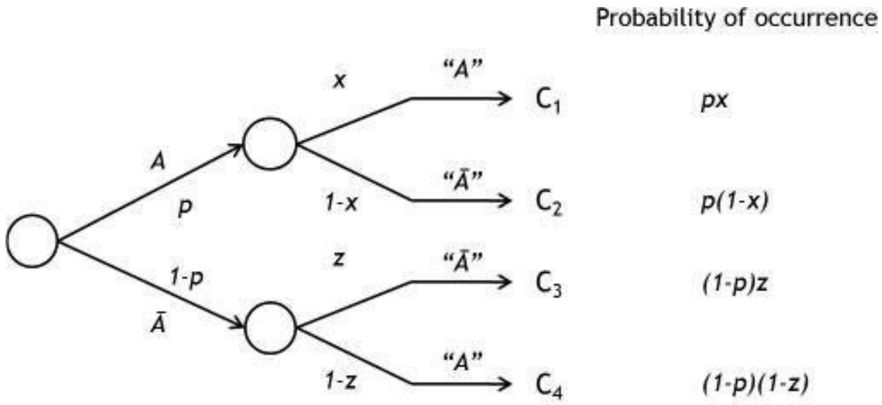


Figure 8. Structure of a decision tree with the probability of occurrence by class. A: Source: Grobbelaar and Visser (2015).

Among DT algorithms, the C4.5 classifier (Quinlan, 1993) is commonly used in GIS (Lee and Park, 2013) and remote sensing (Peña *et al.*, 2014) applications. C4.5 builds a decision tree using the concept of information entropy. The gain ratio is used as an attribute selection measure to build a decision tree. This step removes the information gain bias and thus gives preference to attributes with few values rather than those with many values (Das, 2015, p. 315). In order to avoid overfitting, C4.5 employs post-pruning which allows building smaller tree models that perform better classification accuracy on new data sets as opposed to just the training data set. The parameter that defines the level of pruning is called confidence level (CF). The default confidence limits used by C4.5 is 25% (Cho and Kurup, 2011).

2.2.1.2 Artificial Neural Networks

Artificial Neural Networks (ANNs), initially termed "nervous activity" by McCulloch and Pitt (1943), have gained attention in ecological and remote sensing as a powerful, flexible, statistical modelling tool for uncovering patterns in data (Olden and Jackson, 2002).

There are many types of ANNs that differ in basic architecture (see Bishop, 1995; Ripley, 1996). The feed-forward network (FNN) with the backpropagation algorithm is the most common ANN architecture. FNN's network structure allows information to flow in only one direction, from input to output. Among different FNNs, Multilayer

Perceptron (MLP) is widely used. In short, a FF-MLP consists of an input layer, one or more hidden layer(s) and an output layer (Figure 9). During the training process, multiple weights are assigned to links connecting nodes across various layers. The overall aim is to find weights that minimize a cost function (usually the error) between real observations and predictions. This aim is embedded in the backpropagation algorithm, which computes the cost function on all the training pairs. The weights are then adjusted iteratively to fit the desired output.

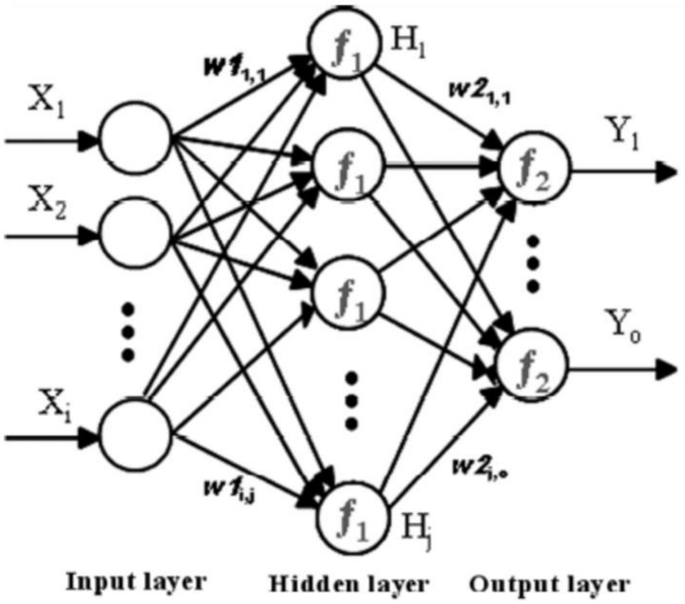


Figure 9. A typical FF-MLP neural network with a single hidden layer architecture with *i*, *j*, and *o* neurons in the input, hidden, and output layers, respectively. f_i stands for the activation function. *w* represents the weights. X_i stands for the input variables. Y_i represents the output variables. Source: Bianconi *et al.* (2010).

In mathematical terms, a single hidden-layer MLP artificial neural network can be represented as:

$$y = \phi(\alpha + \sum_{h=1}^{s} w_h \phi(\alpha_h + \sum_{i=1}^{p} w_{i,h} x_i)) \quad (1)$$

where x is the input, $y$is the output,$p$the number of variables, s is the number of neuron units in a single hidden layer,ϕ is a linear or logistic function, $w_{i,h}$and α_h are the weights of the link between the input layer and the hidden layer, and $w_h$and α are the weights of the link between the hidden layer and the output layers (Yan, 2008).

3

Methodology

3.1 Study Area

The target area was the Amazon region delineated by RAISG (2012) (Figure 10). From this boundary, the humid forest ecoregion delineated by Olson *et al*. (2001) was selected due to a high consistency of the deforestation datasets used and availability of previous studies documenting various spatial configurations and composition of deforested areas, mostly on the Brazilian side.

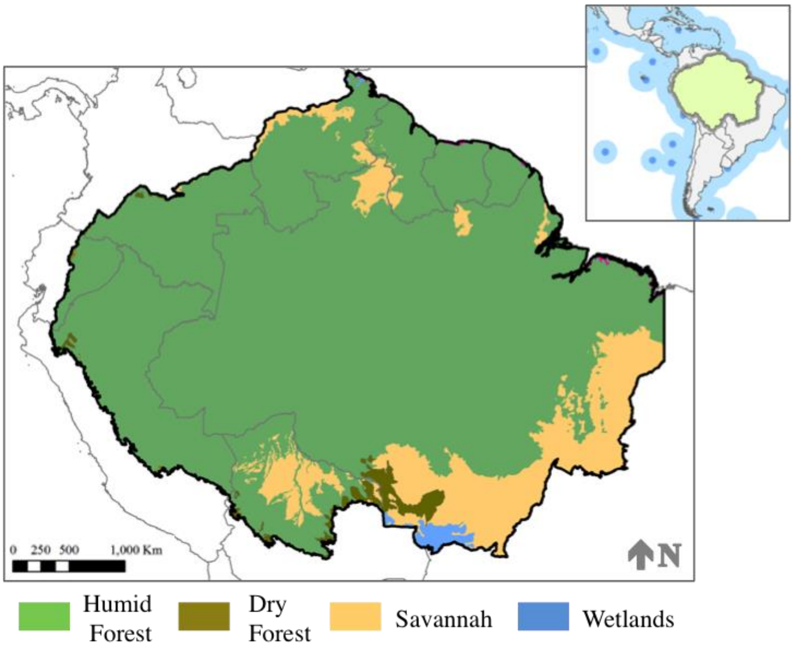


Figure 10. Amazon region delineated by RAISG (2012) including the WWF (2009) Major Habitat Types.

3.2 Deforestation datasets

Two multi-temporal deforestation datasets, Terra-i (Reymondin *et al.,* 2012) and Global Forest Change (GFC) (Hansen *et al.*, 2013), with different spatial resolutions (250m and 30m, respectively) were used. Both datasets have supported the exploration and identification of temporal and spatial trends of deforestation at large geographical extents (continental and global, respectively). Further details about their main features, including the version acquired for this study, are summarised in Table I.

For purposes of this study, a binary forest/non-forest map was created for each dataset. Each map assumed that cumulative detections from 2004 to 2013 were non-forest. This adjustment to a common period between datasets improved comparisons and consistency for *a posteriori* analyses performed from the forest/non-forest maps.

Table I. Features and versions for the two deforestation datasets (monitoring systems) used in this research.

Feature	*Terra-i*	*Global Forest Change (GFC)*
Spatial resolution (in degrees and nominal value in metres)	0.00208333 degrees 250m	0.00025 degrees 30m
Temporal resolution	16 days	Annual
Satellite(s)	MODIS (MOD13Q1) and TRMM (3b42 v7)	Landsat (Circa Landsat 7 mosaics)
Modelling approach	Bayesian neural networks to predict vegetation changes based on (a) vegetation index (MODIS) and (b) precipitation data (TRMM)	Landsat time-series and algorithm (bagged CART). Changes based on image interpretation on-screen training data (QuickBird and 2000 Tree cover from Landsat and MODIS)
Model steps	TRAINING: 2000-2003 (using MODIS vegetation clustering) MODELING: 2004-Present	TRAINING: 2000 (using %tree cover threshold <25% not forest) MODELING: 2001-2013
Version used	Version 2004 to Feb 2015	Version 1.1 (2001-2013)

The term “deforestation” in these datasets, indicated here by *non-forest,* represents replacement of tree cover (planted or natural) by anthropogenic factors (e.g. cultivated pastures, agricultural fields, urban areas and/or human settlements) or by natural events (e.g. flooding, fire). The clarification is relevant as neither the Terra-i nor GFC methodologies is yet able to fully discriminate human disturbances in natural forest cover. Further implications of these definitions are given by TNC (2015).

3.3 Data preparation

Datasets were downloaded and masked with the target area and cumulative 2004-2013 forest/non-forest maps were generated. Both Terra-i and GFC forest/non-forest maps were reprojected from their native geographical projection (WGS84) to the Interrupted Goode Homolosine (IGH) projection using nearest-neighbor resampling. The cell sizes assigned in the reprojection for the former and latter datasets were 240m and 30m, respectively. These reprojecting values, which allowed 1 Terra-i pixel to be equated with 8 GFC pixels and aligned the GFC reprojected pixels based on the Terra-i reprojected pixels, ensured consistent computation of the fractal dimension and extraction FRAGSTAT-like metrics according to the fractal analysis and mining land-use patterns-based methods described in the next sections.

The IGH projection was selected due to its minimized distortion and equal-area characteristics, which lend to its visual appeal and make it a good choice for analysis purposes (Steinwand, 1994). Additionally, this projection has proven feasible at regional levels as demonstrated by previous studies in Latin America (Clark *et al.*, 2012, Clark *et al.,* 2010, Sanchez *et al.*, 2012).

3.4. Spatial deforestation pattern mapping methods

3.4.1 Fractal analysis approach

The box-counting approach with the bottom-up method using fixed scan grids, proposed by Sun *et al.* (2013) and Sun *et al.* (2014), was implemented to compute the fractal dimension from Terra-i and GFC forest/non-forest projected maps. In practice, a series of grids was recursively generated within an AOI box which had been laid over non-forest areas. In each iteration k, the minimum N_k squares of side $\epsilon_k = 2^k$ (pixels, in this case 1 pixel = 30m or 240m depending on the dataset) needed to encompass the deforested areas were recorded, where $k = [0, \cdots, m]$. After defining an optimal value of maximum iterations m, that maintained stable results (Encarnação *et al.*, 2012 cited by Sun *et al.* (2014), the fractal dimension D was estimated by linear regression:

$$\log N_k = -D \log \epsilon_K + c \qquad (2)$$

where c is a constant. The bottom-up method with the fixed scan strategy, used to represent fractal structures over large and heterogeneous landscapes, pixelates the entire landscape and then calculates the fractal dimension of each pixel. In this method, it is important to designate the pixel size (the length of one side of the pixel) before computing D because the box-counting method relies on dividing pixels recursively, which can lead to inexact divisions (Sun *et al.*, 2014).

With the aim of determining the effect of various grid sizes (pixels) on fractal dimension computation, four increasing matrices of 15360 m × 15360 m, 30720 m × 30720 m, 61440 m x 61440 m and 122880 m x 122880 m non-overlapping fixed grids were superimposed on the reprojected Terra-i and GFC forest/non forest maps. These sizes were determined based on considerations of Landsat and MODIS image projection, which have minimum pixel sizes of 30 m and 240 m, respectively. Defining $m=5$ as the optimal value for interactions k, the finest grid, which is given after recursively dividing the maps ($\epsilon_k = 2^k$ or in this case 2^5 = 32) to calculate the fractal dimension, has more than 1 pixel (Table II). For instance, a grid size of 7680 m, which is equivalent to 32 and 256 Terra-i and GFC pixels, can be exactly divided by $\epsilon_k = 2^k$, using $m=5$ as the maximum value possible for perfoming the box counting process.

Table II. List of grid sizes assessed and number of pixels analysed by dataset at the finest grid, with *m = 5* to compute the fractal dimension using the box counting technique.

Grid size (m)	*Numbers of pixels (length) by grid by dataset*		*Numbers of pixels analysed at the finest grid (m = 5 or 2^5 = 32) by dataset*	
	Terra-i (240m)	*GFC (30m)*	*Terra-i (240m)*	*GFC (30m)*
15360	64	512	2	16
30720	128	1024	4	32
61440	256	2048	8	64
122880	512	4096	16	128

3.4.2 The mining land-use patterns-based approach

Based on earlier work by the Brazilian National Institute for Space Research (Silva *et al.* 2011) and derived researches mainly by Saito (2011) and Gavak (2011), the mining land-use patterns-based approach was implemented to map deforestation patterns in the study area. Figure 11 illustrates the workflow of this methodology adapted for this research under a KDD process scheme.

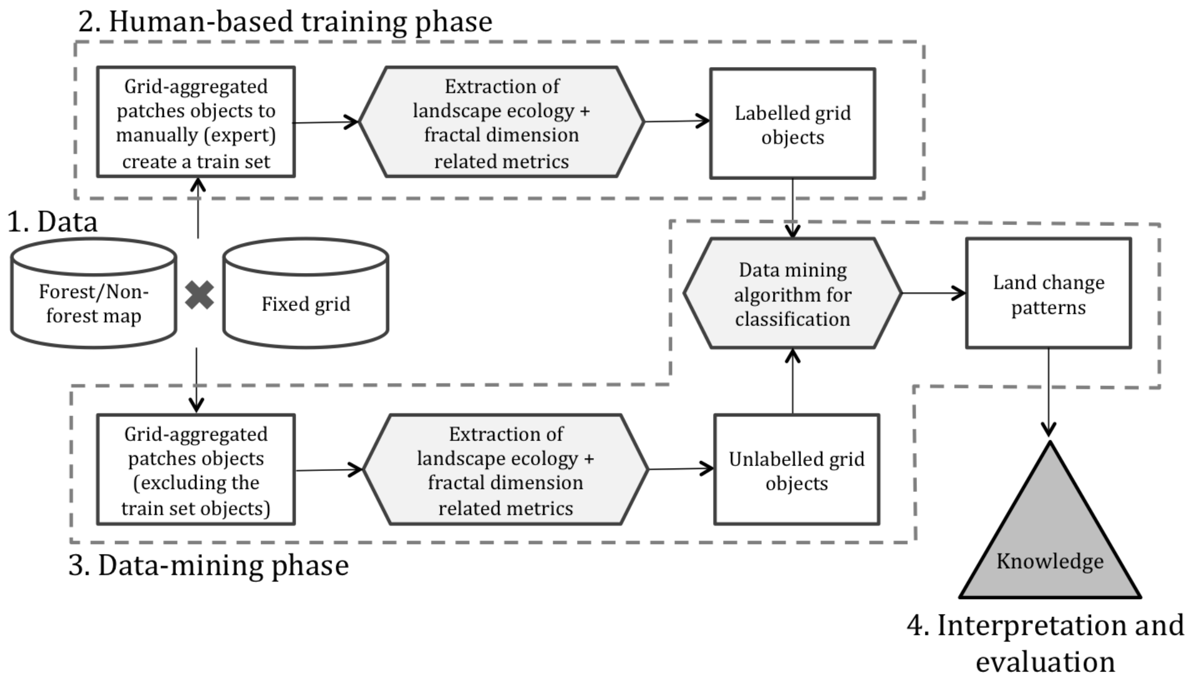

2. Human-based training phase
Grid-aggregated patches objects to manually (expert) create a train set
Extraction of landscape ecology + fractal dimension related metrics
Labelled grid objects
1. Data
Forest/Non-forest map
Fixed grid
Data mining algorithm for classification
Land change patterns
Grid-aggregated patches objects (excluding the train set objects)
Extraction of landscape ecology + fractal dimension related metrics
Unlabelled grid objects
3. Data-mining phase
Knowledge
4. Interpretation and evaluation

Figure 11. Workflow of the mining land-use patterns-based method using fixed grids strategy implemented in this research. Adapted from Silva *et al*. (2011).

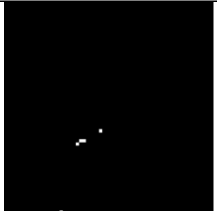
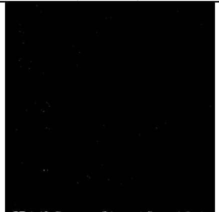
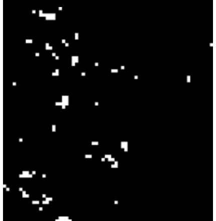
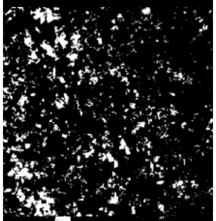
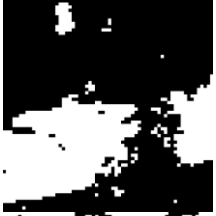
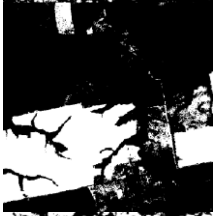
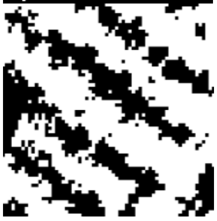
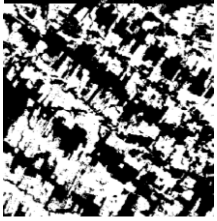
3.4.2.1 Data

The input data consisted of forest/non-forest maps derived from preprocessed (reprojected) deforestation datasets. Using the grid-aggregated patches approach, the four increasing, non-overlapping fixed grid sizes in which the fractal dimensions were computed were also superimposed on this analysis. As a result, each database containing grid-based forest/non-forest objects was split into two subsets by dataset and grid size. One subset was used for a supervised learning, where a set of representative samples was collected from common pattern spatial typologies identified across all fixed grid sizes. The remaining grid-based objects were used as target data to classify automatically using the data mining models trained by dataset and grid size.

3.4.2.2 Human-based training phase

Four spatial pattern types were initially defined and associated with potential land-use activities across all fixed grid sizes assessed from preprocessed Terra-i and GFC forest/non-forest maps (Table III). The selection of these deforestation spatial pattern types was supported with previous mining pattern analyses that used the fixed grid scan strategy on a deforestation dataset derived from Landsat remote-sensing data (30m) in the Brazilian Amazon (Gavak, 2011; Saito, 2011; Saito *et al.*, 2011).

Table III. Description of the deforestation spatial pattern typologies using the minor fixed grid size (15360 m x 15360 m) assessed in the study area. Appendix 2 contains the visualisation for the remaining grid sizes not illustrated in this table.

Pattern	*Visualization using the 15360 m x 15360 m grid size by dataset (aggregated 2004-2013 forest/non-forest map)* ■ **Forest** □ **Non-Forest**		*Description (scale 1:75,000)*	*Land-use agents associated / Spatial distribution / Accessibility (transport infrastructure)*
	Terra-i (240 m)	*GFC (30 m)*		
Diffuse extensive			Small-scale clearings	Smallholder subsistence agriculture / Dispersed or scatter distribution / Low accessibility
Diffuse extensive			Small to medium scale (irregular or geometric) clearings	Roadside or riverside colonization by spontaneous migrants / Clustered distribution / Low to medium accessibility
Geometric			Large-scale (geometric) clearings	Modern and industrial sector activities / Clustered distribution / Medium to high accessibility
Multi-directional			Corridor-like clearings (irregular or geometric) perpendicular to a main corridor clearing	Planned resettlement schemes (mostly in the Brazilian Amazon) / Clustered and/or scatter distribution / High accessibility

After identifying and describing common pattern typologies across all grid sizes, a strategy for collecting representative samples by pattern was performed in two steps. First, a set of fifteen representative grid-based objects organised by pattern (60 samples in total for the four patterns) was manually collected using the largest grid size (122880 m). A key criteria for selecting this initial set of samples was visual agreement in representation of the pattern between Terra-i and GFC forest/non-forest maps. Second, samples were collected for the remaining fixed grid sizes using a stratified sampling strategy. A sample of the target grid size was selected from the sample site in a major grid size (Figure 12).

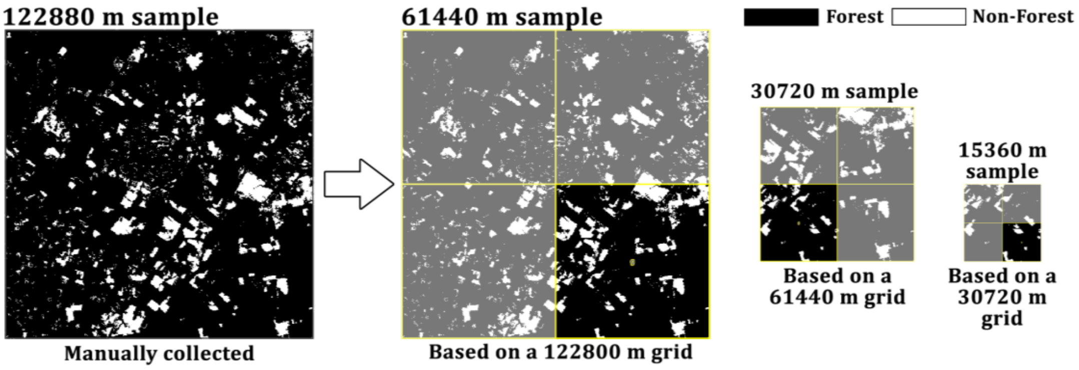


Figure 12. Example from the Terra-i dataset of the sampling strategy. A labelled train set of the geometric pattern was generated and later used in the data-mining phase.

With the set of samples collected across all pattern typologies, grid sizes and datasets, a group of thirteen FRAGSTAT-like metrics at class level were extracted using the grid-based object as the unit of analysis and non-forest as the target class. These metrics were selected on the basis of results of reference studies by Gavak (2011) and Saito, (2011) who discriminated multiple deforestation pattern typologies with the grid-based approach and data mining techniques in the Brazilian Amazon. To complement these metrics, the fractal dimension (FD) determined in the fractal analysis approach was added to the FRAGSTAT-like metrics database according to the grid size and dataset. From this metric, a new variable was derived and added to each database under the assumption of a close relationship between FD and proportion of areal measures. The resulting variable was the ratio between the fractal dimension and the FRAGSTAT-like metric related with the percent of land occupied by the non-forest class.

Appendix 3A contains a basic description and definition (for the purposes of this research) of the set of FRAGSTAT and fractal-like metrics selected and extracted from grid-based objects according to grid size and dataset. Appendix 3B describes further details about their formulas and ranges.

3.4.2.3 Data mining phase

There were two types of datasets used in this phase. One, the “gold-standard” (human-labelled) dataset, prepared by grid size and dataset, was used to build data mining algorithms models. Another set, consisting of the remaining unlabelled grid-based objects, was then automatically classified within the four spatial pattern typologies (described in Table III) using the built models with the best performance (Kappa value) from the gold-standard set.

Regarding the data mining algorithms, the feasibility of C4.5 decision trees and feed-forward multilayer perceptron ANNs with back propagation methods was assessed to classify the four spatial pattern typologies identified from Terra-i and GFC forest/non-forest maps at multiple grid sizes. The workflow of the data mining phase was divided into five sub-phases: data preprocessing, model construction (learning), model evaluation (accuracy), model sensitivity analysis and model use (classification). The first three steps were examined for all C4.5 decision trees and artificial neural networks models generated. The remaining steps (model sensitivity analysis and model use) were exclusively performed for the built neural networks models with the best performance according to the Kappa value. The exclusion of the C4.5 decision tree models in those steps was on the basis of using exclusively them as reference to proof or reject the goodness of neural networks for discriminating deforestation patterns.

