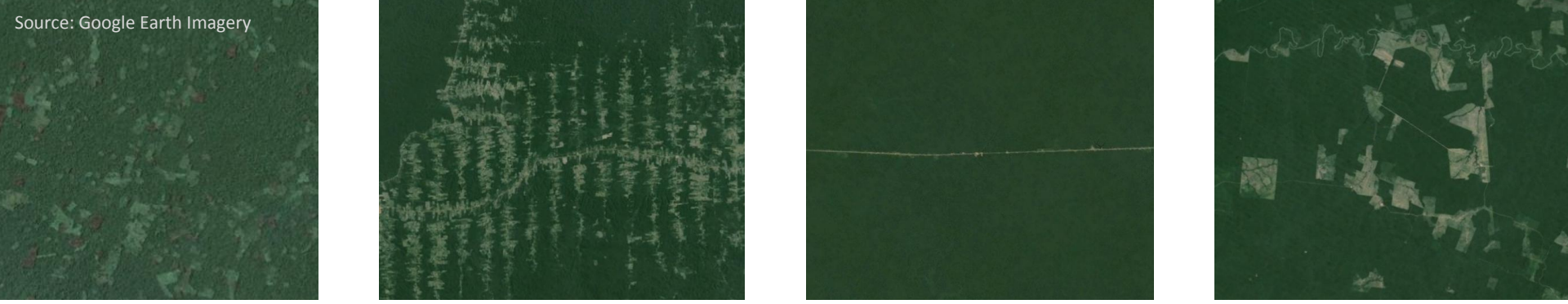


"Mapping spatial patterns of deforested areas monitored by Terra-i and GFC datasets" (Theory)

Alejandro Coca-Castro

PhD(c) in Geography

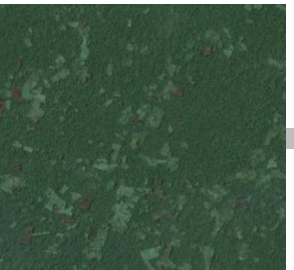
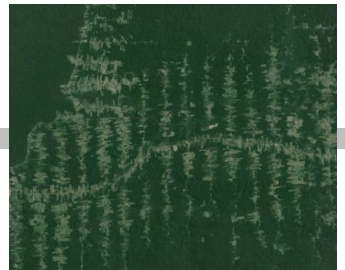
MSc in Environmental Modelling, Monitoring and Management

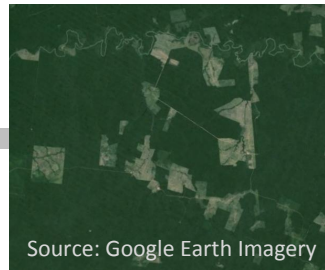
Research supervised by Mark Mulligan, Reader KCL Geography

KING'S College LONDON
Geography

Outline

- State of the art
  - The fractal analysis-based approach
  - The mining land-use patterns-based approach
- Research aim and objectives
- Methods
- Results
- Conclusion
- Limits and future research


Source: Google Earth Imagery

State of the art

There are multiple set of techniques to characterise spatial patterns, sizes and structures of deforested areas:

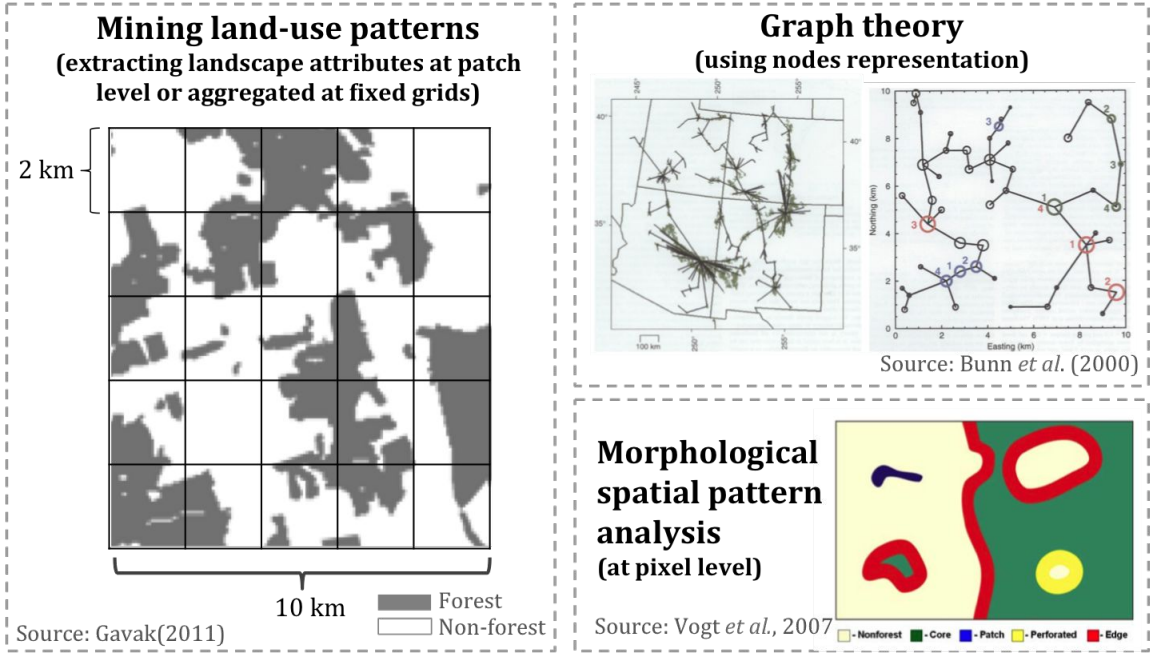


Figure 1. Some examples of methodologies for analysing spatial patterns and elements of deforested areas. Refer to the sources provided for further information on these methodologies.

The fractal analysis-based approach as a screening and complementary method to study deforestation stages

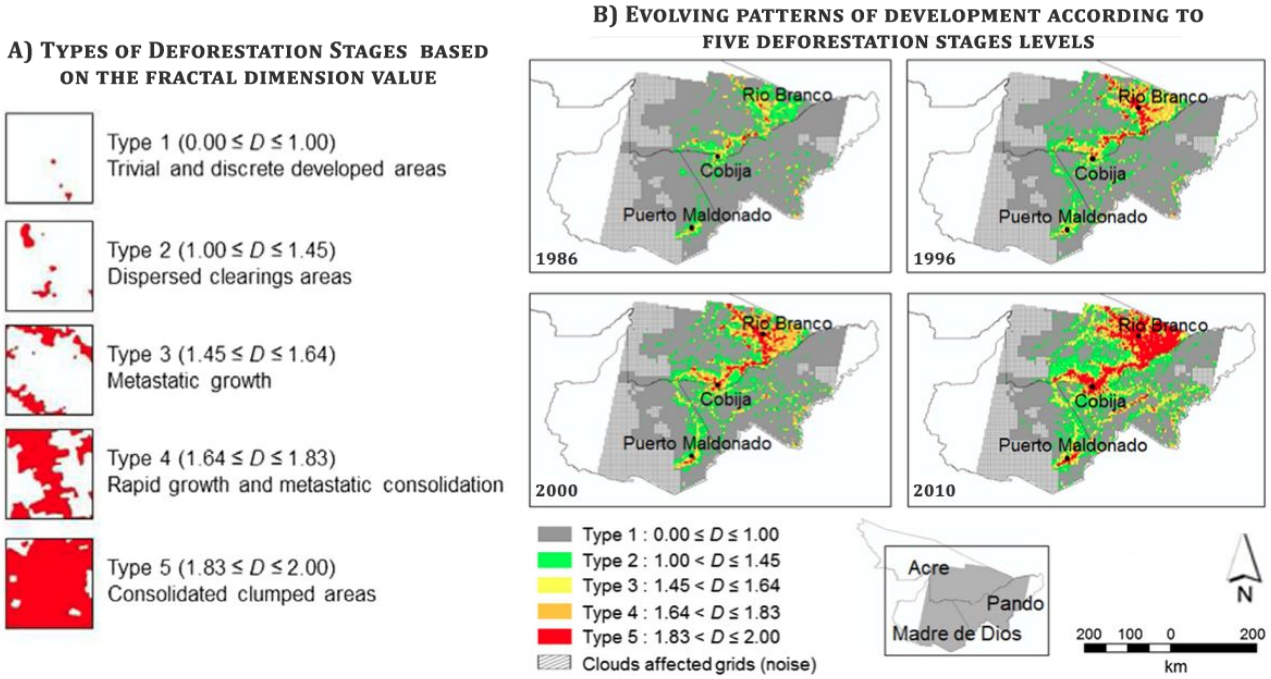


Figure 2. Types of deforestation stages (A) according to the fractal dimension D computed using the box-counting approach and bottom-up method. These stages were mapped on multi-temporal Landsat forest/non-forest maps in the Western Amazon (B). Adapted from Sun *et al.* (2014).