3.4.2.3.1 Data preprocessing

Preprocessing (normalisation) of input data is a particularly important step for built neural networks models with backpropagation. According to Kim (1998), input and output vectors for backpropagation need to be normalised properly in order to achieve the best performance of the network. The author claimed if the activation function used is the standard sigmoid (which was the case in this research) each input should be normalized between 0 and 1. The formula for this normalization is given by Eq. (3)

$$\tilde{a}_{pi} = \frac{a_{pi} - a_{pi\min}}{a_{p\max} - a_{p\min}} \qquad (3)$$

where $a_{p\max} = \max(a_{pi}; i = 1, \cdots, m)$, $a_{p\min} = \min(a_{pi}; i = 1, \cdots, m)$ and $p(p = 1, \cdots, P)$ are the input patterns. In Eq. (3), $\tilde{a}_{pi}$ denotes the normalized value of the unit i of input vector $a_p = (a_{p1}, a_{p2}, a_{p3}, \cdots, a_{pm})$, and a_{pi} denotes the original value of the input i in the pattern p (Kim, 1998).

3.4.2.3.2 Model construction

Several architectures can be derived from the two data mining algorithms tested. As part of the algorithm feasibility assessment, the C4.5 structure used by previous pattern mapping methodologies (Silva *et al.*, 2011) was confronted with FF-MLP ANN architectures according to training set, grid size and deforestation dataset.

In the case of the C4.5 decision tree, the traditional settings given by previous pattern mapping used a default confidence level (CF) of 25% for post-pruning. The smaller the confidence limit, the higher the chances of pruning and vice versa (Cho and Kurup, 2011).

Due to the lack of studies implementing neural networks for mining deforestation spatial patterns classification, this research implemented the grid-search procedure with *k*-fold cross validation for selecting the candidate parameters in the neural networks (Castro *et al.*, 2013, p. 239). The weight decay and hidden nodes number were the parameters tuned according to the type of neural network implemented. Based on the suggestion by Ripley (1996, p. 162), the weight decay when a small number of inputs is scaled to range from 0 to 1 is usually set between 0.001 and 0.1. A typical number of hidden nodes in a single hidden layer can range from 5 to 100 (Yan, 2008). For this study, the number of nodes in a single hidden layer was changed from 1 to 50 by 1. Five values for weight decay were used as a basis for exploration: 0.00001, 0.0001, 0.001, 0.01 and 0.1. As result, there were 250 different parameter combinations of the values for weight decay and the number of hidden-layer nodes.

3.4.2.3.3 Model evaluation

The performance of each model (also denoted as surrogate model in this step) was evaluated using a five-fold cross validation with 40 iterations, similarly to Beleites and Salzer (2008) who also used this setting for a small training set size with the presence of extreme values. Besides informing the surrogate models' performance for C4.5-based and neural networks-based algorithms applied on different training sets, the iterated *k*-fold cross-validation technique allowed exploration of the stability of each surrogate model and identification of optimal parameter combinations for the neural networks algorithm.

In the cross-validation technique, *k* indicates the number of equal subsets that split a full training set. For each run, the training process is run on *k*-1 subsets and the validation is done on the remaining subset. An iteration refers to a permutation of k-1 subsets of the k subsets, followed by a repetition of the same validation scheme. For instance, under a five-fold classification with 40 iterations, the training set was split into 5 parts, 80% training and 20% test, repeated 40 times.

Both the comparison of data mining algorithms and selection of the best parameter combinations for the neural network models assessed was performed using the Kappa (κ) statistical metric as the error measure. This metric was selected based on its feasibility for categorical data (Williamson *et al.,* 2000) and also due to its previous use in the reference patterns mapping studies (Gavak, 2011; Saito, 2011) that used the C4.5 algorithm. A value of 0 for κ indicates disagreement unlikely to be due to chance, and a value of 1 indicates perfect agreement (Williamson *et al.*, 2000). In addition to Kappa, the overall accuracy, which is usually considered as an overestimation as it does not account for agreements that would have occurred by chance, was explored as a complementary reference measure for the correctness of classification. The use of both metrics together has been reported in the machine learning literature for ecological and land cover applications (Lippitt *et al*., 2008).

3.4.2.3.4 Model sensitivity analysis

The sensitivity analysis is a key procedure in model development to detect input-output dependencies. In the case of artificial neural networks, the generalisation of ANNs as measured by an error function (the Kappa metric in this research) depends upon the ratio of the number of training data to the number of ANN parameters, and the ANN parameters depend upon the number of input variables (Yan *et al*., 2012, p. 255). It is important to note that as ANNs are a data-driven approach rather than a statistical approach, the important inputs are selected based on the performances of ANN models or by sensitivity analysis using several techniques such as tested by Olden *et al.* (2004). In Olden *et al.* (2004), the connection weights method was highlighted as the best methodology for accurately quantifying importance of variables.

The connection weights algorithm, originally proposed by Olden and Jackson (2002), calculates the sum of products of final weights of the connections from input neurons to hidden neurons with the connections from hidden neurons to output neuron(s) for all input neurons. The relative importance of a given input variable can be defined as (Olden *et al.,* 2004):

$$RI_x = \sum_{y=1}^{m} \omega_{xy}\omega_{yz} \qquad (4)$$

where RI_x is the relative importance of input neuron x, $\sum_{y=1}^{m} \omega_{xy}\omega_{yz}$ is sum of the product of final weights of the connection from input neuron to hidden neurons with the connection from hidden neurons to output neuron, y is the total number of hidden neurons, and z is output neuron(s). This approach is based on estimates of final network weights obtained by training the network. One of the added advantages of the connections weights method is that it distinguishes if the input has a positive or negative effect on each target output (Yan *et al.*, 2012).

3.4.2.3.5 Model use

The final step in the data mining phase used the best fitted ANN models to automatically classify unlabelled sets distributed in the grid sizes and deforestation datasets assessed. Each unlabelled observation contained the same normalised features (FRAGSTAT and fractal-like metrics) used in the training set. The total number of grid-based objects classified was reported for the whole target study area. Additionally, in order to approximate to the true performance of the best model among the grid sizes assessed for each dataset, a basic balanced sampling scheme was performed that consisted of 30 samples (double the amount used for training) by pattern from the classified unlabelled dataset. This amount was empirically selected and supported by the reference mining pattern study by Saito (2011), which validated the C4.5 models using a random unbalanced sample of 90 grid-based objects.

Finally, following the procedures of reference studies by Gavak (2011) and by Saito (2011) who also used the mining land-use patterns-based approach, in this research was uniquely present and visualise the fitted model(s) with the grid size holding the best Kappa value.

3.5 Software and implementation

Quantum GIS v.2.1.4 was primarily used for vector and raster operations (masking datasets with the target region and creating the forest/non-forest maps) as and to produce cartographical outputs. This software also assisted in the visual selection of the target pattern typologies, the samples of which were then used to create the training sets by grid size for each dataset. The grid sizes vectors (also denoted here as fishnets) were created using the Geospatial Modelling Environment tool v.0.7.2 (Beyer, 2015).

R software v.3.2 was used for extracting the FRAGSTAT and fractal-like metrics from labelled and unlabelled grid-based objects across all grid sizes and datasets. FRAGSTAT-like metrics were selected using FRAGSTAT v.4.2 (McGarigal, 2015), called within an R code. Part of the FRAGSTAT settings empirically configure an edge depth value equal to 0 m to compute edge-related metrics. For the FD, this value was extracted using the raw programming codes provided by Sun *et al.* (2014). After merging both FRAGSTAT and fractal-like metric types by grid size for each deforestation dataset, an exploratory analysis using box-plots was performed over the normalised metrics for each training set. This analysis allowed visual inspection of each variable's behaviour by grid size by deforestation dataset.

Data mining projects were run using the caret v.6.0-52 R package (Kunh, 2015), which contains multiple preprocessing functions (normalisation), algorithms (e.g. C4.5 DTs and FF-ANNs), model evaluation techniques (grid-search procedure with iterated *k*-fold cross validation) and tools and statistics (overall accuracy and kappa metrics) for visualising and checking the models' behaviour. For the ANN models, the default settings for certain model parameters by the caret R-package were maintained for range (0.7) and modified for maximum number of iterations (5000) and maximum allowable number of weights (2000). The latter modifications were based on preliminary tests over the most complex ANN architectures (hidden-layer node numbers = 50), which required the modified values in order to converge. The sensitivity analyses were performed independently using the NeuralNetTools v.1.3.1. R-package (Beck, 2015), which performed the connection weight method.

Finally, R software was also used to export the classification results by fishnet for each dataset. The results were then used to visually assess the spatial distribution of pattern typologies using QGIS. All cartographical outputs of this research maintained the IGH projection. The main R programming codes can be accessed in Appendix 4. Parallel processing using the doParallel v.1.0.8 R package (Weston, 2015) was implemented for efficient processing of the large databases created and/or manipulated in this study. Additionally, to produce reproducible results, the set.seed function in R was added to all code sets. This function guarantees that the random numbers generated in each code are the same for each run.

4

Results

4.1 Fractal analysis approach

Figure 13 illustrates the percent contribution of each deforestation stage out of all grid-based objects analysed, arranged by grid size for each dataset. An effect of the extent of the unit of analysis (grid-based object) on the presence or absence of the five deforestation stages was observed. For instance, the presence of the type 4 (rapid growth and metastatic consolidation) and type 5 (clearing consolidation) deforestation stages was only apparent in the grid-based objects at 15360 m and 30720 m grid sizes. The Terra-i dataset had a higher proportion of type 5 deforestation stage than the GFC dataset. In terms of fluctuations, the type 1 (no or trivial clearing areas) and type 2 (dispersed clearings areas) stages presented the largest changes by grid size, a trend that was especially evident in the GFC forest/non-forest maps. The maps that illustrate the spatial distribution of the five deforestation stages by grid size for each dataset are presented in Figure 14.

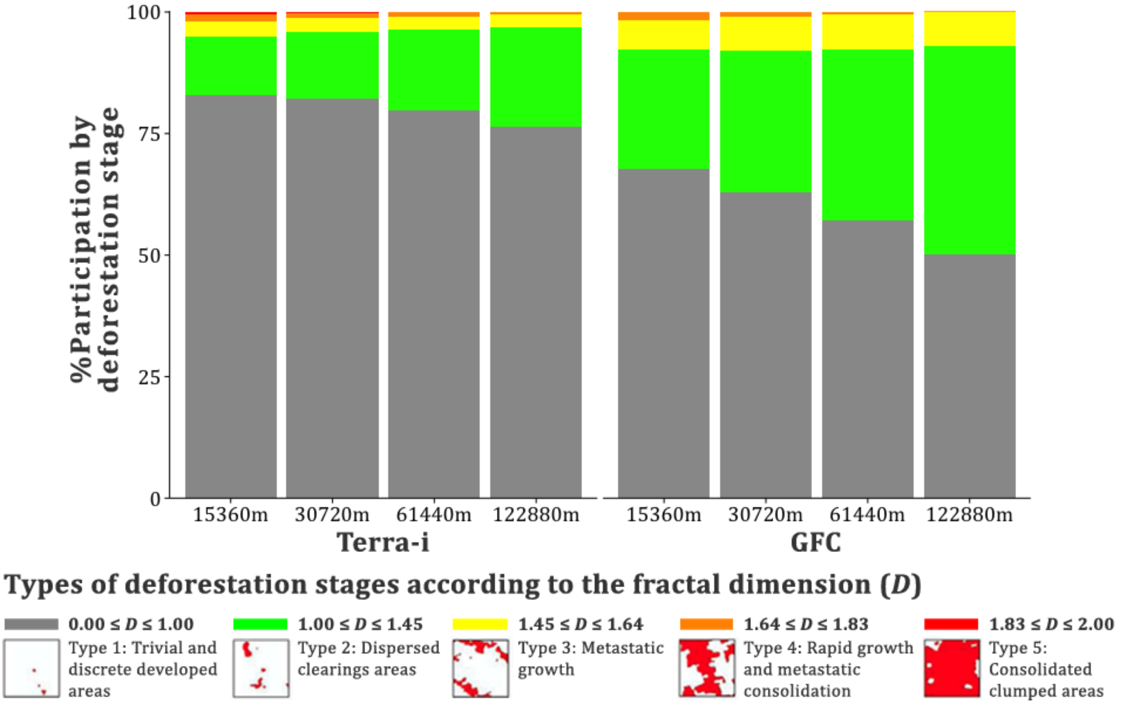


Figure 13. Contribution of the deforestation stages proposed by Sun *et al*. (2014) to all grid-based objects, aggregated by four grid sizes and two deforestation datasets.

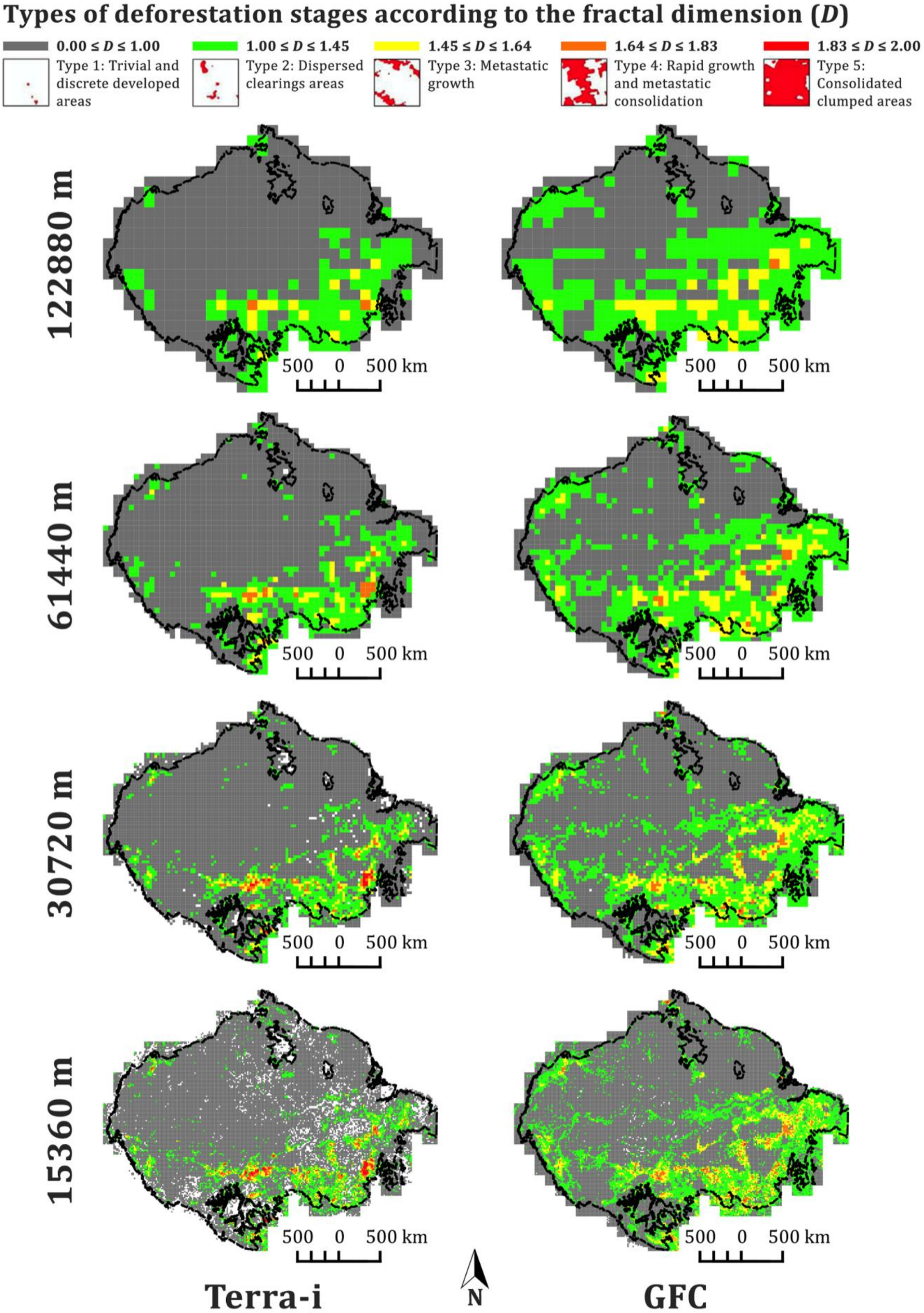

Types of deforestation stages according to the fractal dimension (D)
$0.00 \leq D \leq 1.00$
Type 1: Trivial and discrete developed areas
$1.00 \leq D \leq 1.45$
Type 2: Dispersed clearings areas
$1.45 \leq D \leq 1.64$
Type 3: Metastatic growth
$1.64 \leq D \leq 1.83$
Type 4: Rapid growth and metastatic consolidation
$1.83 \leq D \leq 2.00$
Type 5: Consolidated clumped areas
122880 m
61440 m
30720 m
15360 m
500 0 500 km
N
Terra-i
GFC

Figure 14. Types of deforestation stages proposed by Sun *et al.* (2014) implemented in the study area over four grid sizes (each row) and two deforestation datasets (Terra-i, left column; GFC, right column).

Table IV summarises all overlapping grid-based objects analysed by grid size for each deforestation stage derived from the Terra-i and GFC forest/non-forest maps. It also details the percent of grid-based objects with and without spatial agreement with regard to deforestation stage between datasets. The type 1 stage represented the largest number of grid-based objects with spatial agreement between the Terra-i and GFC deforestation stages maps in all grid sizes assessed. In contrast, almost a third of the overlapping objects disagreed in deforestation stage type, mostly for the type 1 and type 2 categories of both datasets (not shown in Table IV).

Table IV. Total number and distribution of grid-based objects with and without spatial agreement on deforestation stages, by dataset and grid size. Deforestation stage types without any spatial agreement are denoted by a hyphen (-).

Grid size (m)	*Distribution (%) of grid-objects with and without spatial agreement by deforestation type*						*Total overlapping grid-based objects*
	Type 1	*Type 2*	*Type 3*	*Type 4*	*Type 5*	*Disagreement*	
15360	62.46	7.19	1.78	0.68	0.02	27.87	28427
30720	60.70	8.38	1.92	0.39	-	28.62	7984
61440	56.02	11.47	1.89	0.24	-	30.38	2067
122880	49.62	15.41	2.07	-	-	32.89	532

4.2 The mining land-use patterns-based approach

4.2.1 Exploratory analysis of inputs variables in train sets

A series of box plot charts arranged by conceptual category (Neel *et al.*, 2004) allows for approximation of the behaviour of input variables (metrics) in training sets by pattern typology for each grid size and dataset. The main results of this analysis can be accessed in the Appendix 5. Overall, all metrics presented different behaviours, being identified those with potential for discriminating the target patterns typologies.

4.2.2 Model evaluation

Four hundred and eight models (400 ANNs and 8 C4.5 models) were evaluated. The large dominance of ANN models, due to the hyperparameter optimization procedure, allowed finding the optimal parameter combinations (number of hidden nodes and weights decay) by grid size and dataset (Table V).

Table V. Best parameter combinations (number of hidden nodes and decay) for ANN models that had the highest Kappa value in each grid size by dataset. The total number of weights and convergence values are also added as complementary features for each best parameter combination.

Grid size	*ANN parameters*				*ANN features*			
	Hidden nodes		*Decay*		*Weights*		*Convergence*	
	Terra-i	*GFC*	*Terra-i*	*GFC*	*Terra-i*	*GFC*	*Terra-i*	*GFC*
15360 m	3	2	0.01	0.1	64	44	6.40	35.15
30720 m	48	3	0.1	0.01	964	64	26.85	9.89
61440 m	3	3	0.01	0.01	64	64	9.78	12.37
122880 m	48	4	0.1	0.01	964	84	34.75	17.14

Figure 15 illustrates boxplots with the distribution of Kappa values determined by iterated *k*-fold cross-validation for each best ANN parameter combination model confronted with the C4.5 reference model at each grid size by dataset.

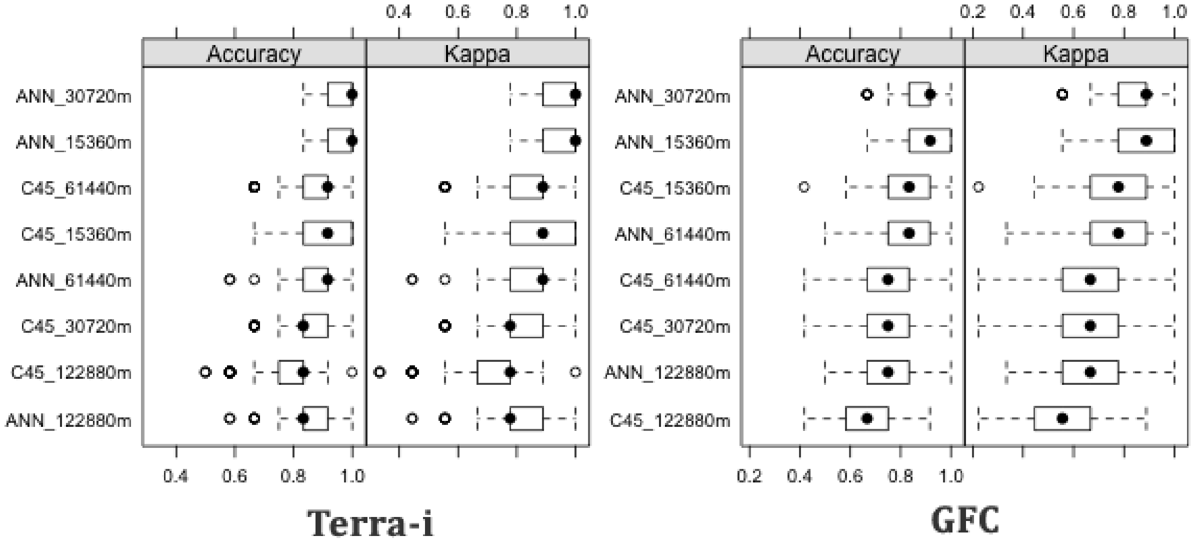


Figure 15. Distribution of overall accuracy and Kappa values from best ANN parameter optimization models and reference C4.5 models by grid size separated by dataset (left: Terra-i, right: GFC). Models are ranked from the highest (upper) to the lowest (bottom) median (black dot) metrics values.

There was agreement between both Terra-i and GFC datasets to rank ANN models of grid size 30,720 m with the highest Kappa median values (1 and 0.85, respectively). Overall, compared to reference C4.5 models, ANNs performed better at discriminating the pattern typologies except for Terra-i's 64,440 m and 122,880 m grid sizes. Comparing models' Kappa median values distribution by dataset, Terra-i models had in general less IQR than GFC, however outliers were dominant for the Terra-i dataset models. Regarding overall accuracy, this metric showed similar behaviour to Kappa except that IQRs were narrower for both Terra-i and GFC datasets. The similar behaviour of both metrics corroborated 30,720 m as the optimal grid size.

For Kappa media values, there were highly significant differences (p < 0.001) between C4.5 and ANN models at each grid size and dataset except for the 61,440 m grid size (p = 0.189) from Terra-i (Table VI). Although median values of ANN models for GFC at 15360 m were higher than the same model type at 30720 m, the latter grid size was selected due to its lesser variation in Kappa median values (shorter IQRs in boxplots figures) than the former.

Table VI. Statistical tests for differences in mean Kappa values between grid sizes and datasets.

Grid size	*Terra-i (mean value and p-value)*			*GFC (mean value and p-value)*		
	C4.5	*Best ANN*	*p-value*	*C4.5*	*Best ANN*	*p-value*
15360 m	0.89	0.95	1.23E-13	0.79	0.89	8.12E-20
30720 m	0.82	0.95	3.89E-30	0.68	0.85	3.64E-41
61440 m	0.84	0.85	0.189	0.70	0.78	9.77E-15
122880 m	0.71	0.81	4.93E-18	0.56	0.64	1.08E-08

4.2.3 Model sensitivity analysis

Figure 16 illustrates the results of the sensitivity analyses for the best ANN models, in this case those with a grid size of 30720 m for both for Terra-i and GFC datasets. It is important to note that differences in the contribution of RI_x values (y-axis) in each chart below are based on estimates of network final weights by pattern typology obtained by training each ANN by dataset as was previously mentioned in the methodology section.

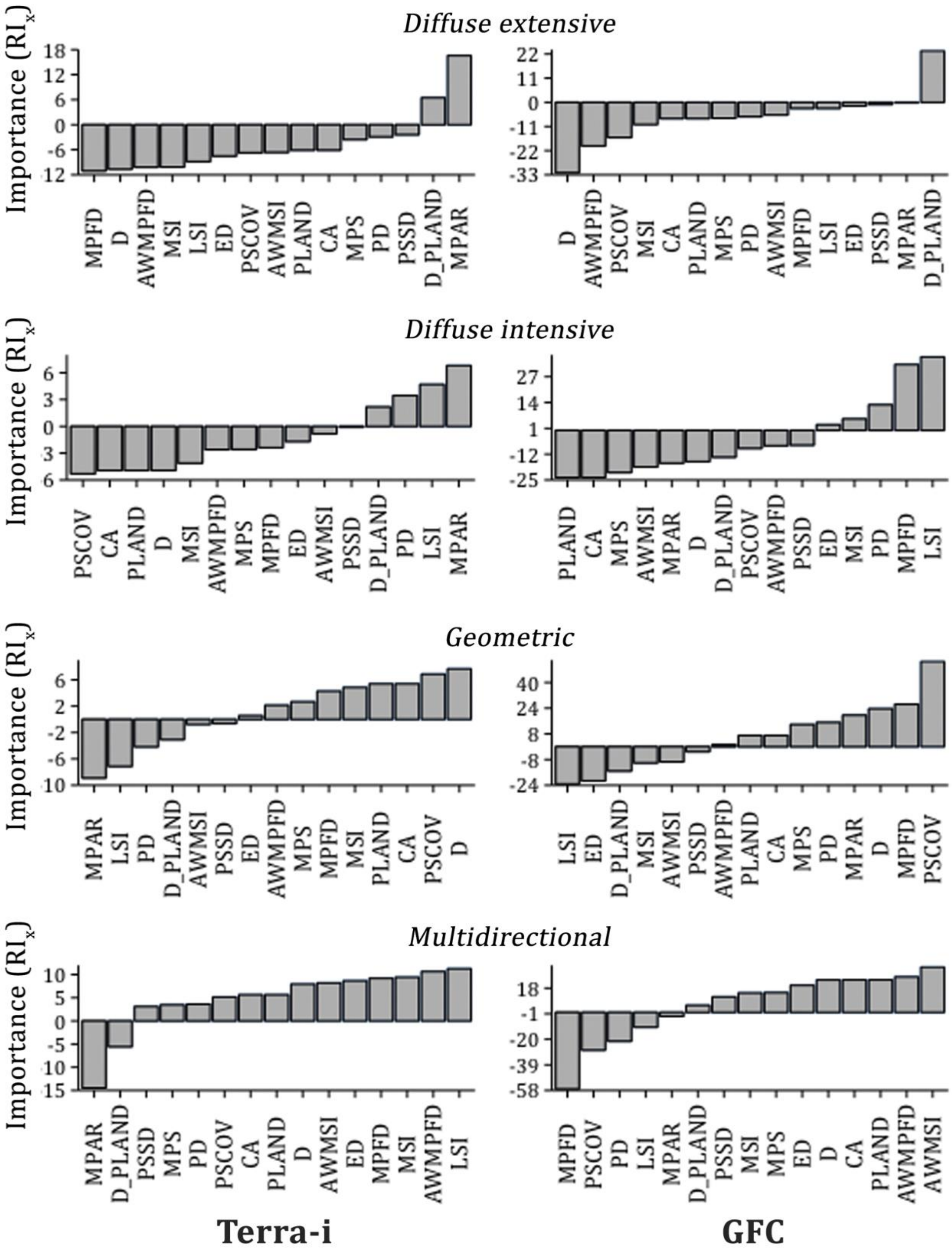


Figure 16. Sensitivity analysis bar plots ranking 15 input variables (normalised FRAGSTAT and fractal-like metrics) for the best ANN models at a grid size of 30720 m by pattern typology (row) and dataset (column).

According to the above charts, the contribution of ANN models inputs, measured with Olden and Jackson's (2002) *RI$_x$* value method, seemed to vary according to pattern typology and dataset. For variation in pattern typology, different variables can contribute positively or negatively in terms of the probability of the presence of each typology. There were only two cases in the Terra-i ANN model where the top ranked variables were the same (MPAR contributing positively for diffuse extensive and intensive patterns and MPAR contributing negatively for geometric and multidirectional patterns). Overall, variables such as PSSD and density-related metrics (edge density and patch density) contributed poorly to the classification of all pattern typologies for both Terra-i and GFC best ANN models. In contrast, MPAR tended to have more marked effect (positive or negative) on the probability of the presence of each pattern typology. Regarding the proposed computed fractal-like metrics (denoted as D and D_PLAND in Figure 16), these seemed to be promissory for deforestation pattern typologies studies due to their relevant contribution to target outputs classification using ANNs models, particularly for the diffuse extensive and geometric patterns.

For variation in dataset type, without considering the contribution rank order, ability to distinguish positive and negative loads from all inputs was essentially the same between both Terra-i and GFC's best ANN models. This variation was accepted in all pattern typologies except for the multidirectional pattern, in which inputs such as PSCOV, PD and LSI had a positive effect in Terra-i's ANN model as opposed to a negative effect for GFC's ANN model.

4.1.4 Model use

Figure 16 and 17 present maps of the spatial distribution of four pattern typologies using the best ANN models for the Terra-i and GFC datasets, respectively. A visual inspection of these maps indicates that the dominant pattern typology in both datasets in the study area belonged to the diffuse-related patterns, extensive and intensive typologies. The former typology predominated the Terra-i forest/non-forest grid-based objects. In contrast, the latter was highly associated with the GFC forest/non-forest grid-based objects. Additionally, it can be observed that a set of grid-based objects classified as diffuse intensive pattern in the Terra-i map (southeast side) were associated with the geometric pattern in the GFC map.

Table VII summarises the total amount of grid-based objects (both unlabelled and training sets) classified under the four deforestation spatial pattern typologies identified using the best ANN models with a grid size of 30720 m created from the Terra-i and GFC forest/non-forest maps. This table also details the percent of grid-based objects with

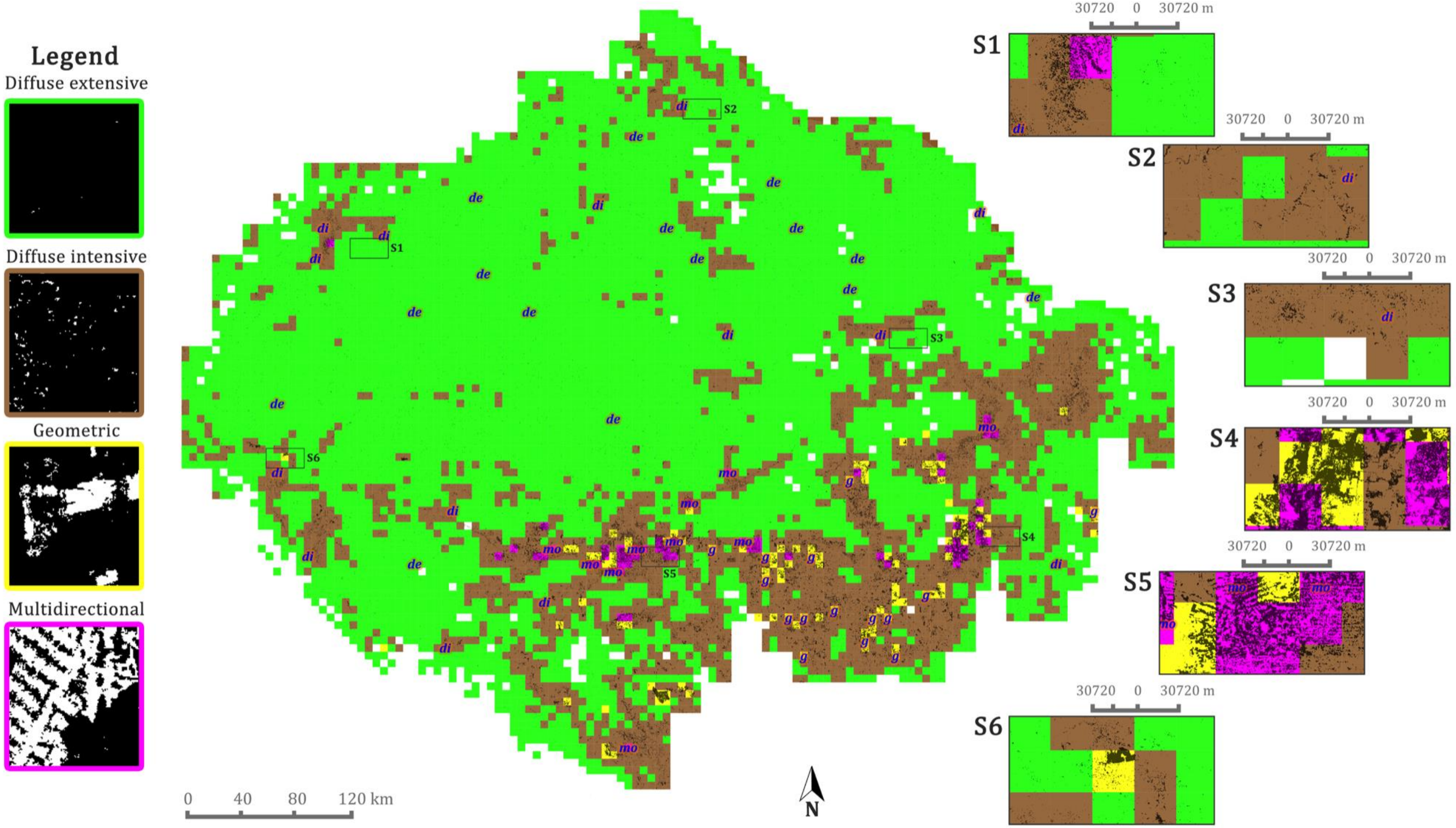


Figure 17. Distribution of deforestation spatial patterns types using the best ANN model at a grid size of 30720 m from forest/non-forest Terra-i maps in the study area (central side). Right side consists of map insets in six random locations. Multi directional, geometric, diffuse intensive and diffuse extensive pattern training sets are denoted using blue italic letters as *mo*, *g*, *di* and *de*, respectively, for all maps.

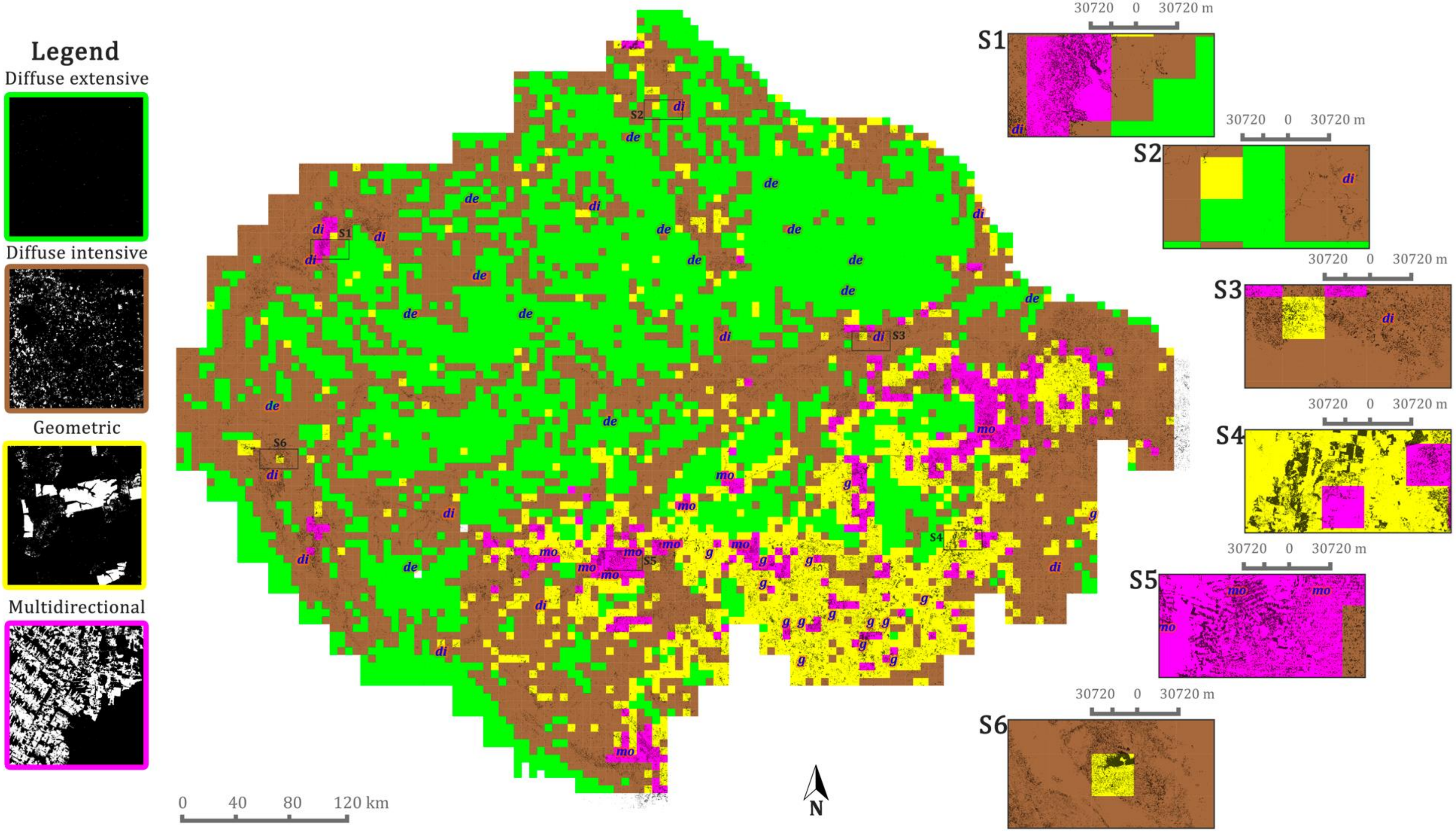


Figure 18. Distribution of deforestation spatial pattern types using the best ANN model at a grid size of 30720 m from forest/non-forest GFC maps in the study area (central side). Right side consists of map insets in six random locations. Multi directional, geometric, diffuse intensive and diffuse extensive pattern training sets are denoted using blue italic letters as *mo*, *g*, *di* and *de*, respectively, for all maps.

and without spatial agreement between deforestation typologies mapped from both datasets. Overall, grid-based objects classified from both Terra-i and GFC datasets agreed that diffuse-related typologies are the most predominant in the study area (97.9% and 81.4%, respectively), followed by the geometric pattern (1.3% and 14.2%, respectively) and multidirectional (0.8% and 4.3%, respectively).

Table VII. Distribution of grid-based objects (includes training sets) and agreement in terms of proportion of four deforestation stages for Terra-i and GFC datasets mapped using the best ANN models (30720 m grid size).

Pattern typology	*Total number (and proportions) of grid-objects classified by dataset*		*Distribution (%) of grid-objects with and without spatial agreement by deforestation type*	
	Terra-i	*GFC*	*In agreement*	*Without agreement*
Diffuse extensive	5840 (73.1%)	3147 (37.8%)	2819 (33.3%)	-
Diffuse intensive	1978 (24.8%)	3629 (43.6%)	805 (9.5%)	-
Geometric	106 (1.3%)	1182 (14.2%)	63 (0.7%)	-
Multidirectional	60 (0.8%)	357 (4.3%)	47 (0.6%)	-
Total	7984	8315	3734 (46.77%)	4249 (53.23%)

Regarding the spatial agreement of mapped typologies between datasets, less than half (46.77%) of total overlapped grid-based objects presented common pattern typologies. The diffuse extensive pattern was the typology with the highest spatial agreement between Terra-i and GFC classified grid-based objects, followed by the diffuse intensive, geometric and multidirectional patterns. Although omitted from Table VII due to the many possible combinations of spatial disagreement, the diffuse extensive vs. diffuse intensive presented in the Terra-i and GFC (or its inverse) was the most common case of disagreement, at 33.2% of all overlapping objects.

A basic assessment of model use over the classified grid-based objects indicated that the GFC ANN model with a grid size of 30720 m had a slightly higher true performance in terms of overall average accuracy (0.800) and Kappa value (0.733) than the Terra-i ANN model, which had average values of 0.772 and 0.696 for the same metrics (Table VIII).

Table VIII. Total number and distribution of grid-based objects with and without spatial agreement on deforestation stages between the two datasets, assessed by grid size. Deforestation stage types with no spatial agreement are denoted by a hyphen (-).

Repetition	*Overall accuracy*		*Kappa value*	
	Terra-i	*GFC*	*Terra-i*	*GFC*
Repetition 1	0.783	0.825	0.711	0.767
Repetition 2	0.775	0.750	0.700	0.667
Repetition 3	0.758	0.825	0.678	0.767
Total	0.772	0.800	0.696	0.733

According to confusion matrices generated from the Terra-i and GFC ANN model evaluations (Appendix 6), confusions between pattern typologies may vary by dataset. Overall, for both dataset models the diffuse extensive typology had the highest chance of being discriminated with regard to the remaining patterns, with some instances of confusion with the diffuse intensive typology. For the diffuse intensive pattern from the GFC model, confusions were equally distributed among the remaining patterns. This result was different for the Terra-i ANN model, in which this pattern was mostly confused with the geometric pattern. The geometric pattern was most often confused with the multi directional pattern in the Terra-i ANN model. In contrast, the multi directional pattern was more often confused with the diffuse intensive pattern in the GFC ANN model. Although the multi directional pattern seemed to be fully discriminated by the Terra-i ANN model (values closer to the 30 samples validated), it was confused with the geometric pattern. This confusion was true for the GFC ANN model as well.

5
Discussion

5.1 Fractal analysis approach

According to Al-Haddam *et al*. (2010), fractal techniques can favour analyses related to the spatial complexity and information content of multi-scale remote sensing data due to the fractal's principle of self-similarity. In this way, multiple spatial and temporal resolutions of remote sensing data could be compared and evaluated based on fractal measurements. In this research, the fractal dimension allowed for comparison of different deforestation stages previously established by Sun *et al.* (2014). Implemented initially for multi-temporal LANDSAT-based forest/non-forest maps, Sun *et al.*'s (2014) bottom-up method using a fixed-grid scan strategy appeared to be also feasible as a means to compute fractal dimension and thus characterise and compare multiple deforestation stages from remote sensing datasets such as Terra-i and GFC. Moreover, the study of FD behaviour under multiple grid sizes and datasets allowed characterizing its sensitivity to both sources.

First, in agreement with previous findings by Sun *et al.* (2014) and related studies (Sun and Southworth, 2013b; Sun *et al.*, 2013), this research found that the number of grid-based objects representing each deforestation stage was affected by the grid size (extent). Coarse grid sizes (122,880 m and 61,440 m) were limited to certain stages, whereas the finest grid size (15,360 m) contains all stages described by Sun *et al*. (2014). Due to its strong relation with the spatial fill capacity of a target class, the FD can be sensitive to increases or decreases in the extent as was previously discussed by Sun and Southworth (2013b), who obtained different FD values depending on multiple extents placed over their study area.

Second, although changes can not be completely attributed to spatial resolution of the Terra-i and GFC datasets due to side effects of other features of the datasets' detection methodologies (e.g. inputs, algorithm and processing), it was found that 30 m GFC grid-based objects contained a higher proportion of type 2 and type 3 stages of deforestation (or FD between 1.00 to 1.64) than 240 m Terra-i grid-based objects. This behaviour agreed with Al-Haddam *et al.*'s (2010) results, where multiscale remotely sensed data (250m MODIS, 30m Landsat and 4m IKONOS) was used to characterise forest canopies (tree crowns) in Guatemala using fractal analyses. The authors found that FD values were highest for the finest spatial resolution (4 metres) and lowest for the coarsest spatial resolution (250 metres). The authors claimed that this result can be explained by the greater detail (complexity) that can be captured within an area at the finest resolution, unlike larger pixel sizes that decrease the complexity of the image as individual clumps of the target class are assimilated into larger blocks.

Finally, in practical terms for the study area, according to the aggregated dataset period (2004-2013) the fractal dimension values extracted and deforestation stages adopted showed that forest changes were spatially disperse, with most of them belonging to the initial deforestation stages (type I and type II) for all grid sizes. This association was reinforced due to a dominant behaviour of these stages in the spatial overlapping analyses between Terra-i and GFC grid-based objects. Additionally, less predominant but relevant in the target region was the presence of deforestation type 3 (metastatic growth). This result indicates that in some zones of the Amazon rainforest recent deforestation has compacted as clearings start to agglomerate.

5.2 The mining land-use patterns-based approach

5.2.1 Exploratory analysis of input variables in training sets

The exploratory analysis of fifteen (thirteen FRAGSTAT and two fractal-like) metrics used as inputs in the KDD approach for deforestation patterns demonstrated that values can fluctuate or remain constant depending on grid size and dataset. Moreover, this step was fundamental for detecting (and subsequently confirming by sensitivity analysis) a variable's contribution to discriminating the target deforestation pattern typologies.

Results were not always consistent with the mining deforestation pattern research conducted by Saito (2011), which is the only previous research that fully reports the exploratory analysis performed over a set of FRAGSTAT-like metrics. Supported by results from Wu (2004), the author initially concluded that the ED and MPS metrics did not vary by grid size. Although this behaviour was true for ED in all patterns, it disagreed for MPS, occurring only for the diffuse-related pattern typologies. Other metrics without effects by grid size according to this research included those related with patch area dispersion (PSSD and PSCOV), shape-related metrics (AWMSI and LSI) and perimeter and area relations metrics (MPFD, AWMPFD and MPAR).

According to Saito (2011), robust metrics that are not affected by changes in the extent are highly relevant for discriminating deforestation typologies at multiple scales.

Regarding the discriminating capacity of the metrics assessed, most of them discriminated at least one pattern typology from the remaining typologies. Area-related (CA and PLAND), fractal dimension (D), shape-related (MSI) and area-perimeter related (AWMPFD) metrics seemed to discriminate all four patterns, suggesting their feasibility for incorporation in the data mining procedures. This result did not agree with Saito (2011) who reported ED, MPS and PSSD were the only metrics that presented different median values among all pattern typologies. On the other hand, in agreement with Saito's (2011) research, it was found that area-weighted metrics such as AWMPFD and AWMSI provided better discrimination results than the raw metrics (MPFD and MSI).

Complementary to the Saito's (2011) findings, the set of fractal-like metrics were useful for discriminating deforestation spatial patterns. In this research, FD values were more distributed along the normalised range (0 and 1) for all patterns than their corresponding area-like metrics (CA and PLAND). These slight differences may confirm Encarnação *et al.*'s (2012) results, which pointed out that FD is a measure of space-filling capacity, not a measure of area. Nevertheless, in order to remove any area side-effect within FD, the ratio variable between FD and PLAND showed that this proposed metric could potentially work for discriminating the diffuse extensive pattern typology.

Finally, comparing the effect of dataset type, the only set of metrics that seemed to remain invariant when the dataset changed in all pattern typologies were the area-related metrics describing patches (MSI, PSSD and PSCOV). For the rest, the median values changed for more than three or all four (D, LSI, AWMPFD, MPAR) patterns or less than two patterns (CA, PLAND, PD, ED, MSI, AWMSI, MPFD). For this research, it is not possible to determine a either dataset due to the caveat that they are differentiated not only by this property but also by other features of their detection methodology.