The mining land-use patterns-based method supports a better understanding of the land-use agents...

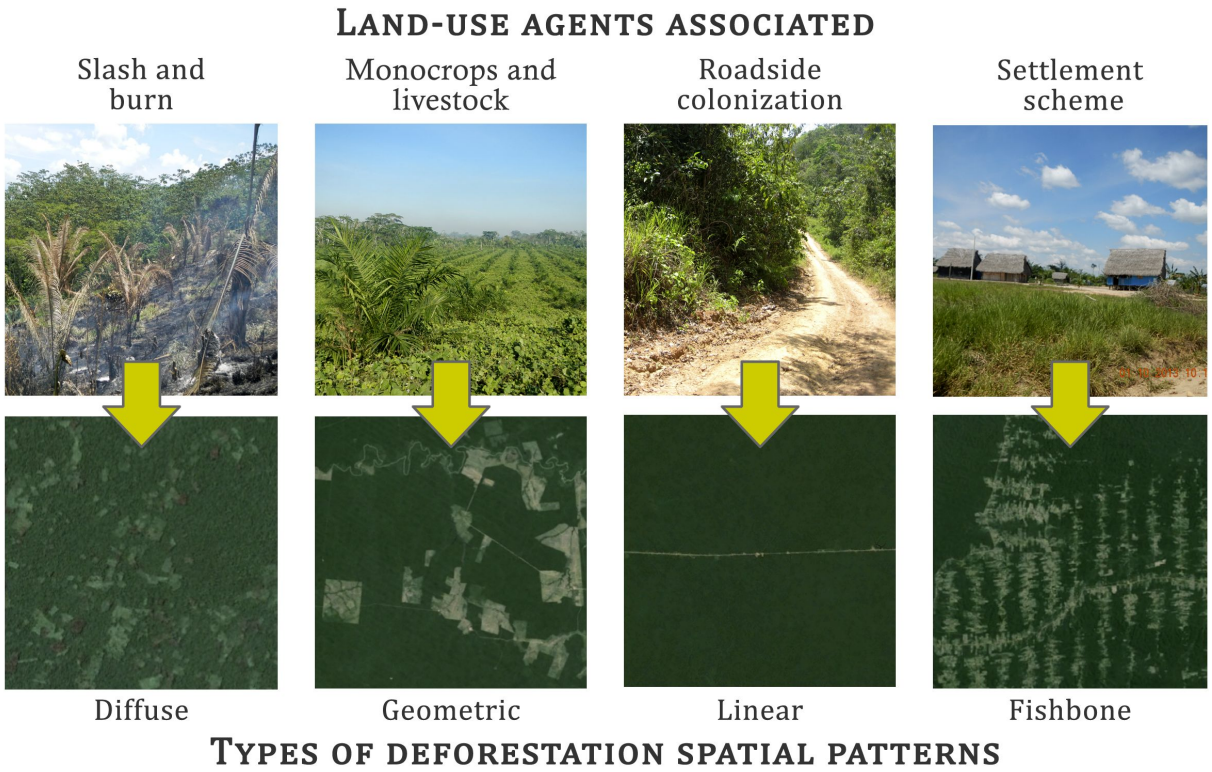


Figure 3. Relationships between four types of deforestation spatial patterns, visually differentiated using Google Earth Imagery, and related land-use agents. Photos provided by the Terra-i project (2015).

The mining land-use approach

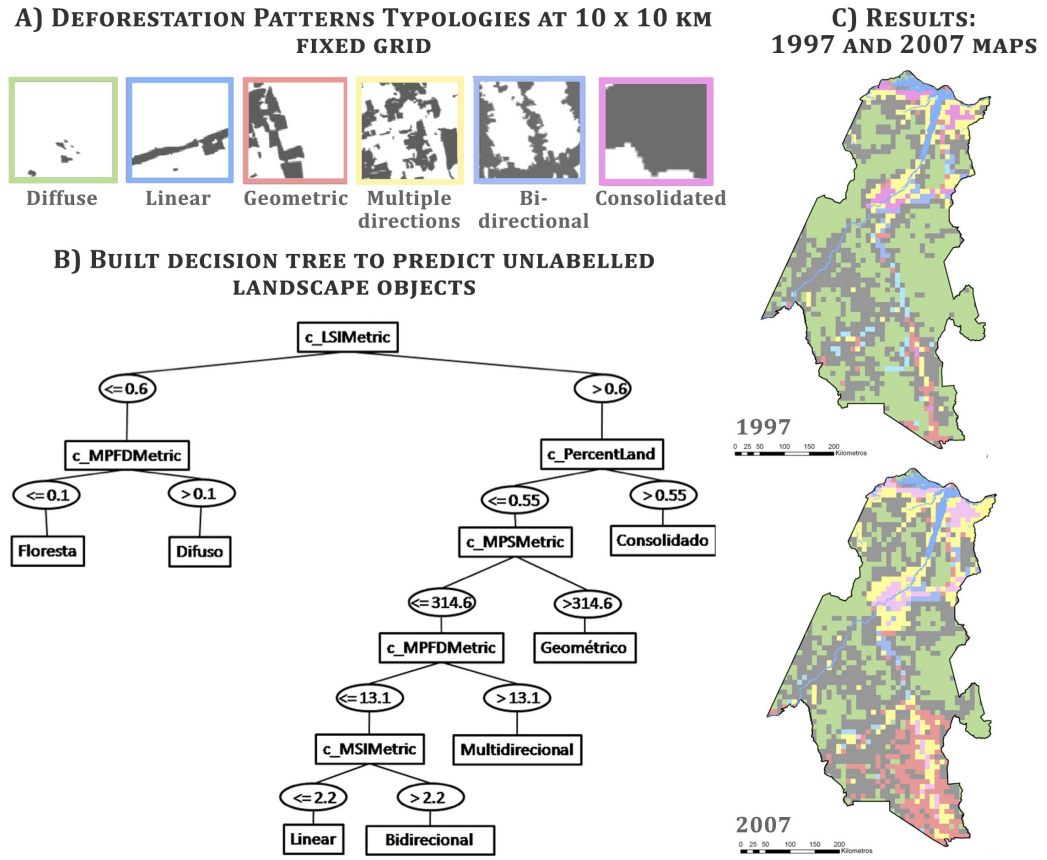


Figure 4. Example of the application of the mining land-use patterns-based analysis. Six spatial patterns (A) were identified at the unit of analysis (10 x 10 km) from PRODES deforestation dataset (30 m). Using a built decision tree (B) these patterns were discriminated for multiple periods (C). Colours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

The mining land-use approach

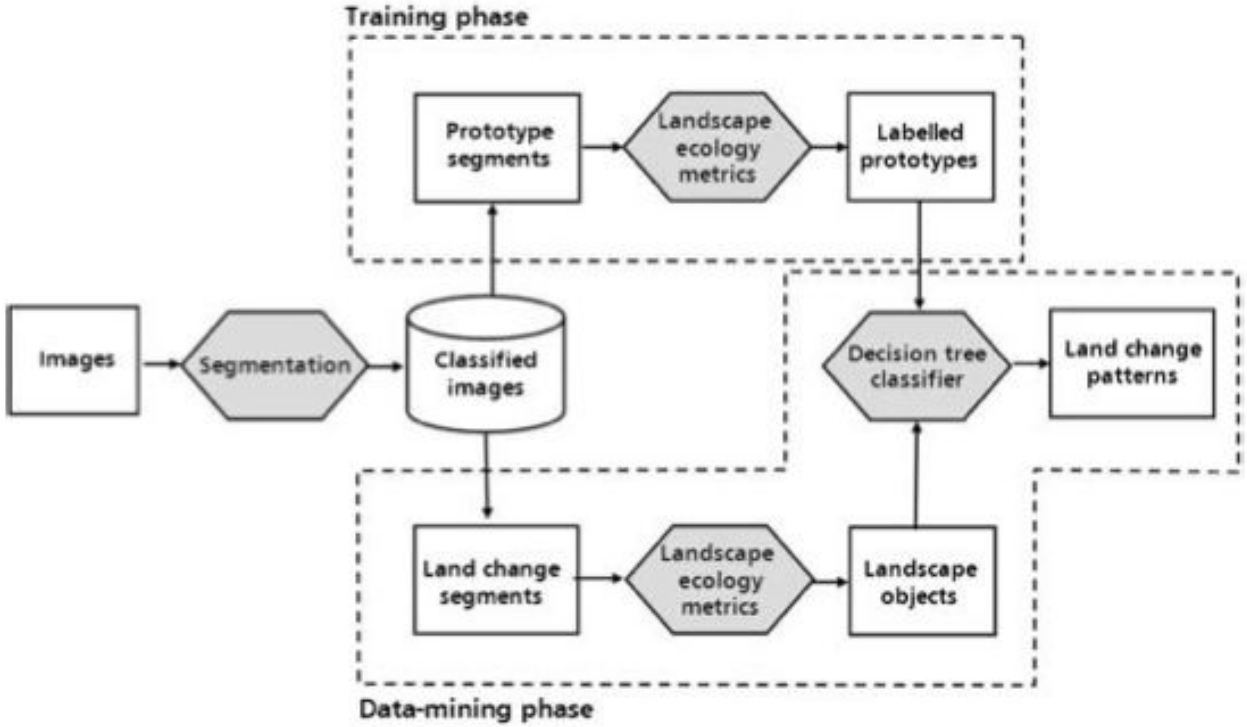
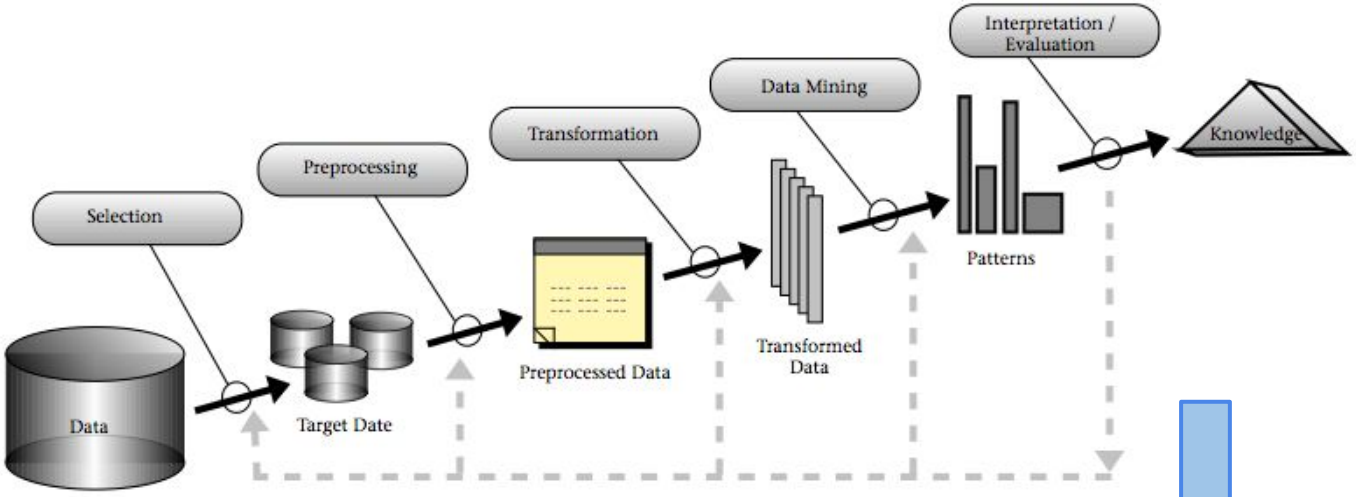
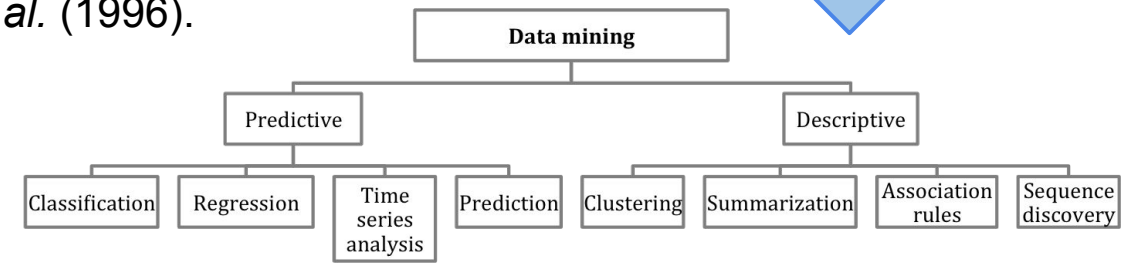


Figure 5. Workflow of the mining land-use patterns-based method. Source: Silva *et al.* (2011).

The Knowledge Discovery in Databases (KDD) as the scientific method behind a data mining analysis



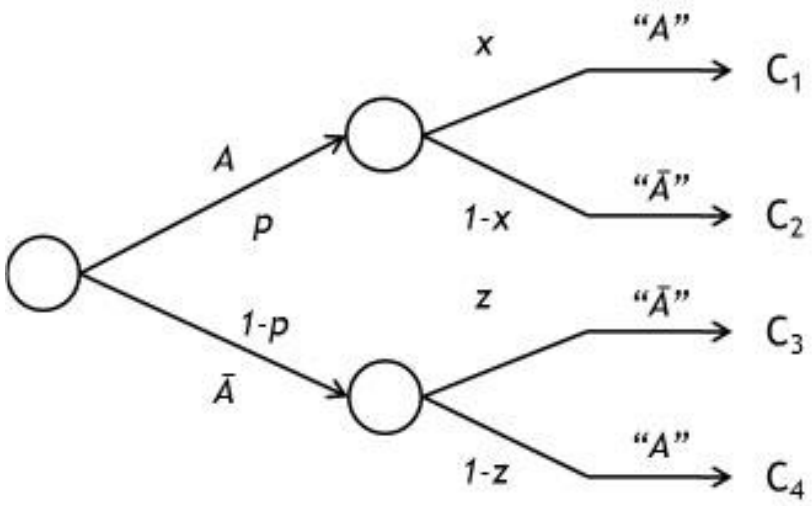
Source: Fayyad *et al.* (1996).



Source: Dunham (2002).

DM algorithms used in the mining land-use approach

Decision Trees (DT)



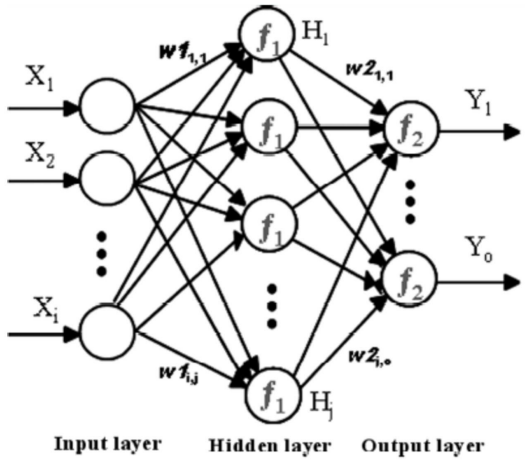
The C4.5 classifier (Quinlan, 1993) as a DT algorithm commonly used in the mining land-use approach.

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.

C4.5 employs post-pruning (set by using the confidence level value) which allows building smaller tree models.

Multi-perceptron feed-forward artificial neural networks (MLP-FF-ANNs)



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.