5.2.2 Model evaluation

Overall, the multilayer perceptron ANNs structures assessed proved statistically superior to the C4.5 decision tree-like data algorithm in terms of Kappa median values for mapping deforestation patterns in the Amazon rainforest using either the Terra-i (except at the grid size of 61440 m) or GFC dataset. Although the model construction is less time efficient for ANNs than C4.5 models, this constraint can be compensated for with the latest advances in CPU hardware which permit more efficient model construction using parallel processing (Long and Gupta, 2008), as was performed in this research. Another consideration that may explain the higher performance in ANNs compared to C4.5 was the step of looking for optimal sets of parameters in the former, while in the latter only the default value of its main parameters (confidence limit) was used. Castro *et al* (2013, p. 237) stated that robust machine learning projects in general stress the necessity of improving the expected performance of a learning algorithm at least for a few or all parameters if the aim is to find an optimal model.

Levin (1992) as cited by Sun *et al* (2014) claimed, "There is no single correct scale or level at which to describe a system, nor does that mean all scales serve equally well". This statement agrees with previous mining pattern reference studies (Gavak, 2011; Saito, 2011; Saito *et al*., 2011; De Oliveira, 2014) that have pointed out the relevance of defining an optimal unit of analysis (extent) that ensures consistent pattern characterisation and mapping. In this research, it was determined that the 30720 m grid size may be the best extent to analyse and map the four deforestation patterns described for both Terra-i and GFC forest/non-forest maps. At the second best performance ranking for either overall accuracy or Kappa values, ANN models with a grid size of 15360 m could also have an acceptable performance (not assessed) for mapping the deforestation patterns identified. In contrast, models generated using grid sizes of 61440 m and 122880 m seemed unfeasible, particularly when ANNs models are implemented.

According to Alsakran *et al.* (2014), achieving a near optimal ANN for a specific task requires prior knowledge of the domain problem and deep understanding in choice of network parameters such as weights, hidden layers size and learning rate, among others. In this research, the grid-search procedure with iterative k-fold cross validation supported understanding the complexity (best parameter combinations) of ANNs for pattern mapping analyses. The best combination of ANN parameters (hidden nodes layer and weight decay) and others post-fitted features (total weights and convergence value) tended to vary slightly between grid sizes and datasets, except for the Terra-i selected combinations for ANN models with grid sizes of 30720 m and 122880 m. These combinations had considerably higher values than the remaining best ANN models generated. The higher values can be directly associated with complex structures that usually have drawbacks such as longer time spent in ANN training (Gnana Sheela and

Deepa, 2013) and the trend of memorising noise in the dataset (Augusto and Shapiro, 2007, p. 116), thus tending towards over-training.

Finally, the implementation of the k-fold cross validation resampling method, besides iteratively validating ANN models created from training sets, was also useful for identifying ANN model variance. While the mining deforestation patterns reference studies from Saito (2011) and Gavak (2011) used an iterated Monte Carlo approach, other previous studies on small data sets (Beleites and Salzer, 2008; Beleites *et al*., 2014) found that resampling strategies like the repeated/iterated k-fold cross validation are most appropriate. The values of the bias and variance for Monte Carlo and k-fold cross validation methods were previously discussed by Burman (1989), who refers to the former method as repeated-learning testing-model.

5.2.3 Model sensitivity analysis

Determining variable importance was a key step in this research because it allowed going beyond the "black box", as ANNs typically have been denoted in comparison to other data mining algorithms like decision trees. The results with the adopted sensitivity analysis method successfully ranked the 15 input variables (normalised FRAGSTAT and fractal-like metrics) by pattern typology and dataset. Results suggested that FRAGSTAT-like metrics such as patch size standard deviation (PSSD) may be omitted for future studies using the same ANN architectures established at the grid size of 30720 m for mapping all pattern typologies identified for Terra-i and GFC forest/non-forest grid-based objects. This finding disagrees with results from Saito (2011) using a C4.5 classifier, who claimed that this metric was relevant for discriminating the deforestation spatial patterns identified. This disagreement may result from the algorithms used to determine the importance of each variable between ANN and C4.5 classifiers.

Part of the main findings of this research were to check the contribution of the fractal dimension and derived ratio variable with percentage of land to the output targets (patterns) in the best ANN models. The sensitivity analyses suggest the feasibility of their incorporation for future incoming analyses about deforestation spatial pattern mapping using data mining techniques. The addition of this set of metrics also indicated that further inputs (FRAGSTAT or fractal-like metrics) used to describe deforested landscapes may be explored and assessed. Turner (2005) stated that no single metric can adequately capture the pattern on a given landscape. In this way, and particularly for future pattern analyses using data mining techniques, a set of metrics can be selected that minimize redundancy while capturing the desired qualities (Riitters *et al.,* 1995).

5.2.4 Model use

The maps detailing the distribution of deforestation pattern typologies in the study region (Figure 18 and 19) contribute to findings from previous studies (see for example Mulligan, 2014) that concluded similar differences in the spatial distribution between Terra-i and GFC deforestation detections.

Although Beleites *et al.* (2013) suggested that "well working classifiers need to be validated with at least 75 test cases to obtain confidence intervals that draw practical conclusions about the model", in this research this recommendation was not possible due to limitations in the classified observations, particularly for the multi directional pattern from the Terra-i ANN model. Considering this caveat, performance assessments of the Terra-i and GFC ANNs models with a limited number of samples (30) indicated good results (average Kappa values of 0.70 and 0.73, respectively) according to the ranges suggested by Monserud and Leemans (1992). It is important to note that Kappa values obtained by the resampling technique from surrogate ANN models (using training sets) were higher than the true model performance, suggesting optimistic bias from the surrogates. Additionally, the slightly better Kappa values in the GFC ANN model seemed to indicate its reliability for mapping and describing the distribution of the spatial pattern typologies in the study region.

The confusion matrices analyses provided a better understanding of the ANN models' discrimination capacity. The diffuse extensive patterns were more easily discriminated than the other patterns, which resulted in greater chances of confusion. The heterogeneity (multiple shapes sizes and distribution) of non-forest objects in the fixed grids could be the main cause of confusion among the remaining typologies, as was determined by Saito (2011). The same author stated that certain patterns can represent transition deforestation stages of other patterns and thus it may be difficult to completely discriminate pattern typologies. For instance, in this research diffuse intensive and multi directional patterns could be related to each other, with the latter considered a transition phase of the former.

Finally, in practical terms for the study area, the classified Terra-i and GFC grid-based objects using the best ANN models pointed out a similar conclusion obtained with the fractal analysis approach. Initial deforestation stages, denoted as diffused-related typologies (extensive and intensive), tended to dominate under the mining approach. The spatial overlapping analyses between classified Terra-i and GFC classified grid-based objects, besides strengthening the identification of this dominance, also permitted the conclusion that the geometric pattern is largely distributed in the region in comparison with the multi directional pattern. It is important to highlight that although the mining method apparently produced similar results to the fractal analysis approach, the former allowed finding an optimal extent (in this research a grid size of 30720 m) for analysing the deforestation patterns from both Terra-i and GFC datasets. Additionally, the mining approach favoured the identification of potential agents of forest change assigned initially to each pattern typology described.

6 Conclusions

From the 2004-2013 aggregated forest/non-forest maps constructed from MODIS-based Terra-i and Landsat-based Global Forest Change deforestation datasets, the fractal analysis and data mining methodological frameworks implemented in this research permitted the mapping of different deforestation spatial stages and patterns in the Amazon humid forest. Both frameworks agreed in the use of the fixed non-overlapping grids strategy, in this case four increasing grid sizes (15360 m, 30720 m, 61440 m and 122880 m), which permitted extraction of framework-specific information. The resulting frameworks were reported, analysed and evaluated considering the effects of dataset type and extent (grid size).

Results from the fractal analysis framework, which was based on the box-counting method with the bottom up approach initially described by Sun *et al*. (2014) for computing the fractal dimension from LANDSAT-based forest/non-forest maps, appeared feasible as a means to characterise five deforestation stages (or ranges of fractal dimension values) using multiple fixed grid sizes over the datasets used. The number of grid-based objects representing each deforestation stage was affected by the grid size (extent), with coarse grid sizes (122,880 m and 61,440 m) limited to certain stages and the finest grid size (15,360 m) containing all five stages. Additionally, although changes could not be completely attributed to spatial resolutions of deforestation datasets due to side effects of others features of their detection methodologies (i.e inputs, algorithm and processing), it was found than GFC grid-based objects contained a higher proportion of advanced stages of deforestation (or FD values between 1.00 to 1.64) than Terra-i objects.

Regarding the data mining framework, adapted from earlier works by the Brazilian National Institute for Space Research (INPE) (Silva *et al.* 2011), analysis suggested the suitability of ANNs in comparison to the traditional C4.5 decision tree algorithm. Additionally, as well as the addition of fractal-like metrics extracted from a fractal analysis methodological framework as inputs to models built, besides the traditional landscape ecology metrics used in previous references studies.

In comparison with the fractal analysis method, the data mining method involved more robust modelling processes (data preprocessing, model construction, model evaluation, model sensitivity analysis and model use) that consequently permitted identification of optimal extent(s) or grid size(s) for mapping the identified pattern typologies using fifteen landscape-related metrics (thirteen FRAGSTAT-like and two fractal-like metrics) per deforestation dataset. Overall, the ANNs performed better than the C4.5 algorithm according to the model's performance metrics (Kappa and overall accuracy values). The surrogate model assessments with the true performance assessments of the best ANN model classified grid-based objects indicated that the GFC dataset with a grid size of 30720 m can produce the most accurate results for mapping the four patterns described in the Amazon rainforest region. Supported by the exploratory analysis of model inputs, the sensitivity analysis indicated that the fractal-like metrics seemed to be useful inputs in pattern mapping research using ANNs due to their contribution to the target outputs from the best ANN models analysed. The same analysis also indicated that FRAGSTAT-like metrics such as patch size standard deviation (PSSD) may be omitted for future studies using the same best ANN architectures indicated in this research.

Finally, both fractal and data mining methods suggested the dominance of initial stages of deforestation, or patterns described as dispersed and/or clustered distributed deforested areas. Less common but also relevant in the target region was the presence of the more developed stages of deforestation (type 3) or pattern typologies such as geometric in the fractal analysis and data mining approaches, respectively. This result indicates that in some parts of the Amazon rainforest recent deforestation has started to compact as clearings start to agglomerate in medium and large shapes. In terms of agents of forest change, the most dominant pattern typologies point to spontaneous and small agricultural colonisations as the main drivers of recent forest-change dynamics in the study area.

7
Limitations and future research

Although this research is the first to report the differences and similarities in feasibility between two methodological frameworks (fractal analysis and data mining) for mapping the spatial stages and patterns of recent deforestation datasets, there are some limitations and complementary research needed to produce conclusive information for decision making in forested areas at the regional scale.

The use of aggregated forest/non-forest maps from recent deforestation datasets may have the caveat that some pattern typologies are not well represented in each area analysed due to lack of information of past events of deforestation outside of the datasets' detection periods. For instance, there could be diffuse-like grid-based objects that can be located on developed areas and thus some associations between agents of forest change and pattern typology, particularly diffuse-related types, may not be appropriate. A similar issue can occur with the fractal analysis approach, wherein deforestation stages (fractal dimension ranges) need to be reevaluated for adjusting them to datasets providing recent deforestation detections such as Terra-i and GFC.

In terms of the feasibility of data mining models, there are several pathways that can strengthen their potential to discriminate and map deforestation spatial patterns as a follow-up to FF-MLP ANN models. First, as corroborated by the sensitivity analysis, the addition of new model inputs such as fractal-like metrics can be relevant to discriminate the output targets (patterns typologies). It is thus suggested to explore more spatial landscape heterogeneity measurements such as FRAGSTAT-like (both at class and landscape level) and fractal-related metrics (i.e. lacunarity) in pattern mapping research using data mining techniques. Second, from the knowledge discovery perspective, although this research reported ANN input contributions and best parameter combinations according to the Kappa value using the iterative k-fold cross validation method, for providing useful information for decision making purposes it will be important to visualise and share the ANN configurations in an easy-to-understand way (see for example Dias *et al.*, 2012), as well as decision tree algorithms (Stiglic *et al.*, 2012). Next, the quality of the classification process depends greatly on the representativeness and sufficiency of the amount of the labelled data used to generate a classifier, a fact that can be considered one of the main limitations of the data mining framework of this research. For this issue, it is suggested to investigate learning curves that permit finding a sufficient level of labelled samples, as was performed by Beleites *et al.* (2013), or to implement semi-supervised methods that can use both labelled and unlabelled data. The use of additional unlabelled data has been shown to offer significant improvements in comparison to classifiers generated only on the basis of labelled data (Gabrys and Petrakieva, 2004). The balanced sampling strategy implemented for model building and validation may limit the amount of validation samples and thus the classifiers' potential performance. Other sampling strategies such as using imbalance datasets (Chawla, 2005) should be considered, as the representative areas for certain patterns (such as multi directional) are smaller than other patterns such as the diffuse-related types.

Finally, in regard to the software and packages used to implement the methodological frameworks, their performance was improved with parallel processing. It is suggested to explore others languages more efficient than R, such as JAVA or C, which may contain libraries with faster processing than R-based packages, particularly for performing fractal dimension extraction and model building procedures.

References

Al-Hamdan, M., Cruise, J., Rickman, D. and Quattrochi, D. (2010) Effects of Spatial and Spectral Resolutions on Fractal Dimensions in Forested Landscapes. *Remote Sensing* 2(3), 611-640.

Alsakran, J., Rodan, A., Alhindawi, N. and Faris, H. (2014) Visualization analysis of feed forward neural network input contribution. *Scientific Research and Essays* 9(14), 645-651.

Augusto, J.C. and Shapiro, D. (2007) *Advances in Ambient Intelligence*. First edition. IOS Press: Amsterdam.

Baggio, J., Salau, K., Janssen, M., Schoon, M., and Bodin, O. (2011). Landscape connectivity and predator prey population dynamics. *Landscape Ecology* 26(1), 33-45.

Batistella M., Robeson S. and Moran E. (2003) Settlement design, forest fragmentation, and landscape structure in Rondonia, Amazonia. *Photogrammetric Engineering and Remote Sensing* 69(7), 805–812.

Beck, M. (2015). *NeuralNetTools: Visualization and Analysis Tools for Neural Networks* [Online]. Available from: https://cran.r-project.org/web/packages/NeuralNetTools/index.html [Accessed 15 July 2015].

Beleites, C. and Salzer R. (2008) Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry* 5, 1261-71.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. and Popp, J. (2013) Sample size planning for classification models. *Analytica Chimica Acta* 760, 25-33.

Beyer, H. L. (2015) *Geospatial modelling environment Version: 0.7.2 RC2* [Online]. Available from: http://www.spatialecology.com/gme/gmehistory.htm [Accessed 10 June 2015].

Bianconi, A., Von Zuben, C.J., Serapião A.B. and Govone J.S. (2010) Artificial neural networks: A novel approach to analysing the nutritional ecology of a blowfly species, Chrysomya megacephala. *Journal of Insect Science* 10(58), 1-18.

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press: Oxford.

Bunn, G., Urban, L., and Keitt, H. (2000) Landscape connectivity: a conservation application of graph theory. *Journal of Environmental Management* 59(4), 265-278.

Burman, P. (1989) A comparative study of ordinary cross-validation, r-fold cross-validation and the repeated learning-testing methods. *Biometrika* 76(3), 503-514.

Castro, F., Gelbukh, A. and González, M. (2013) *Advances in Soft Computing and Its Applications*. 12th Mexican International Conference, MICAI 2013, Mexico City.

Chawla, N. V. (2005) *Data Mining for Imbalanced Datasets: An Overview. Chapter: Data Mining and Knowledge Discovery Handbook*. pp 853-867.

Cho, J. H. and P.U. Kurup (2011) Decision tree approach for classification and dimensionality reduction of electronic nose data. Sensors and Actuators B: Chemical 160(1).

Clark, T. Aide, M., Ricardo Grau, H. and Riner, G. (2010). A scalable approach to mapping annual land cover at 250 m using MODIS time series data: A case study in the Dry Chaco ecoregion of South America. *Remote Sensing of Environment* 114(11), 2816-2832.

Clark,, M. L., Aide, T. M. and Riner, G. (2012) Land change for all municipalities in Latin America and the Caribbean assessed from 250-m MODIS imagery (2001–2010). *Remote Sensing of Environment*

Coca, A. 2015. *How can the shapes and distribution of deforested areas inform us about the agents of changes on the ground?* [Online]. Available from: http://goo.gl/T9pSdt [Accessed 21 July 2015].

Das, A. (2015) *Guide to Signals and Patterns in Image Processing. Foundations, Methods and Applications*. Springer: UK.

Dias, M. M., Yamaguchi, J. K., Rabelo, E. and Franco, C. (2012) *Visualization Techniques: Which is the Most Appropriate in the Process of Knowledge Discovery in Data Base?* [Online]. Available from: http://goo.gl/4a258F [Accessed 15 July 2015].

Dunham, M. H. (2002) *Data Mining, Introductory and Advanced Topics*. Prentice Hall: USA.

Encarnação, S., Gaudiano, M., Santos, F., Tenedório, J. and Jorge, M.P. (2012) Fractal cartography of urban areas. *Scientific Reports* 2, 57.

Ewers, R. M., and Laurance, W. F. (2006) Scale-dependent patterns of deforestation in the Brazilian Amazon. *Environmental Conservation* 33(3), 203-211.

Fayyad, U, Haussler, D. and Stolorz, P. (1996) KDD for Science Data Analysis: Issues and Examples. *In Proceedings of KDD-96*, Menlo Park, CA: AAAI Press.

Feigenbaum, E.A., and Simon, H.A. (1963). Performance of a reading task by an elementary perceiving and memorizing program. *Behavioral Science* P-2358.

Gabrys, B. and Petrakieva, L. (2004) Combining labelled and unlabelled data in the design of pattern classification systems. *International Journal of Approximate Reasoning* 35(3), 251-273.

Gavak, E. S. (2011) *Dinâmica de padrões de mudança de uso e cobertura da terra na região do Distrito Florestal Sustentável da BR-163*. Thesis (Master). Brazil's National Institute for Space Research's Centre.

Geist H. J. and Lambin E. F. (2001) What drives tropical deforestation? A meta-analysis of proximate and underlying causes of deforestation based on subnational case study evidence. *In*: LUCC Report Series No. 4, CIACO, Louvain-la-Neuve, Belgium. 118 p.

Gnana Sheela, K.G. and Deepa, S.N. (2013) Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Mathematical Problems in Engineering* ID 425740.

Grobbelaar, S. and Visser, J.K. (2015) Determining the cost of predictive component replacement in order to assist with maintenance decision-making. *South African Journal of Industrial Engineering* 26(1), 150-162.

Han, J. and Kamber, M. (2006) *Data Mining: Concepts and Techniques*. Second edition. Elsevier: San Francisco.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A.., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O. and Townshend, J. R. G. (2013) High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* 342(6160), 850-853.

Hudson, A., Fontès, J., Jeanjean, H., Miquel, C., Puig, H. and Solier, C. (1995) Study of forest non-forest interface: Tipology of fragmentation of tropical forest. *In*: Tropical Ecosystem Environment Observation by Satellite (TREES) Series B: Research Reports No.2. European Commission Joint Research Centre and European Space Agency, Luxembourg. EUR-16291EN. 90 p.

Kim, D. (1998) Normalization methods for input and output vectors in back propagation neural networks. International Journal of Computer Mathematics 71(2), 161-171.

Kuhn, M. (2015) *caret: Classification and Regression Training* [Online]. Available from: https://cran.r-project.org/web/packages/caret/index.html [Accessed 15 June 2015].

Lambin, E. F., Geist H. J. and Lepers, E. (2003) Dynamics of land-use and land-cover change in tropical regions. *Annual Review of Environment and Resources* 28, 205-241.

Lee, S. and I. Park (2013) Application of decision tree model for the ground subsidence hazard mapping near abandoned underground coal mines. J*ournal of Environmental Management* 127(30), 166-176.

Li, H. and Reynolds, J. F. (1994) A simulation experiment to quantify spatial heterogeneity in categorical maps. *Ecology* 75(8), 2446-2455.

Li, J., Du, Q. and Sun, C. (2009) An improved box-counting method for image fractal dimension estimation. *Pattern Recognition* 42(11), 2460-2469.

Lippit, C. D., Rogan, J., Li, Z., Eastman, R. J. and Jones, T. J. (2008.) Mapping Selective Logging in Mixed Deciduous Forest: A Comparison of Machine Learning Algorithms. *Photogrammetric Engineering and Remote Sensing* 74(10), 1201-1211.

Long, L. N. and Gupta, A. (2008) Scalable Massively Parallel Artificial Neural Networks. *Journal of Aerospace Computing, Information, and Communication* 5(1), 3-15.

Maciel, A. M. (2014) *Mineração de grafos em dados espaciais de desmatamento*. Thesis (Master). Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semiárido

Mandelbrot, B. (1975) Stochastic models for the earth's relief, the shape and fractal dimension of coastlines, and the number area rule for islands. *Proc. National Acad.-Sc* 72(10), 2825-2828.

McCulloch, W. S. and Pitts, W. H. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 7(1), 115-133.

McGarigal, K. (2015) *Fragstat help v. 4.2.* [Online]. http://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf [Accessed 10 June 2015].

McGarigal, K. and Marks, B. J. (1995) *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure*. [Online]. Available from: http://www.umass.edu/landeco/pubs/mcgarigal.marks.1995.pdf [Accessed 15 February 2015].

Mertens, B. and Lambin E. F. (1997) Spatial modelling of deforestation in southern Cameroon: spatial disaggregation of diverse deforestation processes. *Applied Geography* 17(2), 143-162.

Minor, E. S., and Urban, D. L. (2008) A graph-theory framework for evaluating landscape connectivity and conservation planning. *Conservation Biology* 22(2), 297-307.

Monserud, R.A. and Leemans, R. (1992) Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling* 62(4), 275-293.

Mulligan, M. (2014) What you see depends on how you see: differences between deforestation monitoring systems as continuous fields [Online]. http://blog.policysupport.org/2014/03/what-you-see-depends-on-how-you-see.html [Accessed 11 July 2015].

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D. (2004) An introduction to decision tree modeling. *J. Chemometrics* 18(6), 275-285

Neel, M.C., McGarigal, K. and Cushman, S.A. (2004) Behavior of class-level landscape metrics across gradients of class aggregation and area. *Landscape Ecology* 19, 435-455.

Olden J. D., and Jackson, D. A. (2002) Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* 154(1–2), 135-150.

Olden, J. D., Joy, M. K. and Death, R. G. (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* 178, 389–397.

Oliveira, F. J. B. and Metzger, J. P. (2006) Thresholds in landscape structure for three common deforestation patterns in the Brazilian Amazon. *Landscape Ecology* 21(7), 1061–1073.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., Kassem, K. R. (2001) Terrestrial ecoregions of the world: a new map of life on Earth. *Bioscience* 51(11), 933-938.

Peitgen, H.O, Jurgens, H. and Saupe (1992) *Chaos and Fractals: New Frontiers of Science*. First edition. Springer:Berlin.

Peña, J.M., Gutiérrez, P.A., Hervás-Martínez, C., Six, J., Plant, R.E. and López-Granados, F. (2014) Object-Based Image Classification of Summer Crops with Machine Learning Methods. *Remote Sensing* 6(6), 5019-5041.

Prist, P. R., Michalski, F. and Metzger, J. P. (2012) How deforestation pattern in the Amazon influences vertebrate richness and community composition. *Landscape Ecology* 27(6), 799-812.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann: Los Altos.

RAISG: Amazonian Network of Georeferenced Socio-environmental Information. (2012) *Amazonia under pression*. [Online]. Available from: http://raisg.socioambiental.org/system/files/Amazonia%20under%20pressure16_05_2013.pdf [Accessed 14 February 2015].

Reymondin, L. Jarvis, A., Perez-Uribe, A., Touval, J., Argote, K., Rebetez, J., Guevara, E. and Mulligan, M. (2012) A methodology for near real-time monitoring of habitat change at continental scales using MODIS-NDVI and TRMM. *Submitted Remote Sensing of Environment*.

Riitters, K. H., O’Neill, R. V., Hunsaker, C. T., Wickham, J. D.,Yankee, D. H., Timmons, S. P., Jones, K. B. and Jackson, B. L. (1995) A factor analysis of landscape pattern and structure metrics. *Landscape Ecology* 10, 23-40.

Riitters, K. H., Wickham, J. D., O'Neill, R. V., Jones, K. B., Smith, E. R., Coulston, J. W., Wade, T.G. and Smith, J. H. (2002). Fragmentation of continental United States forests. *Ecosystems* 5(1), 815-822.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press: Cambridge.

Roiger, R.J. and Geatz, M.W. (2003) *Data Mining - A Tutorial-Based Primer*. Addison Wesley: Boston.

Saito, E. S. (2011) *Caracterizacão de trajetórias de padrões de ocupacão humana na amazônia legal por meio de mineracão de dados*. Thesis (Master). Brazil's National Institute for Space Research's Centre.

Saito, E. S., Fonseca, L. M. G., Escada, M. I. S. and Korting, T. S. (2011) Efeitos da mudança de escala em padrões de desmatamento na Amazônia. *Revista Brasileira de Cartografia* 63(3), 401-414.

Sánchez-Cuervo A. M., Aide , Clark M. L. and Etter A. (2012) Land Cover Change in Colombia: Surprising Forest Recovery Trends between 2001 and 2010. *PLoS ONE* 7(8), e43943.

Saura, S. and Pascual-Hortal, L. (2007) A new habitat availability index to integrate connectivity in landscape conservation planning: comparison with existing indices and application to a case study. *Landscape and Urban Planning* 83(1), 91-103.

Silva, M. P. S. (2005) Analysis of amazon deforestation patterns and their relation to social economical and ecological process [Online]. Available from: http://goo.gl/jDSJ74 [Accessed 21 June 2015].

Silva, M. P. S., Câmara, G., Escada, M. I. S., Souza, R. C. M. and Valeriano, D. M. (2005) Mining patterns of change in remote sensing image databases. *In:* IEEE*,* ed. *5th IEEE Internacional Conference on data mining, Houston*. 27-30 November 2005. Houston: IEEE, 362-369.

Steinwand, D. R. (1994) Mapping raster imagery to the Interrupted Goode Homolosine projection. *International Journal of Remote Sensing* 15(17), 3463-3471.

Stiglic, G., Kocbek, S., Pernek, I., Kokol, P. (2012) Comprehensive Decision Tree Models in Bioinformatics. *PLoS ONE* 7(3), e33812.

Sun, J. and Southworth, J. (2013a) Retrospective analysis of landscape dynamics using normalized spectral entropy. *Remote Sensing Letters* 4(11), 1049-1056.

Sun, J. and Southworth, J. (2013b) Remote Sensing-Based Fractal Analysis and Scale Dependence Associated with Forest Fragmentation in an Amazon Tri-National Frontier. *Remote Sensing* 5(2), 454-472.

Sun, J., Huang, Z., Southworth, J. and Qiu, Y. (2013). Mapping fractality during the process of deforestation in an Amazon tri-national frontier. *Remote Sensing Letters* 4(6), 589-598.

Sun, J., Huang, Z., Zhen, Q., Southworth, J. and Perz, S. (2014) Fractally deforested landscape: Pattern and process in a tri-national Amazon frontier. *Applied Geography* 52(1) 204-211.

Terra-i Project (2015). *Panoramio channel of the Terra-i project* [Online]. Available from: http://www.panoramio.com/user/7916628 [Accessed 15 July 2015].

TNC: The Nature Conservancy (2015) *Applicability of the Hansen Global Forest Data to REDD+ Policy Decisions* [Online]. Available from: http://goo.gl/0cp5SP[Accessed 15 July 2015].

Turner, M. 2005. Landscape ecology: What is the State of the Science? *Annual Review of Ecology , Evolution, and Systematics* 36, 319-344.

Turner, M. G., Gardner, R. H. and O'Neill, R. V. (2001) *Landscape ecology in theory and practice: Pattern and process*. New York: Springer.

Vogt, P., Riitters, K., Estreguil, C., Kozak, J., Wade, T., and Wickham, J. (2007) Mapping spatial patterns with morphological image processing. *Landscape Ecology* 22(2), 171-177.

Weston, S. (2015). *doParallel: Foreach parallel adaptor for the parallel package* [Online]. Available from: https://cran.r-project.org/web/packages/doParallel/index.html [Accessed 15 July 2015].

Williamson, J. M., Manatunga, A. and Lipsitz, S. R. (2000) Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* 1-2, 191-202.

Wu, J. (2004) Effects of changing scale on landscape pattern analysis: scaling relations. *Landscape ecology* 19(1), 125-138.

WWF: World Wide Foundation (2009). *Terrestrial habitats* [Online]. Available from: http://wwf.panda.org/about_our_earth/ecoregions/about/habitat_types/selecting_terrestrial_ecoregions [Accessed 15 February 2015].

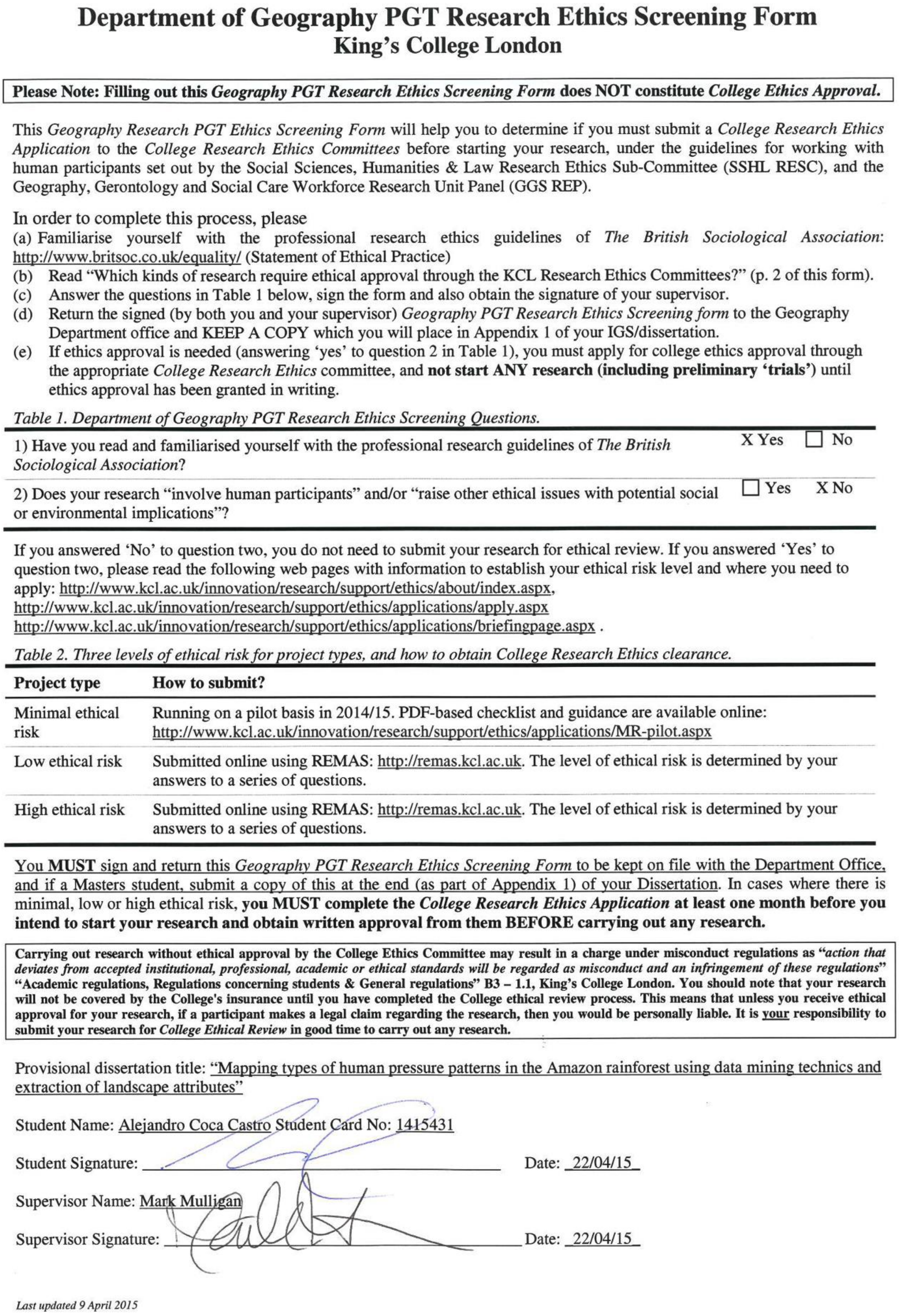
Yan, A. (2008) *Analysis on protein structures using statistical and bioinformatical methods*. Thesis (Doctor of Philosophy). Iowa State University.

Yan, X. -S., Yang, Gandomi, A.H., Talatahari, S. and Alavi, A. H. (2012) *Metaheuristics in Water, Geotechnical and Transport Engineering* [Online]. Available from: https://goo.gl/MSYu5E [Accessed 10 July 2015].

Zaccarelli, N., Li, B. -L., Petrosillo, I. and Zurlini, G. (2012) Order and disorder in ecological time-series: introducing normalized spectral entropy. *Ecological Indicators* 28(1), 22-30.

Zurlini, G., Riitters, K., Zaccarelli, N. and Petrosillo, I. (2007). Patterns of disturbance at multiple scales in real and simulated landscapes. *Landscape Ecology* 22(5), 705-721.

Appendix 1. KCL Ethics Screening and Risk Assessment Forms



Department of Geography PGT Research Ethics Screening Form

King's College London

Please Note: Filling out this *Geography PGT Research Ethics Screening Form* does NOT constitute *College Ethics Approval.*

This *Geography Research PGT Ethics Screening Form* will help you to determine if you must submit a *College Research Ethics Application* to the *College Research Ethics Committees* before starting your research, under the guidelines for working with human participants set out by the Social Sciences, Humanities & Law Research Ethics Sub-Committee (SSHL RESC), and the Geography, Gerontology and Social Care Workforce Research Unit Panel (GGS REP).

In order to complete this process, please

(a) Familiarise yourself with the professional research ethics guidelines of *The British Sociological Association*: http://www.britsoc.co.uk/equality/ (Statement of Ethical Practice)

(b) Read "Which kinds of research require ethical approval through the KCL Research Ethics Committees?" (p. 2 of this form).

(c) Answer the questions in Table 1 below, sign the form and also obtain the signature of your supervisor.

(d) Return the signed (by both you and your supervisor) *Geography PGT Research Ethics Screening form* to the Geography Department office and KEEP A COPY which you will place in Appendix 1 of your IGS/dissertation.

(e) If ethics approval is needed (answering 'yes' to question 2 in Table 1), you must apply for college ethics approval through the appropriate *College Research Ethics* committee, and **not start ANY research (including preliminary 'trials')** until ethics approval has been granted in writing.

Table 1. Department of Geography PGT Research Ethics Screening Questions.

1) Have you read and familiarised yourself with the professional research guidelines of *The British Sociological Association*?	X Yes □ No
2) Does your research "involve human participants" and/or "raise other ethical issues with potential social or environmental implications"?	□ Yes X No

If you answered 'No' to question two, you do not need to submit your research for ethical review. If you answered 'Yes' to question two, please read the following web pages with information to establish your ethical risk level and where you need to apply: http://www.kcl.ac.uk/innovation/research/support/ethics/about/index.aspx, http://www.kcl.ac.uk/innovation/research/support/ethics/applications/apply.aspx http://www.kcl.ac.uk/innovation/research/support/ethics/applications/briefingpage.aspx .

Table 2. Three levels of ethical risk for project types, and how to obtain College Research Ethics clearance.

Project type	How to submit?
Minimal ethical risk	Running on a pilot basis in 2014/15. PDF-based checklist and guidance are available online: http://www.kcl.ac.uk/innovation/research/support/ethics/applications/MR-pilot.aspx
Low ethical risk	Submitted online using REMAS: http://remas.kcl.ac.uk. The level of ethical risk is determined by your answers to a series of questions.
High ethical risk	Submitted online using REMAS: http://remas.kcl.ac.uk. The level of ethical risk is determined by your answers to a series of questions.

You **MUST** sign and return this *Geography PGT Research Ethics Screening Form* to be kept on file with the Department Office, and if a Masters student, submit a copy of this at the end (as part of Appendix 1) of your Dissertation. In cases where there is minimal, low or high ethical risk, **you MUST complete the *College Research Ethics Application* at least one month before you intend to start your research and obtain written approval from them BEFORE carrying out any research.**

Carrying out research without ethical approval by the College Ethics Committee may result in a charge under misconduct regulations as "*action that deviates from accepted institutional, professional, academic or ethical standards will be regarded as misconduct and an infringement of these regulations*" "Academic regulations, Regulations concerning students & General regulations" B3 – 1.1, King's College London. You should note that your research will not be covered by the College's insurance until you have completed the College ethical review process. This means that unless you receive ethical approval for your research, if a participant makes a legal claim regarding the research, then you would be personally liable. It is your responsibility to submit your research for *College Ethical Review* in good time to carry out any research.

Provisional dissertation title: "Mapping types of human pressure patterns in the Amazon rainforest using data mining technics and extraction of landscape attributes"

Student Name: Alejandro Coca Castro Student Card No: 1415431

Student Signature: __________ Date: 22/04/15

Supervisor Name: Mark Mulligan

Supervisor Signature: __________ Date: 22/04/15

Last updated 9 April 2015

3 RISK ASSESSMENT FORM AND ASSOCIATED DOCUMENTATION

After reading through ALL risk categories, please select RISK TYPE A or B below.

RISK TYPE A

You are only eligible for RISK TYPE A if all of the following are true:

- **Your work takes place within: college premises or home or within organizations/premises that have their own clear risk assessment in place.**
- **Your work involves ONLY library/archival data or existing on-line/other data.**
- **Your work WILL NOT expose you to risks greater than in everyday life.**

DECLARATION: I have considered ALL categories in this form (see page 4 onwards) and I declare that I am undertaking a student project/dissertation where: a) NONE of my research will be outside of college premises or home or organizations/premises that have their own clear risk assessment in place; and b) it does not involve ANY of the risks identified in ANY of the categories of this risk assessment form. Should my research project change, such that there are now risks involved, then it is my responsibility to resubmit this form after completing an assessment for Risk Type B.

SIGNATURES OF PERSON FILLING IN A RISK ASSESSMENT AND COUNTERSIGNATURE.
A. Person filling in this risk assessment
Signature **(TYPE YOUR NAME AND STAFF OR STUDENT ID IN PLACE OF A SIGNATURE):** ALEJANDRO COCA CASTRO
Date: 29ND JUNE 2015
B. Countersignature and date. I sign to indicate that I have read this and consider it an appropriate assessment. (Students – Research Supervisor; Research Staff – Project Leader; Academic Staff – Head of Department)
Signature **(TYPE YOUR NAME AND STAFF OR STUDENT ID IN PLACE OF A SIGNATURE):**
Date:

RISK TYPE B

Fill out THIS PAGE and ALL OTHER PAGES in this form.

DECLARATION: I have considered ALL categories in this form and have indicated which risks apply to me that are greater than in everyday life and normal activities (writing yes/no for every section). Where I have answered 'yes' then I have also indicated the degree of risk from 1–5 (1=low, 5=high) and, where appropriate, added notes or comments relating to the level of risk. I have identified and added any additional risks not explicitly covered by this form in the final section.

SIGNATURES OF PERSON FILLING IN A RISK ASSESSMENT AND COUNTERSIGNATURE.
A. Person filling in this risk assessment
Signature **(TYPE YOUR NAME AND STAFF OR STUDENT ID IN PLACE OF A SIGNATURE):**
Date:
B. Countersignature and date. I sign to indicate that I have read this and consider it an appropriate assessment. (Students – Research Supervisor; Research Staff – Project Leader; Academic Staff – Head of Department)
Signature **(TYPE YOUR NAME AND STAFF OR STUDENT ID IN PLACE OF A SIGNATURE):**
Date:

PROJECT DETAILS

Please ensure you complete the TRAVEL NOTIFICATION AND INSURANCE FORM (available at http://bit.ly/kclgeotravel) at least two weeks before leaving. For work outside of the UK, please do not forget to obtain insurance in accordance with College regulations.

A. The project title.
Mapping types of human pressure patterns in the Amazon rainforest using data mining technics and extraction of landscape attributes

B. The overall aim of the project.
To map types of human pressure patterns in the Amazon rainforest through the integration of landscape fragmentation metrics and data mining techniques.

C. Can you identify any risk to others of the activity that you propose? *(If yes, please elaborate.)*
This research is entirely desk-based.

D. Can you identify any reputational or business risk to the college of the activity that you propose? *(If yes, please elaborate.)*
This research does not consider any reputational or business risk to the college. Any credit for findings from others authors will be acknowledged or properly reference in this work.

E. The location(s) where the data will be collected.
The analyses will be performed within college facilities and at home. The input data and literature will be gathered through library/archival data or online databases.

Are you planning to work alone? (***If yes,*** *please justify this decision and detail the crisis management measures in place if something goes wrong in the field. How will you summon help? Who will know where you are and raise the alert if you are not back when planned and what are their contact details?* ***If no****, please provide name and contact numbers of colleague(s) who will be assisting you with the fieldwork if applicable, and/or named contact who will be provided with information on the timings and locations of your work.*) N/A

G. Do you have permission to do this work from the landowner or person responsible for the site you are working in? *Please elaborate*. N/A

H. Are you leaving any equipment on site? If so, do you have permission to do so and who is overseeing the equipment locally when you are not there? *Please elaborate*. N/A

I. Please briefly list the methods or techniques that will be employed for data collection during the project.

- Creating multi-temporal annual forest/non-forest maps from two deforestation datasets (Terra-i and Global Forest Cover);
- Characterizing and mapping human pressure spatial patterns by dataset integrating landscape fragmentation metrics and data mining techniques;
- Comparing and discussing the spatial patterns mapped results between both deforestation datasets.

J. Please elaborate on other relevant Project Details not mentioned above or in the Personal Details (below).
This research is entirely desk-based.

PERSONAL DETAILS

Data Protection: This information will remain confidential and will only be used by the Department in the event of an emergency or urgent need to contact you.

Please note that, other than for very low risk environments, we do not sanction lone field working. In all cases, you must have a nominated contact who should know where you are, be in potentially immediate contact with you (by phone) and who should have the timings of your departure and return from work, so that an alarm might be raised.

Your Name: Alejandro Coca Castro

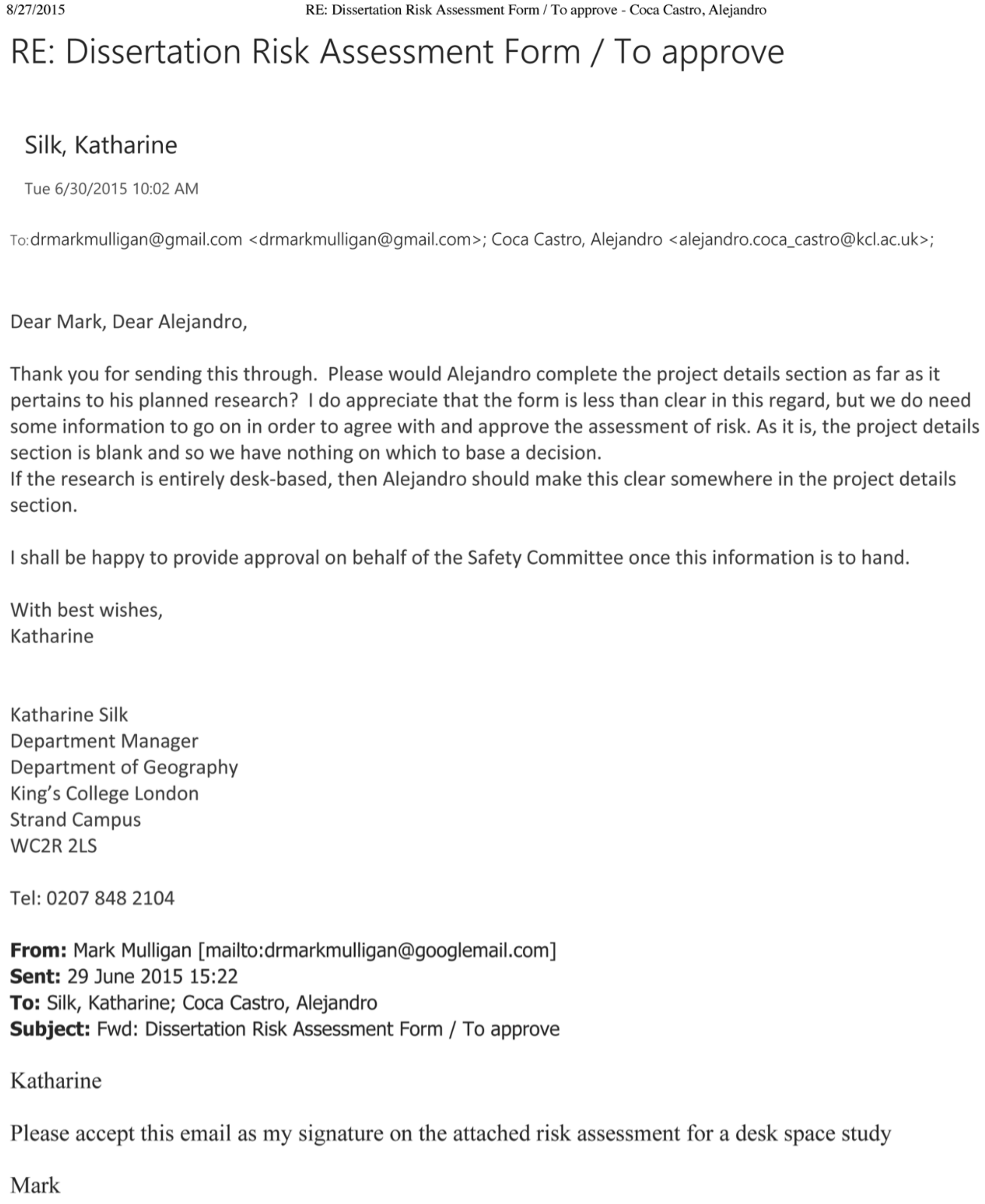
Next of Kin (name): Liliana Castro

Next of Kin contact details: phone (07891581380) / address (30C Northampton street) / postal code (N12UX)

Any relevant medical information which might impact upon your study or safety: N/A

Address of nearest local hospital(s) to project or fieldwork site(s): N/A

Name of local contact person(s) if applicable: N/A



8/27/2015 RE: Dissertation Risk Assessment Form / To approve - Coca Castro, Alejandro

RE: Dissertation Risk Assessment Form / To approve

Silk, Katharine

Tue 6/30/2015 10:02 AM

To:drmarkmulligan@gmail.com <drmarkmulligan@gmail.com>; Coca Castro, Alejandro <alejandro.coca_castro@kcl.ac.uk>;

Dear Mark, Dear Alejandro,

Thank you for sending this through. Please would Alejandro complete the project details section as far as it pertains to his planned research? I do appreciate that the form is less than clear in this regard, but we do need some information to go on in order to agree with and approve the assessment of risk. As it is, the project details section is blank and so we have nothing on which to base a decision.
If the research is entirely desk-based, then Alejandro should make this clear somewhere in the project details section.

I shall be happy to provide approval on behalf of the Safety Committee once this information is to hand.