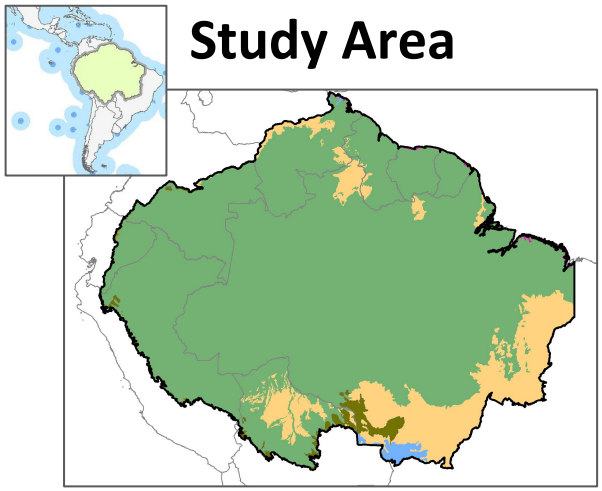
Research aim

To explore techniques to map types of deforestation spatial patterns and stages in the Amazon rainforest from existing remote-sensing image databases

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

Methods

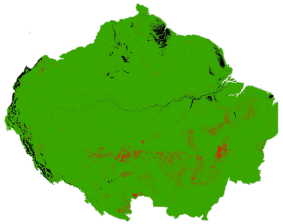
Study Area



Humid Forest Savannah

INPUT DATA

Terra-i, MODIS, 250m



2004-2013
Terra-i data
(v. 2004-March 2014)

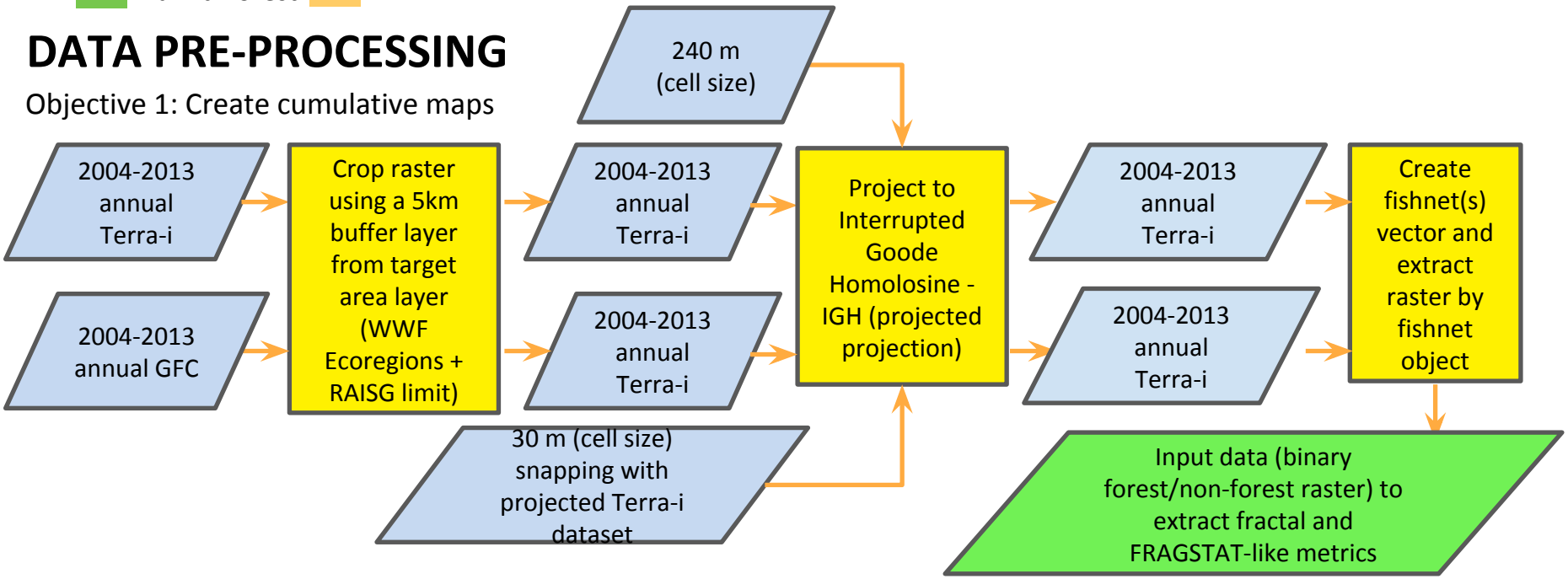
Hansen, Landsat, 30 m

2001-2013
Hansen data
(v. 2013)

Geodata visualization / pre-processing using **ArcMap v.10.2**
Amazon Boundary according to RAISG (Biophysical + Political criteria)

DATA PRE-PROCESSING

Objective 1: Create cumulative maps



Methods

Objective 2 and 3: Mapping techniques and Analysis - The fractal analysis approach

Based on the R-code provided by Sun *et al*. (2014), the fractal dimension (*D*) was computed using the Box-counting method (fixed grid). In accordance to this method, an iterative number, *k*, must be defined k= [0,....m]

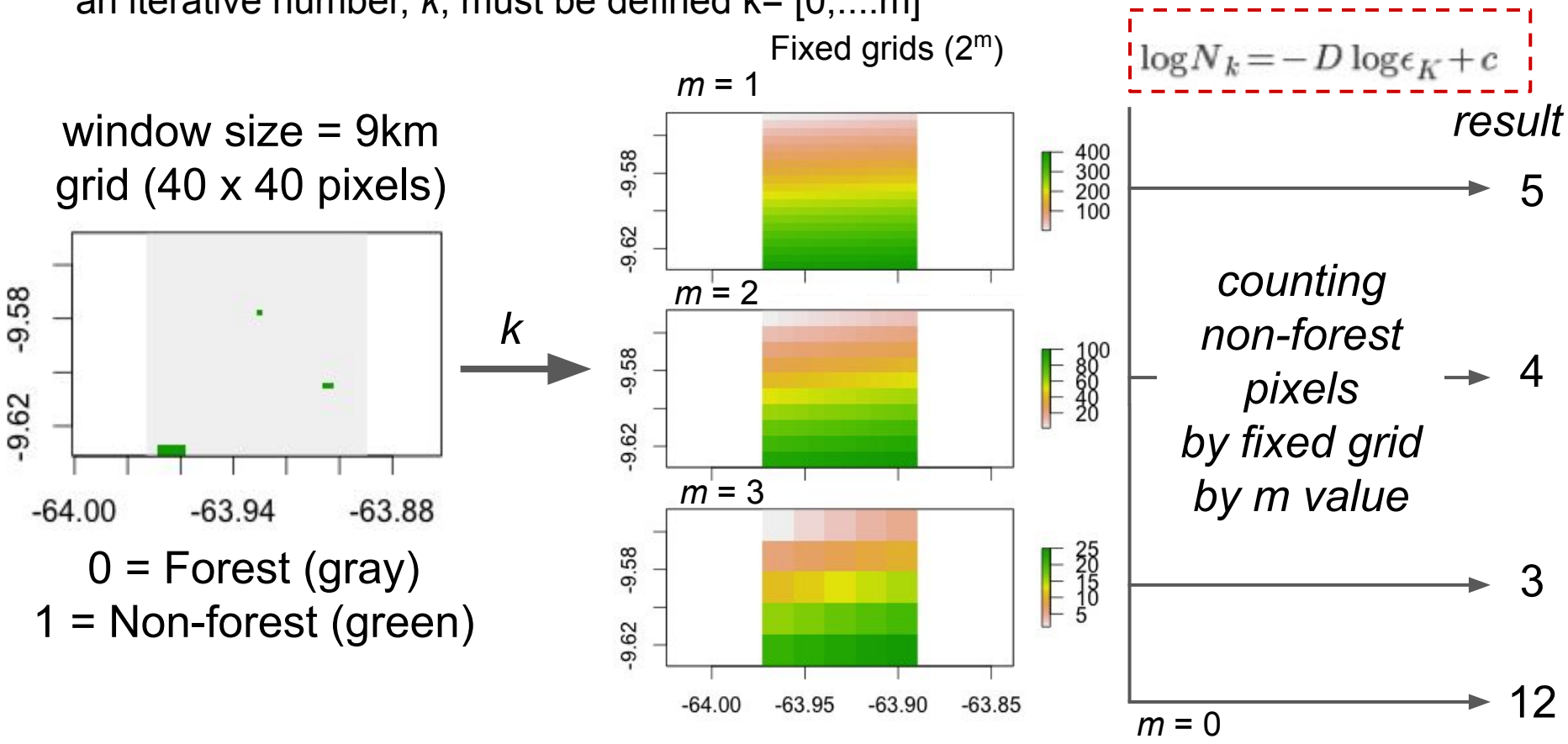


Figure 5. Graphical representation of the parameters and data used in the R-code by Sun *et al*. (2014) using Terra-i data (250m).