With best wishes,
Katharine

Katharine Silk
Department Manager
Department of Geography
King's College London
Strand Campus
WC2R 2LS

Tel: 0207 848 2104

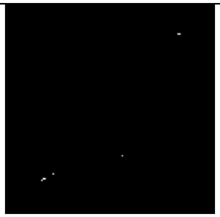
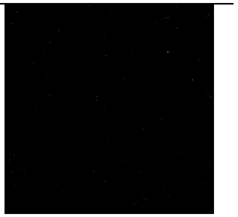
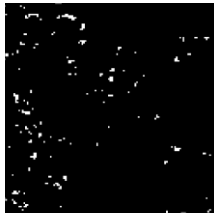
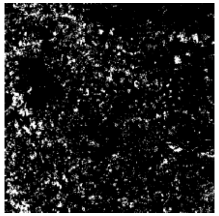
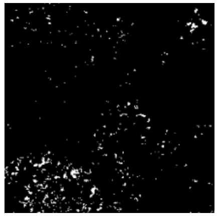
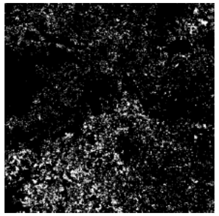
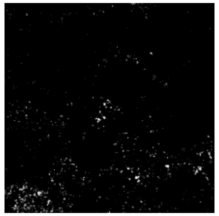
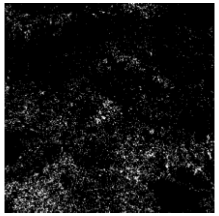
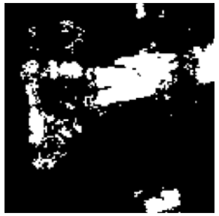
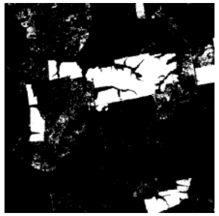
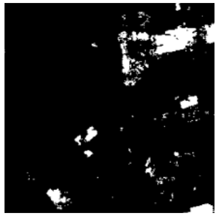
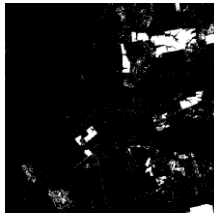
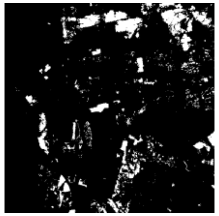
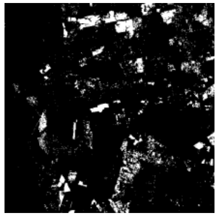
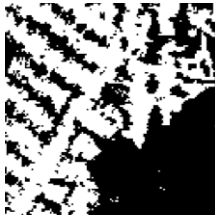
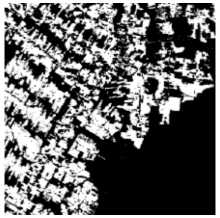
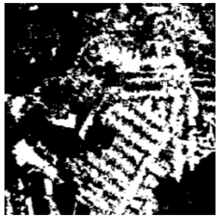
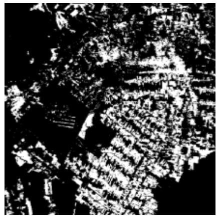
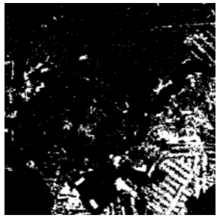
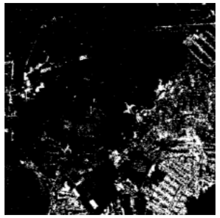
From: Mark Mulligan [mailto:drmarkmulligan@googlemail.com]
Sent: 29 June 2015 15:22
To: Silk, Katharine; Coca Castro, Alejandro
Subject: Fwd: Dissertation Risk Assessment Form / To approve

Katharine

Please accept this email as my signature on the attached risk assessment for a desk space study

Mark

Appendix 2. Representative grid-based objects for deforestation spatial pattern typologies in three out of four fixed grid sizes.

	Grid size (and scale) by deforestation dataset ■ **Forest** □ **Non-Forest**					
	30720 m x 30720 m		*61440 m x 61440 m*		*122880 m x 122880 m*	
Pattern	*(scale 1:150,000)*		*(scale 1:300,000)*		*(scale 1:600,000)*	
	Terra-i (240 m)	*GFC (30 m)*	*Terra-i (240 m)*	*GFC (30 m)*	*Terra-i (240 m)*	*GFC (30 m)*
Diffuse extensive						
Diffuse extensive						
Geometric						
Multi-directional						

Appendix 3A. Description of the selected set of metrics by conceptual category (Neel *et al.*, 2004) used the mining land-use patterns-based approach. Their meaning was adapted to the research context (non-forest as the target class; and grid-based objects at multiple fixed grid sizes as units of analysis or landscape).

FRAGSTAT-like Area/Density/Edge category metrics

- Class area (*CA*) is the sum of the area of all non-forest patches in a grid-based object. It is useful to compare non-forest surfaces of multiple grid-based objects with the same extent (a fixed grid size);
- Percentage of landscape (*PLAND*) reflects the contribution of the non-forest patches in terms of area on a grid-based object. *PLAND* is an appropriate measure for comparison non-forest surfaces of multiple grid-based objects among landscapes of varying sizes (i.e multiple fixed grid sizes);
- Patch density (*PD*) indicates ratio of number of patches by class area, it facilitates comparisons among landscapes of varying sizes (multiple fixed grid sizes);
- Edge density (ED) is the total length of the patch edge per unit area within each landscape (grid-based object). ED increases with forest disturbance, it is a direct measure of forest fragmentation. It is useful to compare non-forest edges of multiple grid-based objects with the same extent (a fixed grid size);
- Mean patch size (MPS) is another measure of forest fragmentation. Patch types with smaller MPS might be considered more fragmented. A given MPS value can refer either to patches of the same size or to patches of very different sizes;
- Patch size standard deviation (*PSSD*) and Patch Size Coefficient of Variation (PSCV) are variability metrics and indicate aspects related to landscape heterogeneity (grid-based object). Thus, grid-based objects with greater PSSD and PSCV are more heterogeneous and grid-based objects with lower PSSD and PSCV are more uniform.

FRAGSTAT-like Shape category metrics

- Landscape shape index (LSI) captures the complexity of all non-forest patches boundaries in a grid-based object by calculating a normalized ratio of their perimeter to their area. LSI equals 1 when non-forest patches are in average maximally compact and increases without limit as their shapes becomes more irregular;

- Mean shape index (*MSI*) is the average perimeter to area ratio for the non-forest class. As LSI metric, MSI provides a relative measurement of shape complexity;
- Area-weighted mean shape index (*AWMSI*) has the same calculation as *MSI*, excepted is weighted by the size of its patches. The particularity about *AWMSI* is that larger patches are weighted more heavily than smaller patches in calculating the average patch shape;
- Mean patch fractal area dimension (MPFD) reveals the relationship between shape and area of grid-based object constituted by non-forest patches. MPFD measures the average shape complexity of all non-forest patches within the grid-based object. It ranges from 1 to 2, with 1 meaning Euclidian geometric shapes such as circles and squares, and 2 meaning a very complex patch shape.
- Area-weighted patch fractal area dimension (*AWMPFD*) has the same calculation as *MPFD*, excepted is weighted by the size of its patches. This improves the measure of class patch forest fragmentation because the structure of smaller patches is often determined more by image pixel size than by characteristics of natural features found in the landscape (Herold, 2004). *AWMPFD* has the same range of values and interpretation as *MPFD*.
- Mean perimeter-area ratio (*MPAR*) is the mean value of the perimeter/area ratio for all non-forest patches constituting a grid-based object. This indicator expresses the complexity of the non-forest patches' elements.

Fractal-like metrics

- Fractal dimension (*D*) reveals the spatial fill capacity, or the extent to which non-forest areas can occupy a grid-based object.
- Ratio fractal dimension by percentage of non-forest in the landscape, is a empirical metric to cancel any proportion or areal effect contained in the raw fractal dimension metric.

Appendix 3B. FRAGSTAT and fractal-like metrics extracted over the non-forest class grid-based objects for implementing the mining land-use patterns-based approach. See further description of symbols and abbreviations in metrics formulas at the end of table. Source of FRAGSTAT-like metrics features: Batistella (2001) and Saito *et al*. (2011).

Metric	*Metric type*	*Formula*	*Range (and units)*
Class area (*CA*)	FRAGSTAT-like	$CA = \sum_{j=1}^{n} \alpha_{ij}(\frac{1}{10{,}000})$	*CA* > 0, without limit (ha)
Landscape percentage (*PLAND*)	FRAGSTAT-like	$PLAND = P_i = \frac{\sum_{j=1}^{n} \alpha_{ij}}{A}(100)$	0 < *PLAND* ≤ 100 (percentage)
Patch density (*PD*)	FRAGSTAT-like	$PD = \frac{n_i}{A}(10{,}000)(100)$	*PD* > 0 (Number per 100 ha)
Edge density (*ED*)	FRAGSTAT-like	$ED = \frac{\sum_{k=1}^{m\prime} e_{ik}}{A}(10{,}000)$	*ED* ≥ 0, without limit (m ha^{-1})
Largest shape index (*LSI*)	FRAGSTAT-like	$LSI = \frac{\sum_{k=1}^{m} e''_{ik}}{2\sqrt{piA}}$	*LSI* ≥ 1, without limit (adimensional)

Mean patch size (*MPS*)	FRAGSTAT-like	$MPS = \frac{\sum_{j=1}^{n} \alpha_{ij}}{n_i} (\frac{1}{10{,}000})$	*MPS* ≥ 0, without limit (ha)
Patch size standard deviation (*PSSD*)	FRAGSTAT-like	$PSSD = \sqrt{\frac{\sum_{j=1}^{n} \left[\alpha_{ij} - \left(\frac{\sum_{j=1}^{n} \alpha_{ij}}{n_i} \right) \right]^2}{n_i}} (\frac{1}{10{,}000})$	*PSSD* ≥ 0, without limit (ha)
Patch size coefficient variation (*PSCOV*)	FRAGSTAT-like	$PSCOV = \frac{PSSD}{MPS}(100)$	*PSCOV* ≥ 0, without limit (percentage)
Mean shape index (*MSI*)	FRAGSTAT-like	$MSI = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{p_{ij}}{2\sqrt{\pi \alpha_{ij}}} \right)$	*MSI* ≥ 1, without limit (adimensional)
Area-weighted mean shape index (*AWMSI*)	FRAGSTAT-like	$AWMSI = \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\left(\frac{p_{ij}}{2\sqrt{\pi \alpha_{ij}}} \right) \left(\frac{\alpha_{ij}}{A} \right) \right]$	*AWMSI* ≥ 1, without limit (adimensional)
Mean patch fractal area dimension (*MPFD*)	FRAGSTAT-like	$MPFD = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{2 \ln p_{ij}}{\ln \alpha_{ij}} \right)}{n_i}$	1 ≤ *MPFD* ≤ 2 (adimensional)
Area-weighted patch fractal area dimension (*AWMPFD*)	FRAGSTAT-like	$AWMPFD = \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\left(\frac{2 \ln p_{ij}}{\ln a_{ij}} \right) \left(\frac{\alpha_{ij}}{A} \right) \right]$	1 ≤ AW*MPFD* ≤ 2 (adimensional)

Mean perimeter area ratio (*MPAR*)	FRAGSTAT-like	$MPAR=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{p_{ij}}{\alpha_{ij}}\right)}{n_i}$	*MPAR* ≥ 0, without limit (m m^{-2})
Fractal dimension (*D*)	Fractal-like	$\log N_k=-D\log\epsilon_K+c$	0 ≤ *D* ≤ 2 (adimensional)
Ratio Fractal dimension by proportion of the landscape occupied by patch type (class) (D_{Pi})	Fractal-like	$D_{P_i}=\frac{D}{P_i}$	*D* ≥ 0.2 (adimensional)

α_{ij} = area (m^2) of patch *j* of class *i*

e_{ik} = total length (m) of edge in landscape between patch types (classes) *i* and *k*; includes landscape boundary segments involving patch type

e''_{ik} = total length (m) of edge in landscape between patch types (classes) *i* and *k*; includes the entire landscape boundary and background edge segments, regardless of whether they represent the true edge

n_i (*n*) = number of patches in the landscape of patch type (class) *i*

m = number of patches types in the landscape, excluding the landscape border if present

m' = number of patches types in the landscape, including the landscape border if present

P_i = proportion of the landscape occupied by patch type (class) *i*;

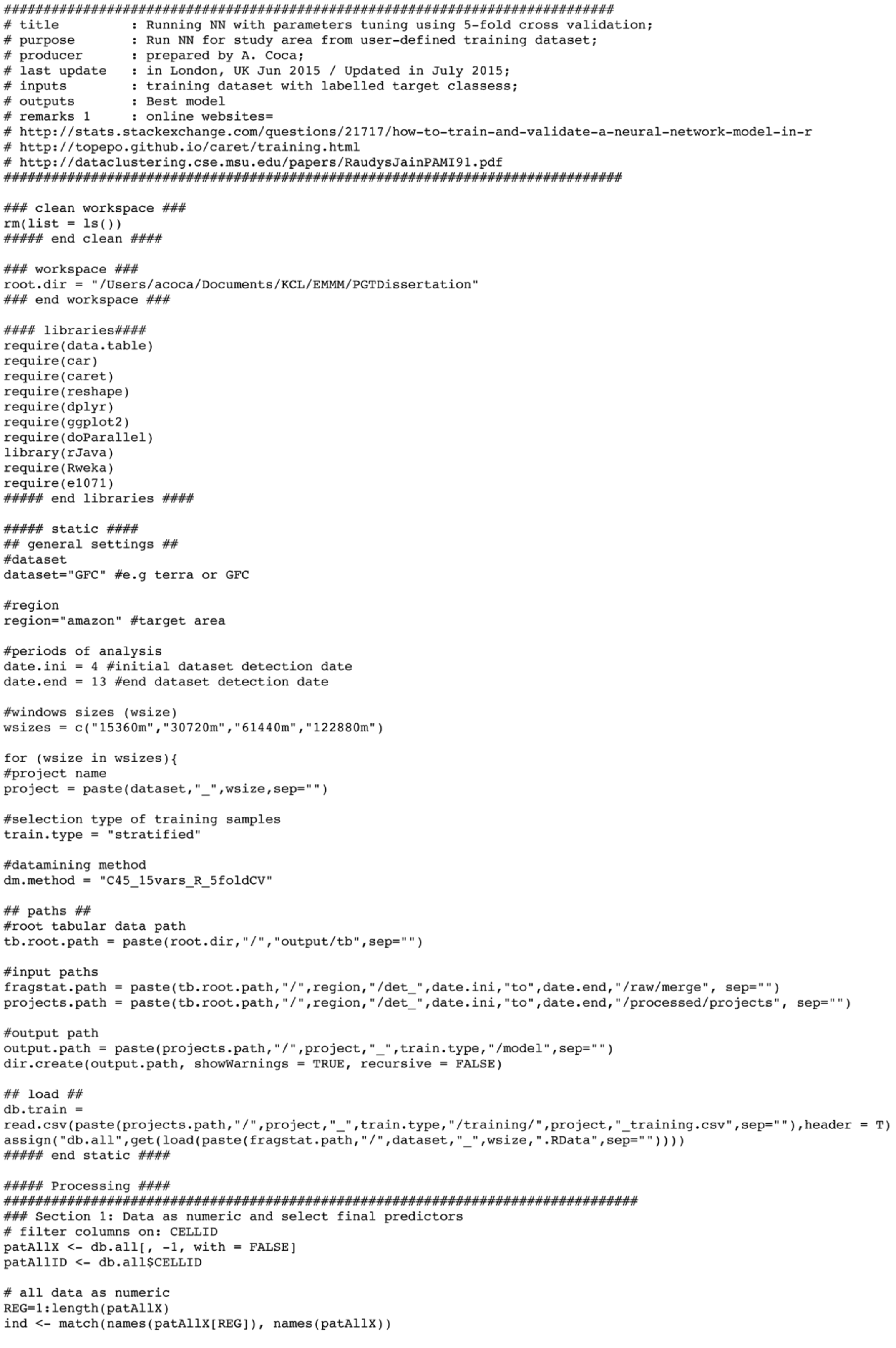
A = total landscape area (m^2)

N_k = number of boxes used in the box-counting technique to compute fractal dimension

ε_K = length (m) of the boxes used in the box-counting technique to compute fractal dimension

Appendix 4. Screenshots of the main R programming codes regarding to data mining modelling framework. The codes for the fractal analysis framework can be requested directly to Sun *et al*. (2014).

Part 1A: Code snippet for model building the C4.5 models



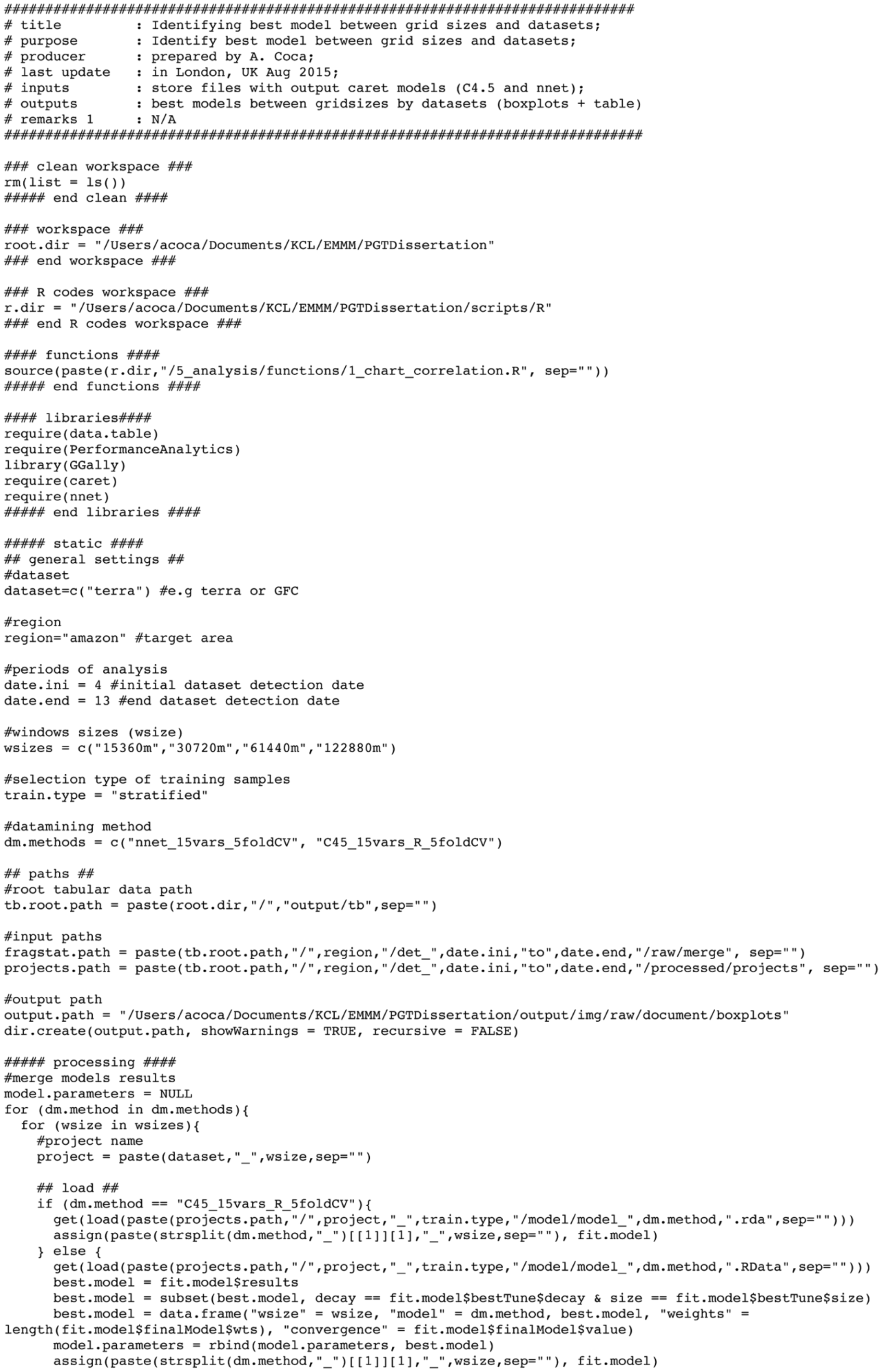
```r
#############################################################################
# title         : Running NN with parameters tuning using 5-fold cross validation;
# purpose       : Run NN for study area from user-defined training dataset;
# producer      : prepared by A. Coca;
# last update   : in London, UK Jun 2015 / Updated in July 2015;
# inputs        : training dataset with labelled target classess;
# outputs       : Best model
# remarks 1     : online websites=
# http://stats.stackexchange.com/questions/21717/how-to-train-and-validate-a-neural-network-model-in-r
# http://topepo.github.io/caret/training.html
# http://dataclustering.cse.msu.edu/papers/RaudysJainPAMI91.pdf
##############################################################################

### clean workspace ###
rm(list = ls())
##### end clean ####

### workspace ###
root.dir = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation"
### end workspace ###

#### libraries####
require(data.table)
require(car)
require(caret)
require(reshape)
require(dplyr)
require(ggplot2)
require(doParallel)
library(rJava)
require(Rweka)
require(e1071)
##### end libraries ####

##### static ####
## general settings ##
#dataset
dataset="GFC" #e.g terra or GFC

#region
region="amazon" #target area

#periods of analysis
date.ini = 4 #initial dataset detection date
date.end = 13 #end dataset detection date

#windows sizes (wsize)
wsizes = c("15360m","30720m","61440m","122880m")

for (wsize in wsizes){
#project name
project = paste(dataset,"_",wsize,sep="")

#selection type of training samples
train.type = "stratified"

#datamining method
dm.method = "C45_15vars_R_5foldCV"

## paths ##
#root tabular data path
tb.root.path = paste(root.dir,"/","output/tb",sep="")

#input paths
fragstat.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/raw/merge", sep="")
projects.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/projects", sep="")

#output path
output.path = paste(projects.path,"/",project,"_",train.type,"/model",sep="")
dir.create(output.path, showWarnings = TRUE, recursive = FALSE)

## load ##
db.train =
read.csv(paste(projects.path,"/",project,"_",train.type,"/training/",project,"_training.csv",sep=""),header = T)
assign("db.all",get(load(paste(fragstat.path,"/",dataset,"_",wsize,".RData",sep=""))))
##### end static ####

##### Processing ####
################################################################################
### Section 1: Data as numeric and select final predictors
# filter columns on: CELLID
patAllX <- db.all[, -1, with = FALSE]
patAllID <- db.all$CELLID

# all data as numeric
REG=1:length(patAllX)
ind <- match(names(patAllX[REG]), names(patAllX))
## Convert all variables (columns) to numeric format
for (i in seq_along(patAllX)) {
  set(patAllX, NULL, ind[i], as.numeric(as.character(patAllX[[ind[i]]])))
}

#variables to select
sel.var.INPE.geoDMA =
"^c_.CA.$|^c_.PLAND.$|^c_.PD.$|^c_.AREA_MN.$|^c_.AREA_CV.$|^c_.LSI.$|^c_.SHAPE_MN.$|^c_.SHAPE_AM.$|^c_.FRAC_MN.$|^

sel.var.fractal = "^f_FD$|^f_FDn$"
sel.var=paste(sel.var.INPE.geoDMA,sel.var.fractal,sep="|")

#filter using selected variables identified for all datasets
filter = grepl(sel.var, names(patAllX))
patAllX = patAllX[, filter, with = F]
patAllX[,"c_.AREA_SD."]= (patAllX$c_.AREA_MN.*patAllX$c_.AREA_MN.)/100

filter = grepl(sel.var, names(db.train))
patTrainX = db.train[, filter]
patTrainX[,"c_.AREA_SD."]= (patTrainX$c_.AREA_MN.*patTrainX$c_.AREA_MN.)/100

###############################################################################
### Section 2: Separate training and unlabelled dataset from final database
###join ID to separate unlabelled data
db.target = cbind(CELLID = patAllID,patAllX)
db.unlabelled=db.target[-which(db.target$CELLID %in% unique(db.train$CELLID)),]

#define input data (train and unlabelled) for model
db.train.model = data.frame("CELLID"=db.train$CELLID,"pattern"=db.train$pattern,patTrainX)
db.unlabelled.model = cbind("pattern"="null",db.unlabelled[,-1, with = F])

#training set without ID (CELLID)
vars.remove <- -grep('^(CELLID)', names(db.train.model))
model.train.input = db.train.model[,vars.remove]

###############################################################################
### Section 3: Model settings
#create a list of seed, here change the seed for each resampling
set.seed(40)
n.repeats = 40
n.resampling = 5
length.seeds = (n.repeats*n.resampling)+1
n.tune.parameters = 1
seeds <- vector(mode = "list", length = length.seeds)#length is = (n_repeats*nresampling)+1
for(i in 1:length.seeds) seeds[[i]]<- sample.int(n=1000, n.tune.parameters) #(n.tune.parameters = number of
tuning parameters)
seeds[[length.seeds]]<-sample.int(1000, 1)#for the last model

#create a control object for the models, implementing 10-crossvalidation repeated 10 times
fit.Control <- trainControl(
  method = "repeatedcv",
  number = n.resampling, ## 5-fold CV
  repeats = n.repeats, ## repeated ten times 100 iterations
  savePred = TRUE,
  seeds = seeds
  )

###############################################################################
### Section 4: Model run
## parallel process ##
#cluster
cl <- makeCluster(detectCores()-2)    #create a cluster
registerDoParallel(cl)                #register the cluster

set.seed(40)
## foreach or lapply would do this faster
fit.model <- train(pattern~.,
                           data=model.train.input,
                           trControl = fit.Control,
                           preProcess=c("range"),
                           method = "J48",
                           metric = "Kappa"
)

#export and store model results
.jcache(fit.model$finalModel$classifier)
save(fit.model, file=paste(output.path,"/model_",dm.method,".rda",sep=""))
cat(paste("wsize done = ", wsize, sep=""))
}
##### end processing ####
```

Part 1B: Code snippet for building the ANN models

```
##############################################################################
# title         : Running NN with parameters tuning using 5-fold cross validation;
# purpose       : Run NN for study area from user-defined training dataset;
# producer      : prepared by A. Coca;
# last update   : in London, UK Jun 2015 / Updated in July 2015;
# inputs        : training dataset with labelled target classess;
# outputs       : Best model
# remarks 1     : online websites=
# http://stats.stackexchange.com/questions/21717/how-to-train-and-validate-a-neural-network-model-in-r
# http://topepo.github.io/caret/training.html
# http://dataclustering.cse.msu.edu/papers/RaudysJainPAMI91.pdf
###############################################################################

### clean workspace ###
rm(list = ls())
##### end clean ####

### workspace ###
root.dir = "Q:/BACKUPS/Portatil_AlejandroCoca/temp/dissertation"
### end workspace ###

#### libraries####
require(data.table)
require(devtools)
require(car)
require(caret)
require(NeuralNetTools)
require(nnet)
require(reshape)
require(dplyr)
require(ggplot2)
require(doParallel)
##### end libraries ####

##### static ####
## general settings ##
#dataset
dataset="terra" #e.g terra or GFC

#region
region="amazon" #target area

#periods of analysis
date.ini = 4 #initial dataset detection date
date.end = 13 #end dataset detection date

#windows sizes (wsize)
wsizes = c("15360m","30720m","61440m","122880m")

for (wsize in wsizes){
#project name
project = paste(dataset,"_",wsize,sep="")

#selection type of training samples
train.type = "stratified"

#datamining method
dm.method = "nnet_15vars_5foldCV"

## paths ##
#root tabular data path
tb.root.path = paste(root.dir,"/","output/tb",sep="")

#input paths
fragstat.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/raw/merge", sep="")
projects.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/projects", sep="")

#output path
output.path = paste(projects.path,"/",project,"_",train.type,"/model",sep="")
dir.create(output.path, showWarnings = TRUE, recursive = FALSE)

## load ##
db.train =
read.csv(paste(projects.path,"/",project,"_",train.type,"/training/",project,"_training.csv",sep=""),header = T)
assign("db.all",get(load(paste(fragstat.path,"/",dataset,"_",wsize,".RData",sep=""))))
##### end static ####

##### Processing ####
################################################################################
### Section 1: Data as numeric and select final predictors
# filter columns on: CELLID
patAllX <- db.all[, -1, with = FALSE]
patAllID <- db.all$CELLID

# all data as numeric
REG=1:length(patAllX)
ind <- match(names(patAllX[REG]), names(patAllX))
```

```
## Convert all variables (columns) to numeric format
for (i in seq_along(patAllX)) {
  set(patAllX, NULL, ind[i], as.numeric(as.character(patAllX[[ind[i]]])))
}

#variables to select
sel.var.INPE.geoDMA =
"^c_.CA.$|^c_.PLAND.$|^c_.PD.$|^c_.AREA_MN.$|^c_.AREA_CV.$|^c_.LSI.$|^c_.SHAPE_MN.$|^c_.SHAPE_AM.$|^c_.FRAC_MN.$|^

sel.var.fractal = "^f_FD$|^f_FDn$"
sel.var=paste(sel.var.INPE.geoDMA,sel.var.fractal,sep="|")

#filter using selected variables identified for all datasets
filter = grepl(sel.var, names(patAllX))
patAllX = patAllX[, filter, with = F]
patAllX[,"c_.AREA_SD."]= (patAllX$c_.AREA_MN.*patAllX$c_.AREA_MN.)/100
#preprocessing input data before modelling
patAllX.m = preProcess(patAllX, c("range"))
patAllX.n = predict(patAllX.m,patAllX)

filter = grepl(sel.var, names(db.train))
patTrainX = db.train[, filter]
patTrainX[,"c_.AREA_SD."]= (patTrainX$c_.AREA_MN.*patTrainX$c_.AREA_MN.)/100
#preprocessing input data before modelling
patTrainX.m = preProcess(patTrainX, c("range"))
patTrainX.n = predict(patTrainX.m,patTrainX)

##############################################################################
### Section 2: Separate training and unlabelled dataset from final database
###join ID to separate unlabelled data
db.target = cbind(CELLID = patAllID,patAllX.n)
db.unlabelled=db.target[-which(db.target$CELLID %in% unique(db.train$CELLID)),]

#define input data (train and unlabelled) for model
db.train.model = data.frame("CELLID"=db.train$CELLID,"pattern"=db.train$pattern,patTrainX.n)
db.train.raw = data.frame("CELLID"=db.train$CELLID,"pattern"=db.train$pattern,patTrainX)
db.unlabelled.model = cbind("pattern"="null",db.unlabelled[,-1])

#training set without ID (CELLID)
vars.remove <- -grep('^(CELLID)', names(db.train.model))
model.train.input = db.train.model[,vars.remove]
model.train.raw = db.train.raw[,vars.remove]

##############################################################################
### Section 3: Model settings
#ANN parameters
decay.tune = c(0.0001, 0.001, 0.01, 0.1)
size = size = seq(1, 50,by=1)
maxit.nnet = 5000
rang.nnet = 0.7
MaxNWts.nnet = 2000

#tuning grid for train caret function
my.grid <- expand.grid(.decay = decay.tune, .size = size)

#create a list of seed, here change the seed for each resampling
set.seed(40)
n.repeats = 40
n.resampling = 5
length.seeds = (n.repeats*n.resampling)+1
n.tune.parameters = length(decay.tune)*length(size)
seeds <- vector(mode = "list", length = length.seeds)#length is = (n_repeats*nresampling)+1
for(i in 1:length.seeds) seeds[[i]]<- sample.int(n=1000, n.tune.parameters) #(n.tune.parameters = number of
tuning parameters)
seeds[[length.seeds]]<-sample.int(1000, 1)#for the last model

#create a control object for the models, implementing 10-crossvalidation repeated 10 times
fit.Control <- trainControl(
  method = "repeatedcv",
  number = n.resampling, ## 5-fold CV
  repeats = n.repeats, ## repeated ten times 100 iterations
  classProbs=TRUE,
  savePred = TRUE,
  seeds = seeds
  )

##############################################################################
### Section 4: Model run
## parallel process ##
#cluster
cl <- makeCluster(detectCores()-2)    #create a cluster
registerDoParallel(cl)                #register the cluster

set.seed(40)
## foreach or lapply would do this faster
fit.model <- train(pattern~.,
                          data=model.train.input,
                          trControl = fit.Control,
                           method = "nnet",
                           maxit = maxit.nnet,
                           rang = rang.nnet,
                           MaxNWts = MaxNWts.nnet,
                           tuneGrid = my.grid,
                           trace = F,
                           metric = "Kappa",
                           linout = F
)

#save and export model results
save(db.train.model, fit.model,
     db.unlabelled.model, db.unlabelled,
     maxit.nnet, rang.nnet, MaxNWts.nnet,
     file=paste(output.path,"/model_",dm.method,".RData",sep=""))
cat(paste("wsize done = ", wsize, sep=""))
}
##### end processing ####
```

Part 2: Code snippet for identifying best models



```
#############################################################################
# title          : Identifying best model between grid sizes and datasets;
# purpose        : Identify best model between grid sizes and datasets;
# producer       : prepared by A. Coca;
# last update    : in London, UK Aug 2015;
# inputs         : store files with output caret models (C4.5 and nnet);
# outputs        : best models between gridsizes by datasets (boxplots + table)
# remarks 1      : N/A
##############################################################################

### clean workspace ###
rm(list = ls())
##### end clean ####

### workspace ###
root.dir = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation"
### end workspace ###

### R codes workspace ###
r.dir = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation/scripts/R"
### end R codes workspace ###

#### functions ####
source(paste(r.dir,"/5_analysis/functions/1_chart_correlation.R", sep=""))
##### end functions ####

#### libraries####
require(data.table)
require(PerformanceAnalytics)
library(GGally)
require(caret)
require(nnet)
##### end libraries ####

##### static ####
## general settings ##
#dataset
dataset=c("terra") #e.g terra or GFC

#region
region="amazon" #target area

#periods of analysis
date.ini = 4 #initial dataset detection date
date.end = 13 #end dataset detection date

#windows sizes (wsize)
wsizes = c("15360m","30720m","61440m","122880m")

#selection type of training samples
train.type = "stratified"

#datamining method
dm.methods = c("nnet_15vars_5foldCV", "C45_15vars_R_5foldCV")

## paths ##
#root tabular data path
tb.root.path = paste(root.dir,"/","output/tb",sep="")

#input paths
fragstat.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/raw/merge", sep="")
projects.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/projects", sep="")

#output path
output.path = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation/output/img/raw/document/boxplots"
dir.create(output.path, showWarnings = TRUE, recursive = FALSE)

##### processing ####
#merge models results
model.parameters = NULL
for (dm.method in dm.methods){
  for (wsize in wsizes){
    #project name
    project = paste(dataset,"_",wsize,sep="")

    ## load ##
    if (dm.method == "C45_15vars_R_5foldCV"){
      get(load(paste(projects.path,"/",project,"_",train.type,"/model/model_",dm.method,".rda",sep="")))
      assign(paste(strsplit(dm.method,"_")[[1]][1],"_",wsize,sep=""), fit.model)
    } else {
      get(load(paste(projects.path,"/",project,"_",train.type,"/model/model_",dm.method,".RData",sep="")))
      best.model = fit.model$results
      best.model = subset(best.model, decay == fit.model$bestTune$decay & size == fit.model$bestTune$size)
      best.model = data.frame("wsize" = wsize, "model" = dm.method, best.model, "weights" =
length(fit.model$finalModel$wts), "convergence" = fit.model$finalModel$value)
      model.parameters = rbind(model.parameters, best.model)
      assign(paste(strsplit(dm.method,"_")[[1]][1],"_",wsize,sep=""), fit.model)
  }
}
}
#export models best combination parameters and features
write.csv(model.parameters,paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/analysis/",dat
 row.names=T)

#resampling for model comparison
resamps <- resamples(list(C45_122880m = C45_122880m,
                          C45_61440m = C45_61440m,
                          C45_30720m = C45_30720m,
                          C45_15360m = C45_15360m,
                          ANN_122880m = nnet_122880m,
                          ANN_61440m = nnet_61440m,
                          ANN_30720m = nnet_30720m,
                          ANN_15360m = nnet_15360m))

models.comparison = data.frame(summary(resamps)$statistics)
#export models comparison (boxplot)
write.csv(models.comparison,paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/analysis/",da
 row.names=T)

#boxplots charts
trellis.par.set(caretTheme())
bwplot(resamps, layout = c(2, 1), box.ratio = 1, auto.key = T)

#determining statistical diffennce   s
statistics.models = NULL
for (wsize in wsizes){
  C45.model = get(paste("C45_",wsize,sep=""))
  ANN.model = get(paste("nnet_",wsize,sep=""))
  sta.results = compare_models(C45.model, ANN.model, metric = C45.model$metric[1])
  sta.results = data.frame(wsize = wsize, p_value = sta.results$p.value)
  statistics.models = rbind(statistics.models, sta.results)
}
#export models statistics (p-values)
write.csv(statistics.models,paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/analysis/",da
 row.names=T)
##### end processing ####
```

Part 3: Code snippet for the sensitivity analysis

```
#######################################################################################
# title         : Sensitivity analyses for the best models between gridsizes by dataset;
# purpose       : Identify variable importance in the best models;
# producer      : prepared by A. Coca;
# last update   : in London, UK Aug 2015;
# inputs        : best models files;
# outputs       : barplots by pattern with variable importance according to Olden et
#                 al (2004)
# remarks 1     : N/A
#############################################################################

### clean workspace ###
rm(list = ls())
##### end clean ####

### workspace ###
root.dir = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation/"
### end workspace ###

#### libraries####
require(data.table)
require(devtools)
require(caret)
require(NeuralNetTools)
require(nnet)
require(reshape)
require(dplyr)
require(ggplot2)
require(gridExtra)
require(grid)
##### end libraries ####

##### static ####
## general settings ##
#dataset
dataset="GFC" #e.g terra or GFC

#region
region="amazon" #target area

#periods of analysis
date.ini = 4 #initial dataset detection date
date.end = 13 #end dataset detection date

#windows sizes (wsize)
wsize = "30720m"

#project name
project = paste(dataset,"_",wsize,sep="")

#selection type of training samples
train.type = "stratified"

#datamining method
dm.method = "nnet_15vars_5foldCV"

## paths ##
#root tabular data path
tb.root.path = paste(root.dir,"/","output/tb",sep="")

#input paths
projects.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/projects",
sep="")

#output path
output.path = paste(projects.path,"/",project,"_",train.type,"/results",sep="")
dir.create(output.path, showWarnings = TRUE, recursive = FALSE)

## load ##
get(load(paste(projects.path,"/",project,"_",train.type,"/model/model_",dm.method,".RData",sep="")))
##### end static ####

##### processing ####
#rename variables for plotting
vars.target <- -grep('^(CELLID|pattern)', names(db.train.model))
colnames(db.train.model)[vars.target] =
c("CA","PLAND","PD","ED","LSI","MPS","PSCOV","MSI","AWMSI","MPFD","AWMPFD","MPAR","D","D_PLAND","PSSD")
```

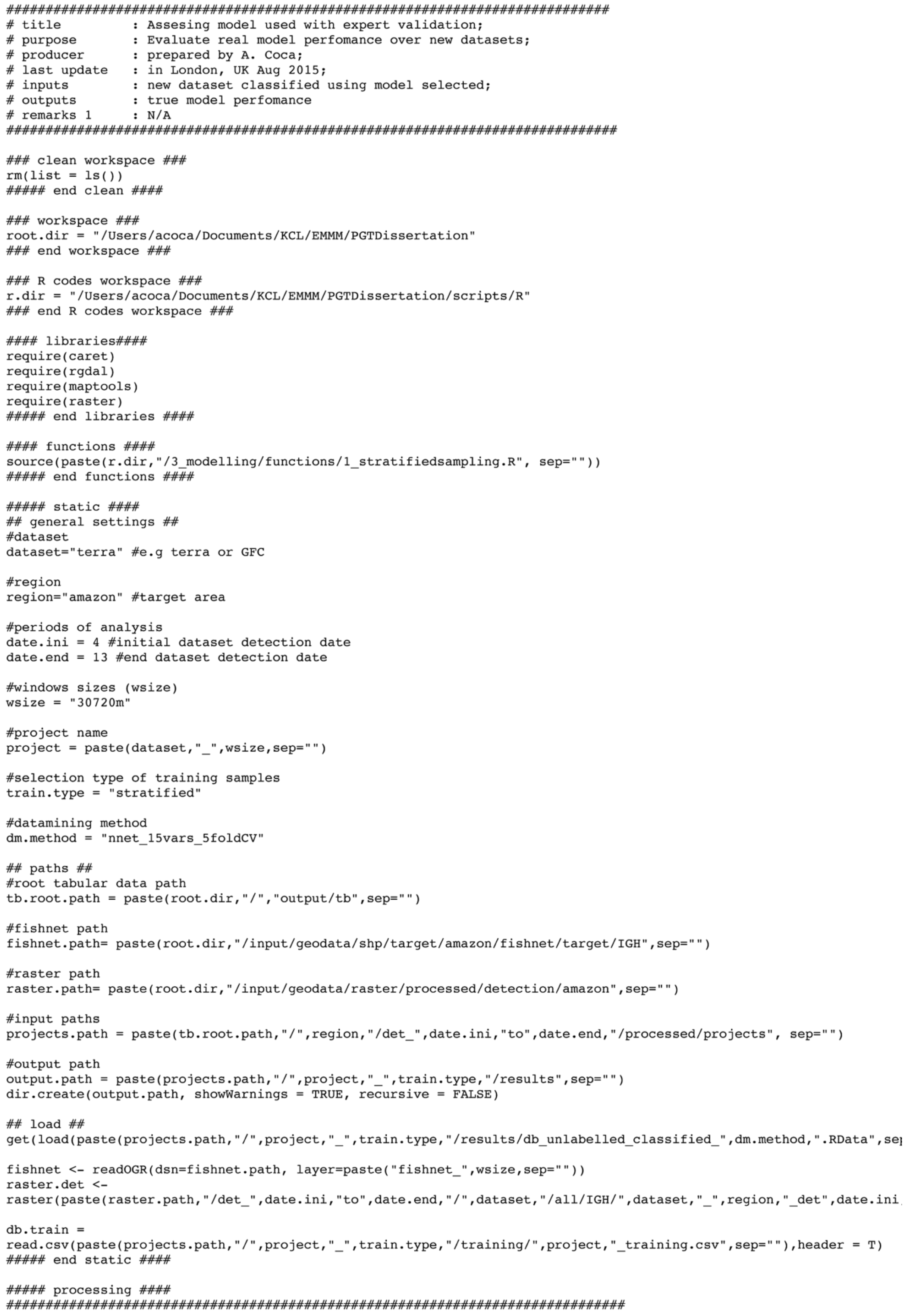
```
#extract best ANN model parameters
wts <- fit.model$finalModel$wts
decay <- fit.model$finalModel$decay
struct <- fit.model$finalModel$n

# recreate
set.seed(40) #seed set for reproducing the same results
recmod <- nnet(pattern~.,
               data=db.train.model[,-1],
               Wts = wts, decay = decay,
               size = struct[2],
               maxit = maxit.nnet,
               rang = rang.nnet,
               MaxNWts = MaxNWts.nnet, linout = F,
               trave = T)

##plotting
#theme
bar.theme_nolegend = theme_classic(base_size = 12, base_family = "Cambria") + theme(plot.title =
element_text(face="bold", colour="#000000", size=14, vjust=1.3), axis.title.x =
element_text(face="plain", colour="#000000", size=15, vjust=-0.3),

axis.text.x  = element_text(angle=90, vjust=0.5, size=14), axis.title.y = element_text(face="plain",
colour="#000000", size=15, vjust=0.8), axis.text.y  = element_text(angle=0, vjust=0.5, size=12)) +
  theme(plot.background = element_blank(),panel.grid.major = element_blank(),panel.grid.minor =
element_blank(),panel.border = element_blank()) +
  theme(axis.line = element_line(color = 'black')) + theme(legend.position = "none")

#loop to create individual plots
for (pattern in levels(db.train.model$pattern)){
  plot.temp = olden(recmod, pattern) + bar.theme_nolegend + theme(legend.position = 'none')
  plot.temp = plot.temp +
    geom_bar(colour="black", fill="gray", stat="identity") +
    scale_y_continuous(breaks = seq(round(min(plot.temp$data$importance)-1),
round(max(plot.temp$data$importance)+1), by = abs(round((round(min(plot.temp$data$importance)-1) -
round(max(plot.temp$data$importance)+1))/5)))) +
    coord_cartesian(ylim=c(round(min(plot.temp$data$importance)-1),
round(max(plot.temp$data$importance)+1)))

  #save plot
  assign(paste("g_",pattern,sep=""), ggplotGrob(plot.temp))
  rm(plot.temp)
}

grid.arrange(g_de, g_di, g_g, g_mo, ncol=1)
##### end processing ####
```

Part 4: Code snippet for the model use step

```
#######################################################################################
# title         : Model use and generation GIS layer for inspection of model true perfomance;
# purpose       : Implement best model to classify unlabelled data and join in a GIS layer (fishnet);
# producer      : prepared by A. Coca;
# last update   : in London, UK Aug 2015;
# inputs        : best models files;
# outputs       : GIS layer (fishnet) with grid-based object classified
# remarks 1     : N/A
#############################################################################

### clean workspace ###
rm(list = ls())
##### end clean ####

### workspace ###
root.dir = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation/"
### end workspace ###

#### libraries####
require(caret)
require(rgdal)
require(maptools)
##### end libraries ####

##### static ####
## general settings ##
#dataset
dataset="GFC" #e.g terra or GFC

#region
region="amazon" #target area

#periods of analysis
date.ini = 4 #initial dataset detection date
date.end = 13 #end dataset detection date

#windows sizes (wsize)
wsize = "30720m"

#project name
project = paste(dataset,"_",wsize,sep="")

#selection type of training samples
train.type = "stratified"

#datamining method
dm.method = "nnet_15vars_5foldCV"

## paths ##
#root tabular data path
tb.root.path = paste(root.dir,"/","output/tb",sep="")

#fishnet path
fishnet.path= paste(root.dir,"/input/geodata/shp/target/amazon/fishnet/target/IGH",sep="")

#input paths
projects.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/projects",
sep="")

#output path
output.path = paste(projects.path,"/",project,"_",train.type,"/results",sep="")
dir.create(output.path, showWarnings = TRUE, recursive = FALSE)

#output root shp dir
outgeo.path = paste(root.dir,"/output/geodata",sep="")

#output project shape path
outshp.project =
paste(outgeo.path,"/",region,"/det_",date.ini,"to",date.end,"/",dataset,"/shp/classification",sep="")
dir.create(outshp.project, showWarnings = TRUE, recursive = FALSE)

## load ##
get(load(paste(projects.path,"/",project,"_",train.type,"/model/model_",dm.method,".RData",sep="")))
fishnet <- readOGR(dsn=fishnet.path, layer=paste("fishnet_",wsize,sep=""))
##### end static ####

##### processing ####
```

```
###############################################################################
### Section 1: Model predictions over unlabelled set
#prediction unlabelled dataset to score
pattern.unlabelled.pred.raw <- predict(fit.model, newdata=db.unlabelled.model, type="raw")

################################################################################
### Section 2: generate classified database for visualisation and inspection
db.training.classified=cbind("CELLID" = db.train.model$CELLID, as.character(db.train.model$pattern))
db.unlabelled.classified=cbind("CELLID" = as.character(db.unlabelled$CELLID),
as.character(pattern.unlabelled.pred.raw))
db.classified = data.frame(rbind(db.training.classified,db.unlabelled.classified))
colnames(db.classified)=c("CELLID","pattern")

# ################################################################################
### Section 3: Join classified database with fishnet layer for visualisation#match
id_m = match(as.character(fishnet$CELLID), as.character(db.classified$CELLID))
xtra1 = db.classified[id_m,]

#spatial cbind
fishnet.merge=spCbind(fishnet,xtra1$pattern)

# ################################################################################
### Section 4: Plot best model
## Best model ##
#line chart
trellis.par.set(caretTheme())
plot(fit.model, metric = "Kappa")

##### export ####
writeSpatialShape(fishnet.merge,paste(outshp.project,"/",dm.method,"_",train.type,"_",project,sep=""))


save(db.unlabelled.classified
     file=paste(output.path,"/db_unlabelled_classified_",dm.method,".RData",sep=""))
##### end processing ####
```