Methods

Objective 2 and 3: Mapping techniques and Analysis - The fractal analysis approach

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

Methods

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (workflow)

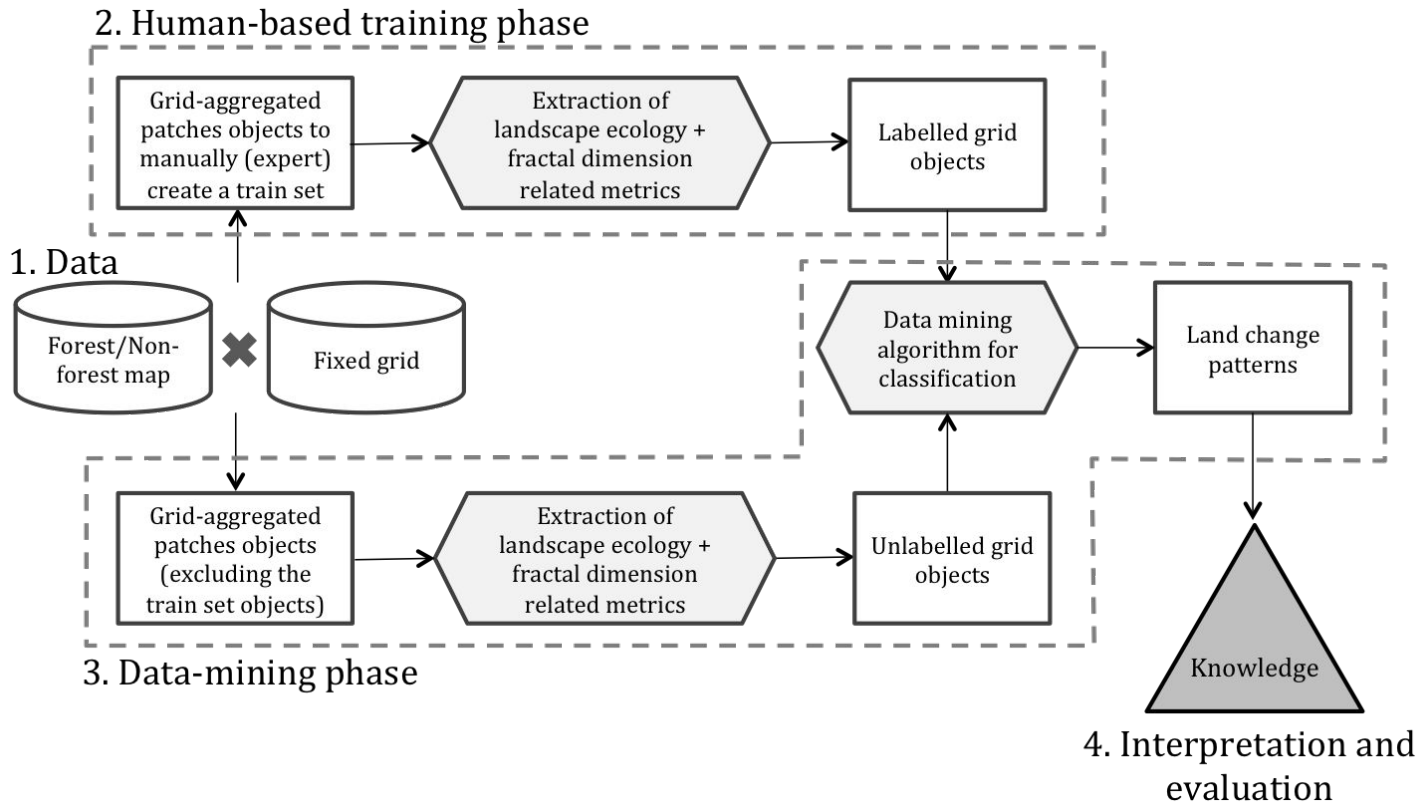
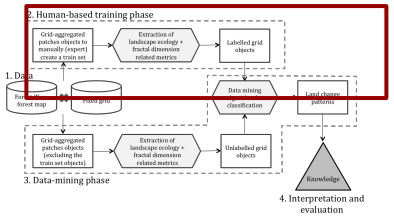
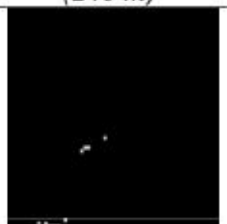
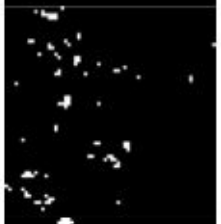
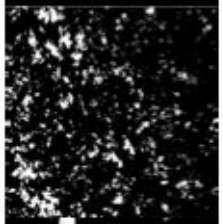
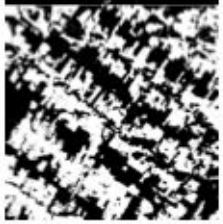


Figure 6. Workflow of the mining land-use patterns-based method. Source: Silva *et al.* (2011).

Methods

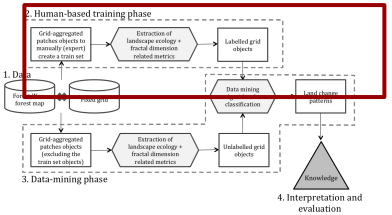
Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Human-based training phase)



Pattern	*Visualization using the 15360 m x 15360 m grid size by dataset (aggregated 2004-2013 forest/non-forest map)* Forest / Non-Forest — *Terra-i (240 m)*	*GFC (30 m)*	*Description (scale 1:75,000)*	*Land-use agents associated / Spatial distribution / Accessibility (transport infrastructure)*
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

Methods

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Human-based training phase)



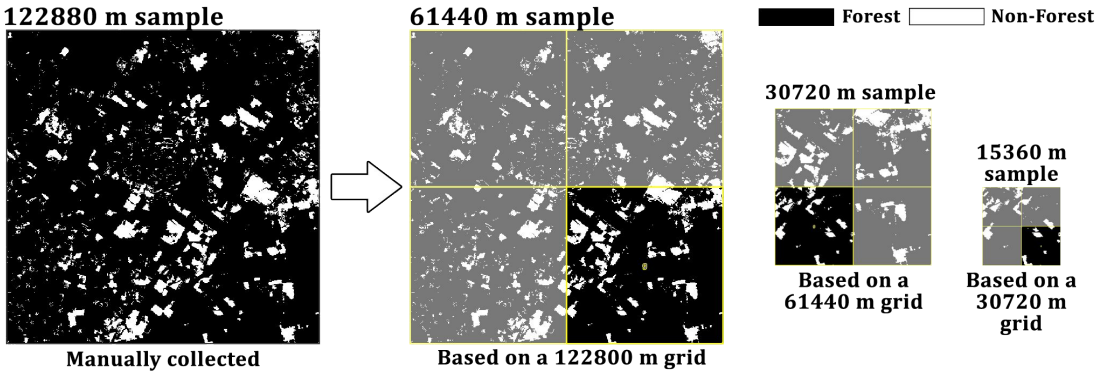


Figure 7. Example from the Terra-i dataset of the sampling strategy. A labelled train set of the geometric pattern was generated and later used in the data-mining phase.

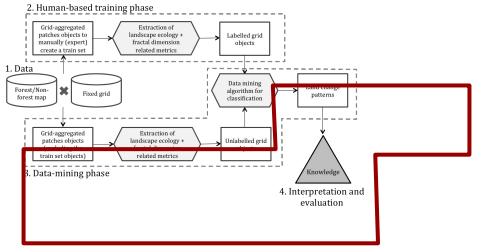
SAMPLES
15 representative grid-based objects organised by pattern (60 samples in total for the four patterns)

METRICS EXTRACTED (15 in total)

- A group of thirteen FRAGSTAT-like metrics at class level extracted by the fishnet object (unit of analysis) - Metrics were selected based on previous studies about the mining land-use approach.
- The fractal dimension (D) determined in the fractal analysis approach was added to the FRAGSTAT-like metrics database.
- Additionally, the ratio between the fractal dimension and the FRAGSTAT-like metric related with the percent of land occupied by the non-forest class (D_LAND).

Methods

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Data mining phase)



There were two types of datasets used in this phase.

- The “gold-standard” (human- labelled) dataset, prepared by grid size and dataset, was used to build data mining algorithms models (C4.5 and ANNs).
- Another set (unlabelled grid-based objects) was then automatically classified within the four spatial pattern typologies using the built models with the best performance (Kappa value) from the gold-standard set.

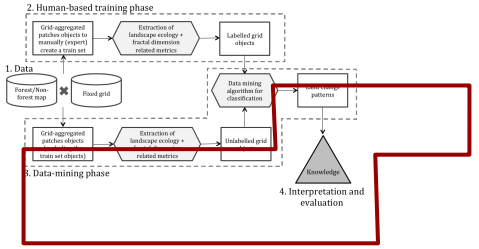
The workflow of the data mining phase was divided into five sub-phases:

- data preprocessing
- model construction (learning)
- model evaluation (accuracy)
- model sensitivity analysis
- model use (classification)

The first three steps were examined for all C4.5 decision trees and ANNs models generated.
The remaining steps (model sensitivity analysis and model use) were exclusively performed for the built ANNs models with the best performance according to the Kappa value.

Methods

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Data mining phase)



- Data preprocessing

Preprocessing (normalisation) of input data as a critical step for built ANNs models with backpropagation.

Each input (metrics) were normalized between 0 and 1.

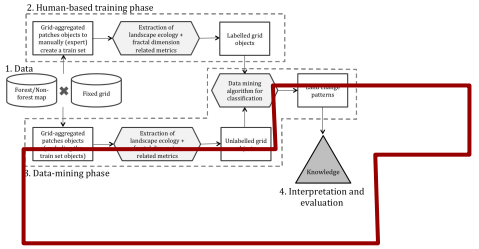
- Model construction (learning)

Several architectures can be derived from the two data mining algorithms tested.

- For C4.5 decision trees, it was used the default confidence level (CF) of 25% for post-pruning in accordance to previous mining patterns studies.
- For ANNs, the weight decay and hidden nodes number were the parameters tuned according to the type of neural network implemented.
  - The grid-search procedure with *k*-fold cross validation for selecting the candidate parameters in the neural networks
  - 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.

Methods

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Data mining phase)



- Model evaluation (accuracy)

The performance of each model (also denoted as surrogate model in this step) was evaluated using a five-fold cross validation with 40 iteration (small training set size with the presence of extreme values).

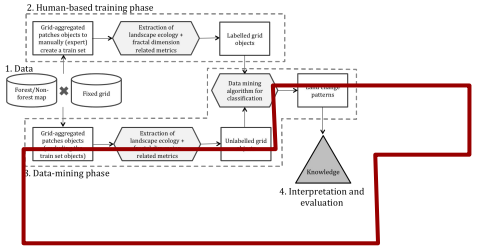
The iterated *k*- fold cross-validation technique allowed exploration of the stability of each surrogate model and identification of optimal parameter combinations for the neural networks algorithm.

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.

Methods

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Data mining phase)



- Model sensitivity analysis

The sensitivity analysis is a key procedure in model development to detect input-output dependencies/There are several techniques such as tested by Olden et al. (2004) being the connection weights (CW) method was highlighted as the best methodology for accurately quantifying importance of variables.

CW 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. This method distinguishes if the input has a positive or negative effect on each target output (Yan et al., 2012).

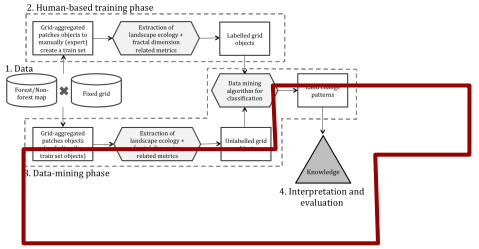
- Model use

Using the ANNs models with best perfomance, the unlabelled sets distributed in the grid sizes and deforestation datasets assessed were automatically classified.

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 tothat consisted of 30 samples (double the amount used for training).

Methods

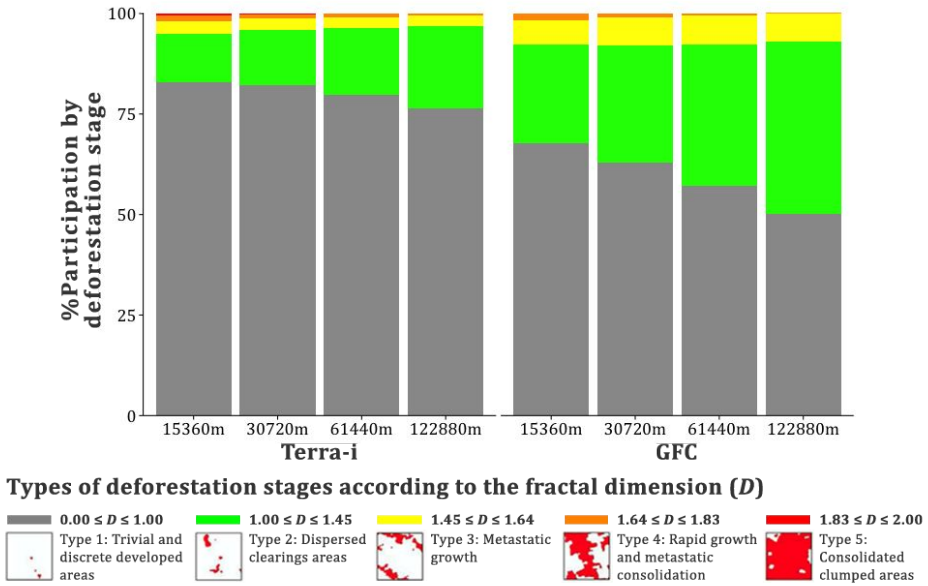
Software and implementation



- Quantum GIS v.2.1.4
  - It was primarily used for vector and raster operations 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.
- R software v.3.2
  - It was used for extracting the FRAGSTAT and fractal-like metrics from labelled and unlabelled grid-based objects across all grid sizes and datasets. An exploratory analysis using box-plots was performed over the normalised metrics for each training set.
    - FRAGSTAT-like metrics were selected using FRAGSTAT v.4.2 (McGarigal, 2015), called within an R code.
    - For the FD, this value was extracted using the raw programming codes provided by Sun et al. (2014).
  - Data mining projects were run using the caret v.6.0-52 R package
    - 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 sensitivity analyses were performed independently using the NeuralNetTools v.1.3.1. R-package (Beck, 2015), which performed the connection weight method.
  - 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.

Results

Objective 2 and 3: Mapping techniques and Analysis - The fractal analysis approach



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.

The Terra-i dataset had a higher proportion of type 5 deforestation stage than the GFC dataset.

The type 1 (no or trivial clearing areas) and type 2 (dispersed clearings areas) stages presented the largest changes by grid size (see for example GFC forest/non-forest maps.

Results

Objective 2 and 3: Mapping techniques and Analysis - The fractal analysis approach

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 I. 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