Part 5: Code snippet for assessing the true models performance



```
#################################################################################
# title         : Assesing model used with expert validation;
# purpose       : Evaluate real model perfomance over new datasets;
# producer      : prepared by A. Coca;
# last update   : in London, UK Aug 2015;
# inputs        : new dataset classified using model selected;
# outputs       : true model perfomance
# remarks 1     : N/A
##################################################################################

### clean workspace ###
rm(list = ls())
##### end clean ####

### workspace ###
root.dir = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation"
### end workspace ###

### R codes workspace ###
r.dir = "/Users/acoca/Documents/KCL/EMMM/PGTDissertation/scripts/R"
### end R codes workspace ###

#### libraries####
require(caret)
require(rgdal)
require(maptools)
require(raster)
##### end libraries ####

#### functions ####
source(paste(r.dir,"/3_modelling/functions/1_stratifiedsampling.R", sep=""))
##### end functions ####

##### static ####
## general settings ##
#dataset
dataset="terra" #e.g terra or GFC

#region
region="amazon" #target area

#periods of analysis
date.ini = 4 #initial dataset detection date
date.end = 13 #end dataset detection date

#windows sizes (wsize)
wsize = "30720m"

#project name
project = paste(dataset,"_",wsize,sep="")

#selection type of training samples
train.type = "stratified"

#datamining method
dm.method = "nnet_15vars_5foldCV"

## paths ##
#root tabular data path
tb.root.path = paste(root.dir,"/","output/tb",sep="")

#fishnet path
fishnet.path= paste(root.dir,"/input/geodata/shp/target/amazon/fishnet/target/IGH",sep="")

#raster path
raster.path= paste(root.dir,"/input/geodata/raster/processed/detection/amazon",sep="")

#input paths
projects.path = paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/projects", sep="")

#output path
output.path = paste(projects.path,"/",project,"_",train.type,"/results",sep="")
dir.create(output.path, showWarnings = TRUE, recursive = FALSE)

## load ##
get(load(paste(projects.path,"/",project,"_",train.type,"/results/db_unlabelled_classified_",dm.method,".RData",sep

fishnet <- readOGR(dsn=fishnet.path, layer=paste("fishnet_",wsize,sep=""))
raster.det <-
raster(paste(raster.path,"/det_",date.ini,"to",date.end,"/",dataset,"/all/IGH/",dataset,"_",region,"_det",date.ini

db.train =
read.csv(paste(projects.path,"/",project,"_",train.type,"/training/",project,"_training.csv",sep=""),header = T)
##### end static ####

##### processing ####
###################################################################################
```

```
### Section 1: Create random sampling from classified objects
db.unlabelled.classified=data.frame(db.unlabelled.classified)
colnames(db.unlabelled.classified)[2] = "pred"

#create repetations and set.seed to replicate
n.repetitions = 3
n.samples = 30
set.seed(40)
seeds <- vector(mode = "list", length = n.repetitions)
for(i in 1:n.repetitions) seeds[[i]]<- sample.int(n=1000, 1)

#set patterns levels
u = as.character(unique(db.unlabelled.classified$pred))

db.validation.all=NULL
for (s in 1:length(seeds)){
  cat(paste("\n ###### repetition no. ",s," ######",sep=""))
  set.seed(as.numeric(seeds[s]))
  sample.tmp = stratified(db.unlabelled.classified,"pred",n.samples)
  sample.shp <- fishnet[fishnet$CELLID %in% sample.tmp$CELLID,]
  db.validation.tmp=db.unlabelled.classified[which(db.unlabelled.classified$CELLID %in%
unique(sample.shp$CELLID)),]
  for (i in 1:dim(sample.shp)[1]){
    db=data.frame(sample.shp[i,])
    # CROP
    tryPoly <- sample.shp@polygons[[i]]@Polygons[[1]]@coords
    MaxY <- max(tryPoly[,2])
    MaxX <- max(tryPoly[,1])
    MinY <- min(tryPoly[,2])
    MinX <-min(tryPoly[,1])
    ext <- extent(cbind(c(MinX,MinY), c(MaxX,MaxY)))

    rawdet.masked <- intersect(raster.det,ext)

    breakpoints <- c(0,1)
    colors <- c("black","white")
    plot(rawdet.masked,breaks=breakpoints,col=colors)

    x <- readline("What is the spatial pattern? ")
    db.validation.tmp[db.validation.tmp$CELLID == db$CELLID,"expert"] = x
    pred.value = as.character(db.validation.tmp[db.validation.tmp$CELLID == db$CELLID,"pred"])
    exp.value = db.validation.tmp[db.validation.tmp$CELLID == db$CELLID,"expert"]
    cat(paste("SAMPLE ",i," out of ",dim(sample.shp)[1], "\n pred: ", pred.value," vs ", "exp: ", exp.value,
sep=""))
  }
  db.validation.tmp[,"rep"] = s
  db.validation.all = rbind(db.validation.all,db.validation.tmp)
}

save(db.validation.all, file=paste(output.path,"/","assessment_modeluse_",dm.method,".rda",sep=""))

#compute Kappa by repetition
metrics.all = NULL
cm.all = NULL
for (r in unique(db.validation.all$rep)){
  kappa.db = subset(db.validation.all, rep == r)
  predicted = kappa.db$pred
  reference = kappa.db$expert
  t = table(factor(predicted, u), factor(reference, u)) #in case a category is missing
  table.cm = confusionMatrix(t)$table
  cm.all = cbind(cm.all,table.cm)
  metrics.tmp = data.frame("rep" = r, t(confusionMatrix(t)$overall[1:2]))
  metrics.all = rbind(metrics.all,metrics.tmp)
}

#export model use evaluation metrics (overall accuracy and kappa)
write.csv(metrics.all,paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/analysis/",dataset,
 row.names=T)
write.csv(cm.all,paste(tb.root.path,"/",region,"/det_",date.ini,"to",date.end,"/processed/analysis/",dataset,"_bes
 row.names=T)
##### end processing ####
```

Appendix 5. Exploratory analysis performed over inputs used in data mining.

The class area (CA) and landscape percentage (PLAND) metrics showed similar behaviours (Figure 1), with different median values for all pattern typologies within each grid size. The multidirectional pattern, besides having the largest median value, also showed the closest median values between datasets.

Although part of the same category, median values among pattern typologies for the patch density (PD) metric were not notably different, except for the diffuse extensive pattern. Median values were similar between the diffuse extensive and multi directional pattern training sets. Additionally, the statistical dispersion between the upper and lower quartiles (interquartile range or IQR) was larger for the diffuse extensive pattern, particularly for the GFC training sets. In regard to the extent of the grid, this metric maintained, in general, a constant median among all grid sizes. The Terra-i training sets presented slightly lower values than GFC only for the multi directional and diffuse extensive training sets.

For the edge density (ED), there were marked differences in median for contrasting pattern typologies such as multi directional and diffuse extensive for both datasets. In contrast, the geometric and diffuse intensive pattern training sets presented closer median values. This behaviour was more predominant in the Terra-i training sets, where diffuse intensive patterns presented the largest IQR in comparison with the remaining pattern typologies within the same dataset. In terms of grid size, median values of this metric only increased or decreased for multi directional and geometric pattern training sets, respectively, when these patterns were analysed from coarser to finer grid sizes.

The pair of fractal-like metrics, the fractal dimension (D) and ratio of fractal dimension by non-forest proportion (D-PLAND), were added into the area category due to the association of the former with the spatial filling of a target class (non-forest) in a landscape. Although the D metric showed a similar behaviour to CA and PLAND, the IQR of each training set was lower in comparison with the FRAGSTAT-like metrics. Regardless of grid size the D metric was almost uniform with few median fluctuations in the multi directional pattern. The Terra-i training sets showed higher overall medians than GFC. The behaviour of the D-PLAND metric was the opposite of the previously discussed metrics; the largest values were held by the diffuse extensive training sets for all grid sizes. For the same pattern, this metric showed a larger IQR in comparison with the remaining pattern typologies.

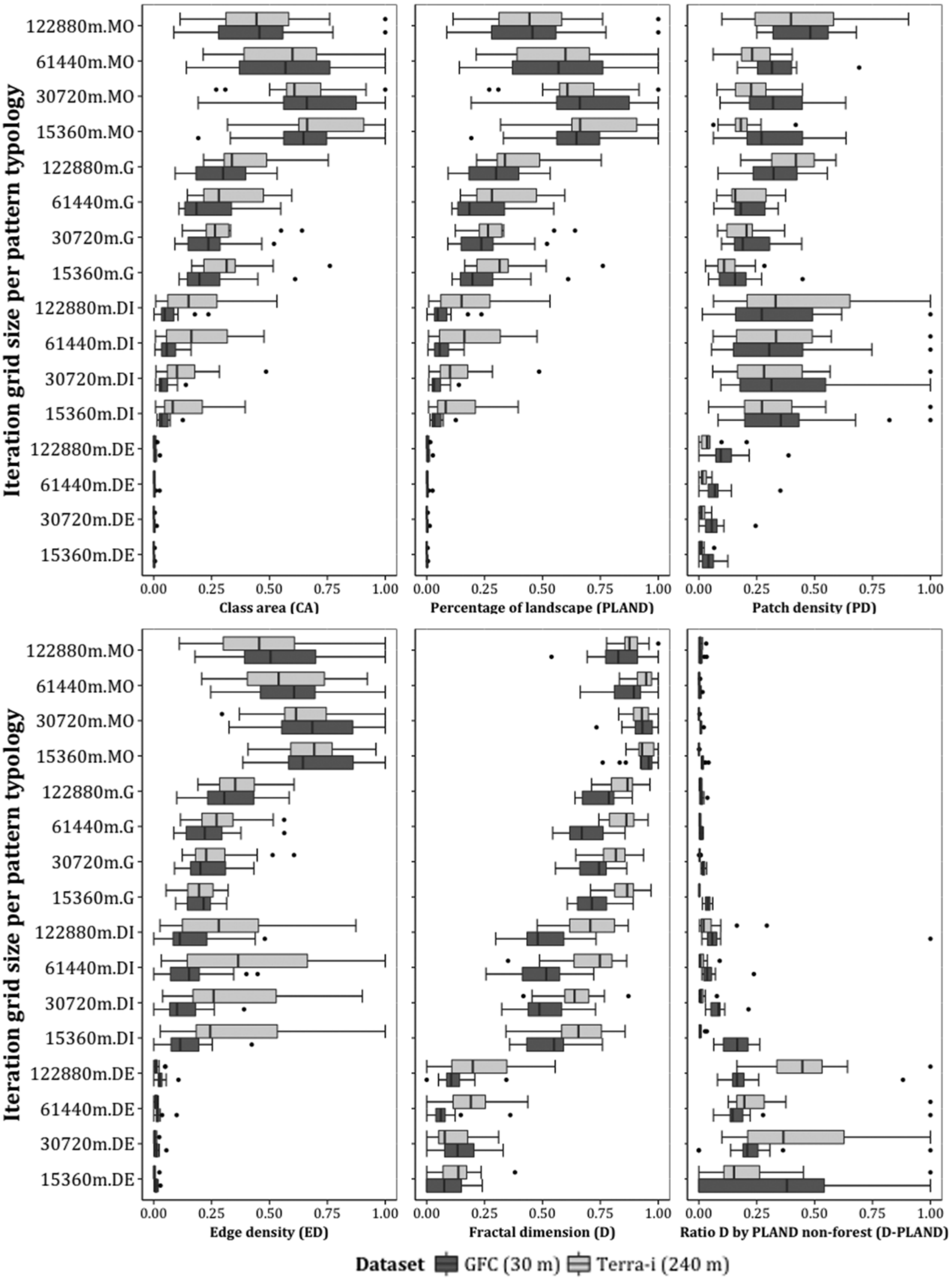


Figure 1. Boxplots showing the distribution of six normalised (0 to 1) input variables (CA, PLAND, PD, ED, LSI and MPAR) in relation to the area/density/edge conceptual category extracted from non-forest class grid-aggregated objects by grid size, pattern typology and dataset. Multidirectional, geometric, diffuse intensive and extensive patterns are denoted at the Y-axis as MO, G, DI, DE, respectively, after grid size.

The behaviours of the other six metrics, which included area-related and shape-related metrics, are illustrated in Figure 2. For the former group, patch size statistics such as the media and dispersion are described. In the case of the mean patch size (MPS) metric, the largest median values were in general given by the multidirectional pattern training sets. As grid size became finer, lower values of this metric were obtained. In addition, there was not a dominant trend of higher or lower MPS values between datasets. Regarding the standard deviation (PSSD) and coefficient of variation (PSCOV), calculated as part of the patch size dispersion statistics, these agreed that multidirectional and geometric pattern training sets were more heterogeneous due to their large variability as compared to the diffuse-related pattern types. These differences in heterogeneity tended to be more marked for PSCOV, which had a uniform median among all grid sizes.

The median of the mean shape index (MSI) metric, which reflects the shape complexity (vertices), was higher for the multidirectional training sets than for the remaining patterns. This behaviour was consistent for both datasets, except for the diffuse intensive pattern from the Terra-i training sets, which had medians closer to or greater than geometric pattern medians within the same dataset. In terms of grid size, MSI was sensitive to the extent of analysis unit for both datasets, except for the diffuse extensive pattern from the GFC training sets which was uniformly distributed.

The area-weighted mean shape index (AWMSI), which, like MSI, also reveals the shape complexity but assigns heavier weighting to large patches, confirmed that the multidirectional pattern training sets were more complex than the other patterns. Although the same order of lower values as the MSI metric was maintained for the remaining patterns, their variability was less marked than for MSI. Regarding grid sizes, AWMSI was in general constant, with a slight trend of higher values in the coarsest size. Although median values of the GFC training sets were higher for the multidirectional pattern than the Terra-i medians, the opposite was true for the remaining grid sizes.

For the landscape shape index (LSI), this presented slight differences between patterns and datasets from multi direction to diffuse intensive, except for the diffuse intensive training sets. The median statistic was not altered by changes in the grid size for the multi directional and diffuse intensive patterns for either datasets. In contrast, this metric seemed to be affected by the extent of analysis unit when considering the geometric and diffuse intensive patterns, especially for the GFC dataset.

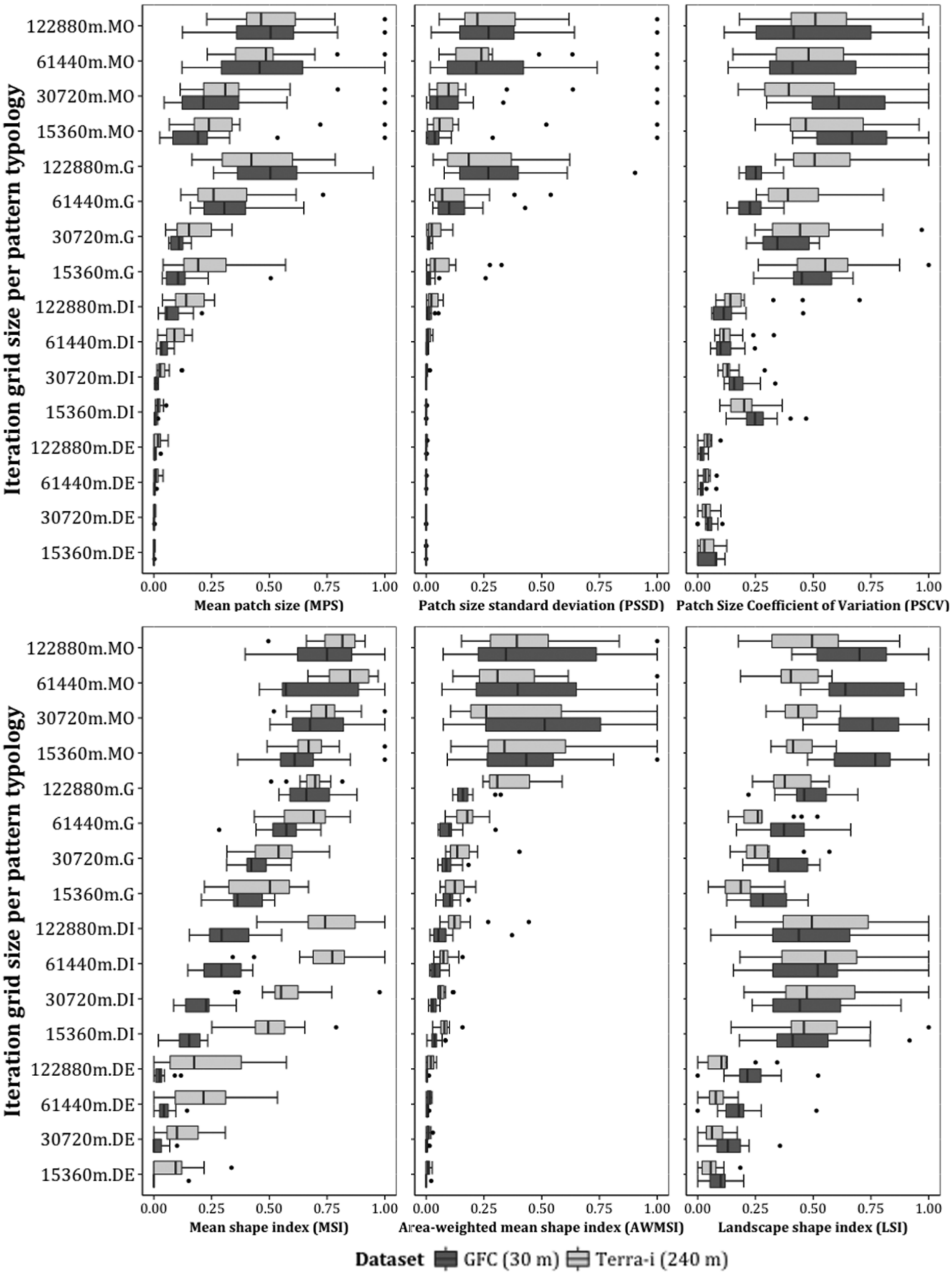


Figure 2. Boxplots showing the distribution of normalised (0 to 1) area-related (MPS, PSSD and PSCV) and shape-related (MSI, AWMSI and LSI) metrics extracted from non-forest class grid-aggregated objects by grid size by pattern typology by dataset. Multidirectional, geometric, diffuse intensive and extensive patterns are denoted at the Y-axis as MO, G, DI, DE, respectively after grid size.

The behaviour of a last set of 3 metrics, grouped together due to their calculation using perimeter and area relations, is explored in Figure 3. Overall, there was not a clear effect of grid size for these metrics, but there were marked behaviours within patterns and datasets. Median values of mean patch fractal area dimension (MPFD) were different between patterns and varied by dataset. For the Terra-i training sets, medians from geometric pattern were higher than the remaining patterns. On the other hand, for GFC training sets there was a gradual decrease from the multi directional to diffuse extensive pattern. Like the MPFD, the area-weighted patch fractal area dimension (AWMPFD) presented a gradual decrease, however training sets had less IQR than MPFD, particularly for the diffuse extensive patterns. Finally, the mean perimeter-area ratio (MPAR) showed differences and similarities between patterns by dataset. For Terra-i training sets, median values at all grid sizes for multi directional and diffuse intensive patterns were closer than for the remaining patterns. In contrast, for GFC training sets there was a gradual increase from the multi directional to diffuse extensive patterns.

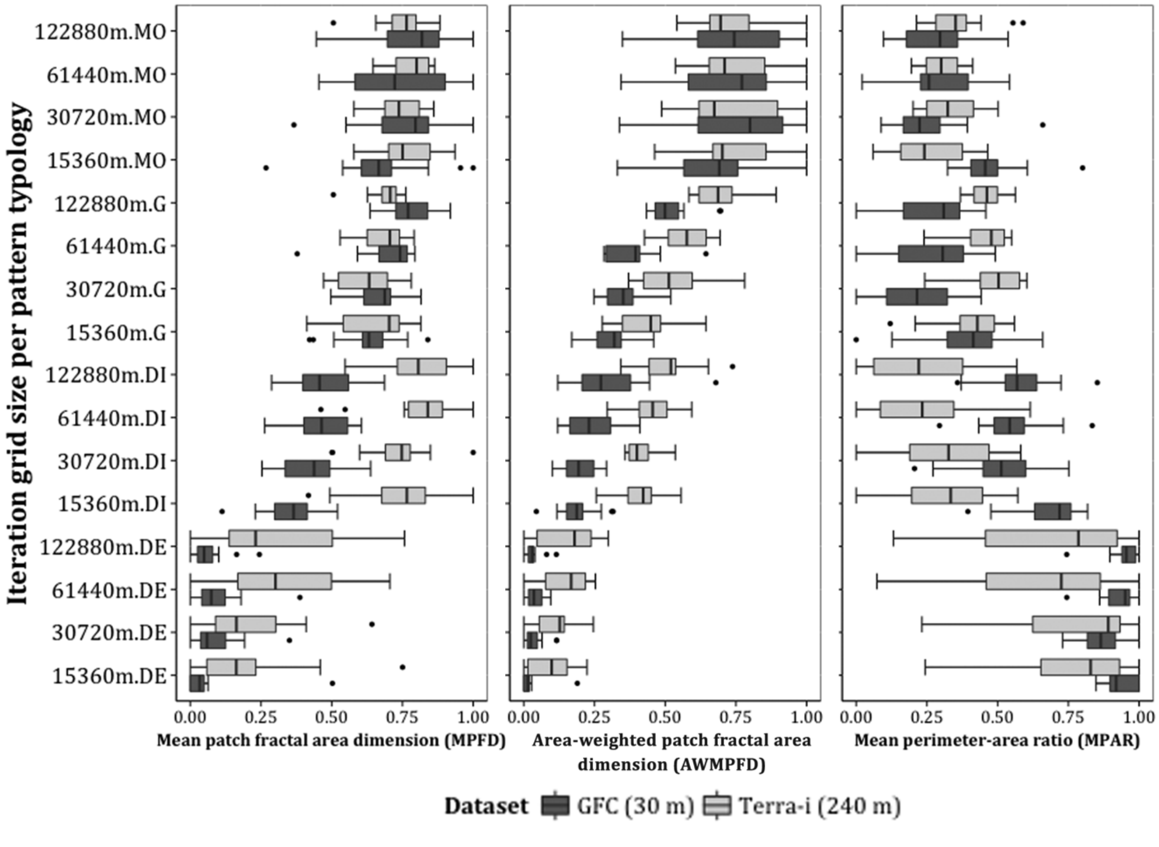


Figure 3. Boxplots showing the distribution of normalised (0 to 1) perimeter-area related (MPFD, AWMPFD, MPAR) metrics extracted from non-forest class grid-aggregated objects by grid size, pattern typology and dataset. Multidirectional, geometric, diffuse intensive and extensive patterns are denoted at the Y-axis as MO, G, DI, DE, respectively, after grid size.

Appendix 6. Confusion matrices for Terra-i (upper) and GFC (bottom) ANN models evaluations. Highlighted numbers indicate grid-based object agreement between predictions and expert observations.

TERRA-I

Pattern	*Repetition 1*				*Repetition 2*				*Repetition 3*			
	de	*di*	*g*	*m*	*de*	*di*	*g*	*m*	*de*	*di*	*g*	*m*
Diffuse extensive (*de*)	27	2	1	0	28	2	0	0	24	6	0	0
Diffuse intensive (*di*)	2	18	10	0	5	18	5	2	4	18	7	1
Geometric (*g*)	1	0	20	9	0	2	20	8	0	0	23	7
Multidirectional (*m*)	0	0	1	29	0	0	3	27	0	1	3	26
Total number (n)	30	20	32	38	33	22	28	37	28	25	33	34

GFC

Pattern	*Repetition 1*				*Repetition 2*				*Repetition 3*			
	de	*di*	*g*	*m*	*de*	*di*	*g*	*m*	*de*	*di*	*g*	*m*
Diffuse extensive (*de*)	30	0	0	0	26	4	0	0	25	5	0	0
Diffuse intensive (*di*)	2	24	1	3	5	20	3	2	4	22	1	3
Geometric (*g*)	2	5	22	1	0	2	23	5	0	3	26	1
Multidirectional (*m*)	0	1	6	23	0	1	8	21	0	0	4	26
Total number (n)	34	30	29	27	31	27	34	28	29	30	31	30