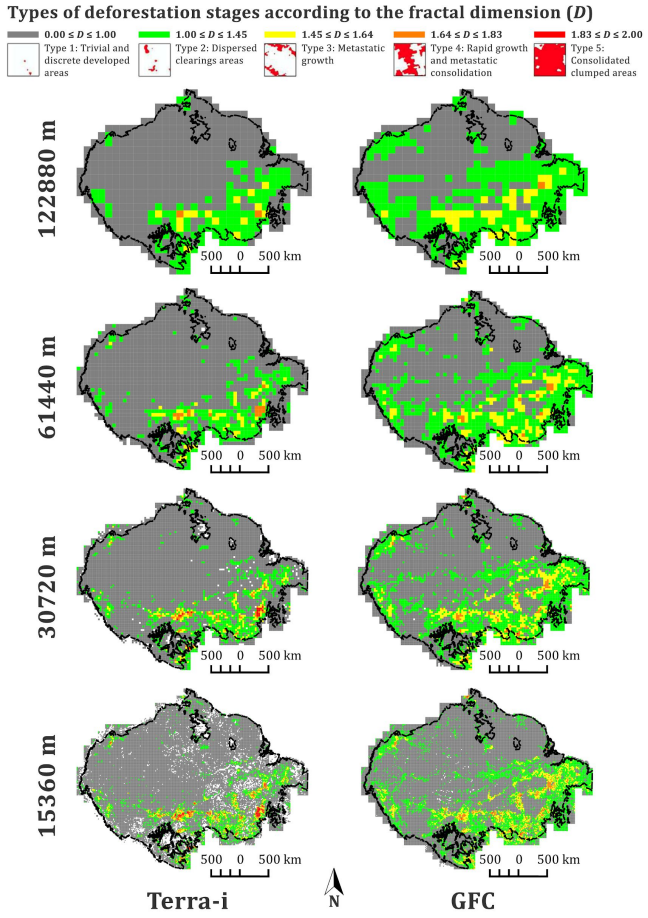


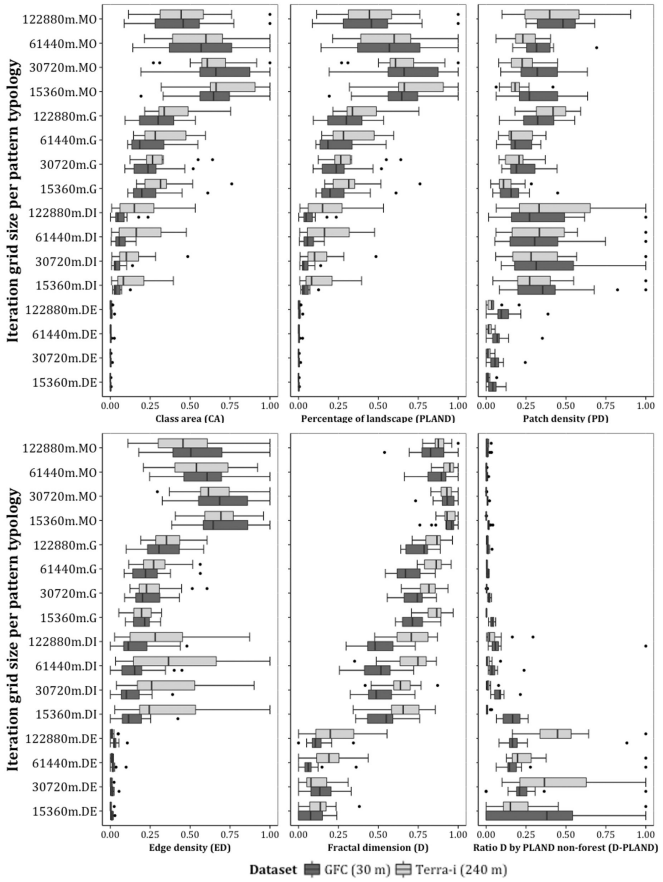
Figure 9. Types of deforestation stages proposed by Sun et al. (2014) implemented in the study area over four grid sizes (each row) and two deforestation datasets (Terra-i, left column; GFC, right column).

Results

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Exploratory analyses)

Overall, all metrics presented different behaviours, being identified those with potential for discriminating the target patterns typologies.

Figure (right). 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.



Results

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Model evaluation)

400 ANNs and 8 C4.5 models were evaluated.

Figure 15 illustrates boxplots each best ANN parameter combination model confronted with the C4.5 reference model at each grid size by dataset.

The table below contains the best parameter combinations (number of hidden nodes and decay) for ANN models that had the highest Kappa value in each grid size by dataset.

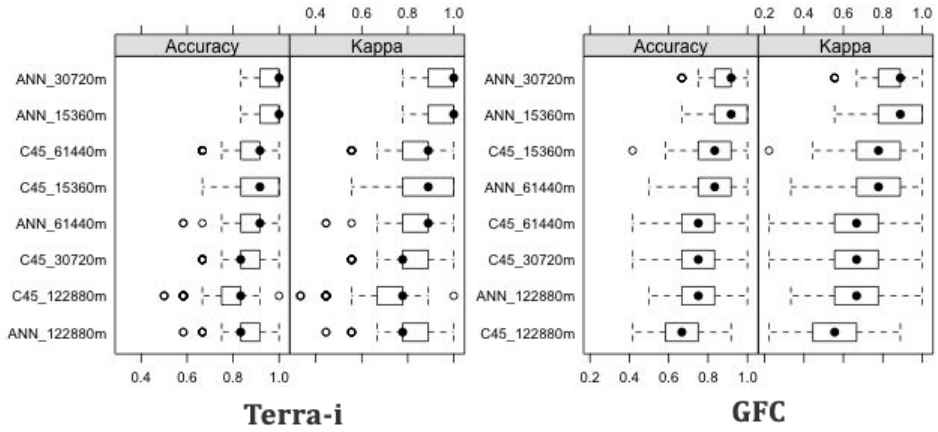


Figure 10. Distribution of overall accuracy and Kappa values from best ANN parameter optimization models and reference C4.5 models by grid size separated by dataset (left: Terra-i, right: GFC). Models are ranked from the highest (upper) to the lowest (bottom) median (black dot) metrics values.

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

Results

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Model evaluation)

Table II. 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

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.

Although mean 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.

Results

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Model sensitivity analysis)

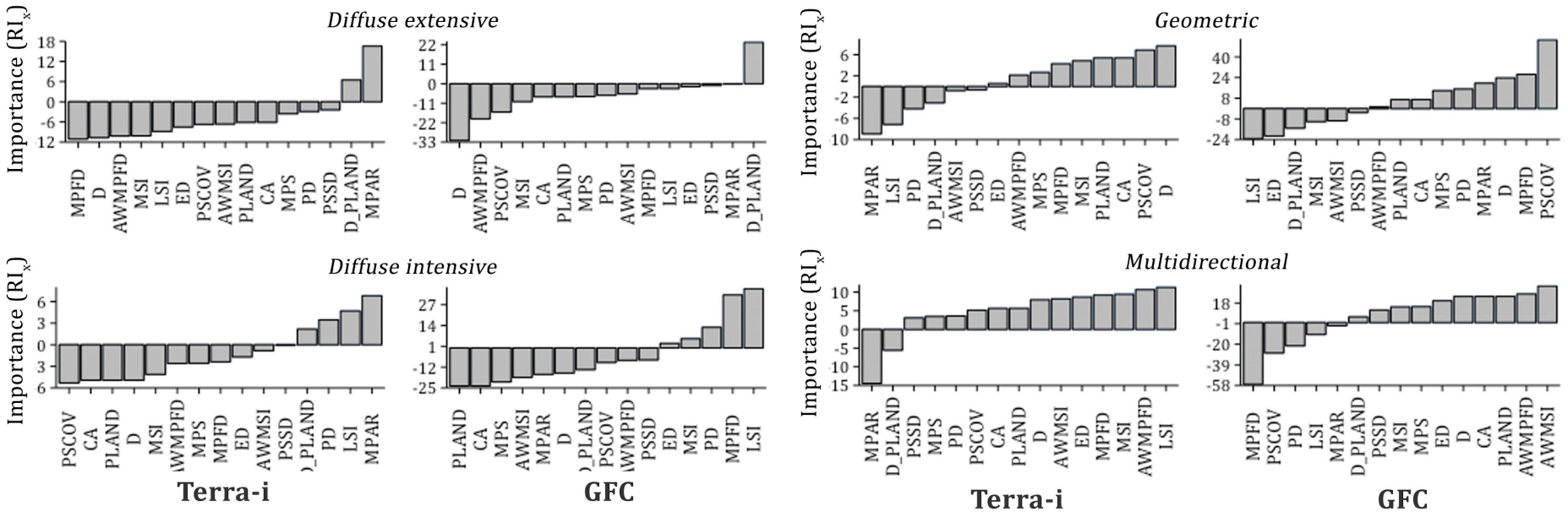


Figure 11. Sensitivity analysis bar plots ranking 15 input variables (normalised FRAGSTAT and fractal-like metrics) for the best ANN models at a grid size of 30720 m by pattern typology (row) and dataset (column).

The contribution of ANN models inputs, measured with Olden and Jackson's (2002) RIx value method, seemed to vary according to pattern typology and dataset.

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.

The proposed computed fractal-like metrics (denoted as D and D_PLAND in Figure 11) seemed to be promissory for deforestation pattern typologies studies using mining patterns-based

Results

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Model use / Terra-i)

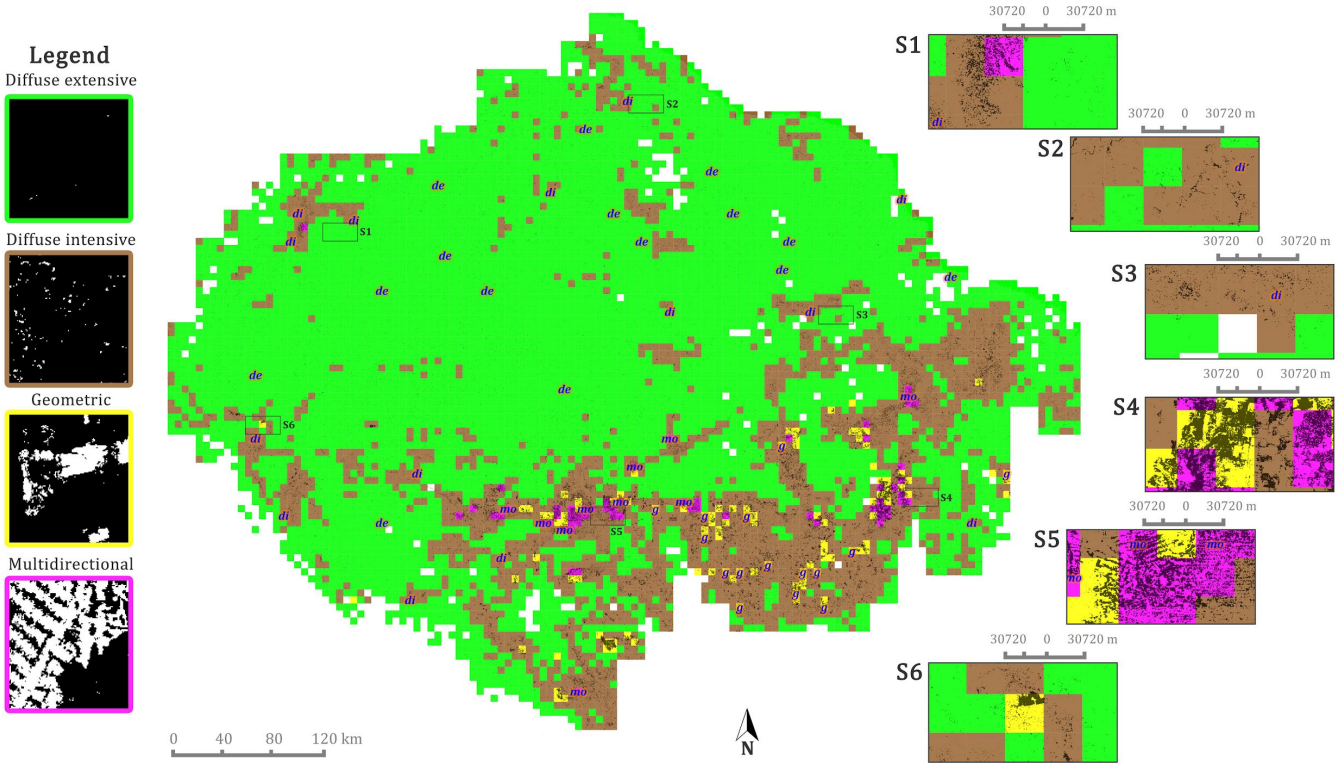


Figure 12. Distribution of deforestation spatial patterns types using the best ANN model at a grid size of 30720 m from forest/non-forest Terra-i maps in the study area (central side). Right side consists of map insets in six random locations. Multi directional, geometric, diffuse intensive and diffuse extensive pattern training sets are denoted using blue italic letters as mo, g, di and de, respectively, for all maps.

Results

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Model use / GFC)

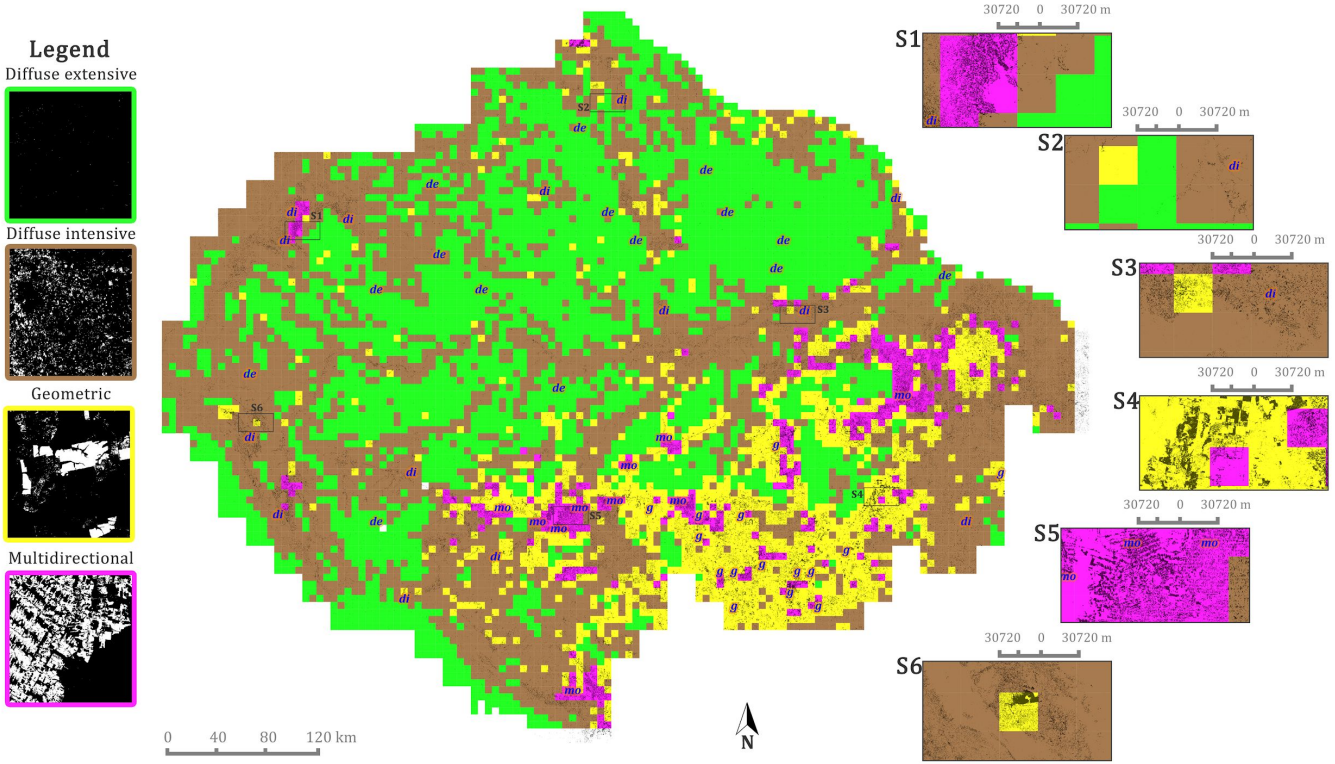


Figure 13. Distribution of deforestation spatial pattern types using the best ANN model at a grid size of 30720 m from forest/non-forest GFC maps in the study area (central side). Right side consists of map insets in six random locations. Multi directional, geometric, diffuse intensive and diffuse extensive pattern training sets are denoted using blue italic letters as mo, g, di and de, respectively, for all maps.

Results

Objective 2 and 3: Mapping techniques and Analysis - The mining land-use approach (Model use / GFC)

Table III. 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%)

Table IV. 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

Conclusions

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 from two remote sensing based datasets (Terra-i and GFC).

Both methods suggested the dominance of initial stages of deforestation, or patterns described as dispersed and/or clustered distributed deforested areas.

For the former framework:

- 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.
- Although changes could not be completely attributed to spatial resolutions, 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.

For the latter framework:

- ANNs are most suitable for mining patterns-based approach in comparison to the traditional C4.5 decision tree algorithm.
- 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.
- The sensitivity analysis indicated that the fractal-like metrics and seemed to be useful inputs in pattern mapping research using ANNs. The patch size standard deviation (PSSD) FRAGSTAT-like metric may be omitted for future studies using the same best ANN architectures indicated here.

Limitations and future research

- Misrepresentation of pattern typologies in each area analysed due to lack of information of past events of deforestation outside of the datasets' detection periods;
- Assess more spatial landscape heterogeneity measurements such as FRAGSTAT-like (both at class and landscape level) and fractal-related metrics (i.e. lacunarity) for pattern mapping research using data mining techniques;
- 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 (see 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.
- 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


Aerial photo - Oil Palm Plantation/Rainforest margin in Ucayali, Peru (Source: CIAT, 2013)
Thanks!
Questions?
KING'S College LONDON