Environmental and Ecological Statistics

Original Research

How well does random forest analysis model deforestation and forest fragmentation in the Brazilian Atlantic forest?

Lisiane Zanella[1,2,3]*, Andrew M. Folkard[2] George Alan Blackburn[2], and Luis M. T. Carvalho[3]

[1] Federal Institution of Rio Grande do Sul (IFRS)
Osório, Rio Grande do Sul, Brazil
e-mail: lisiane.zanella@osorio.ifrs.edu.br
phone: +55 51 99866 7272
ORCID: 0000-0002-6830-6896

[2] Lancaster Environment Centre
Lancaster University,
Lancaster, Lancashire, UK

[3] Ecology and Conservation Sector, Department of Biology / Department of Forestry Sciences
Federal University of Lavras (UFLA)
Lavras, Minas Gerais, Brazil

* Corresponding author

Abstract

We assessed the value of applying random forest analysis (RF) to relating metrics of deforestation (DF) and forest fragmentation (FF) to socio-economic (S-E) and bio-geophysical (BGP) factors, in the Brazilian Atlantic Forest of Minas Gerais, Brazil. A vegetation-monitoring project provided land cover maps, from which we derived DF and FF metrics. An ecologic-economical zoning project provided more than 300 S-E and BGP factors. We used random forest analysis (RF) to identify relationships between these sets of variables, and compared its performance in this task to that of a more traditional multiple linear regression approach. We found that RF modelled relatively-well variance in all metrics used (the rate of deforestation, the amount of forest, and the density and isolation of forest patches), presenting a better performance when compared to the classical approach. RF also identified geographical location and topographic factors as being most closely associated with patterns of DF and FF. Both analyses found factors associated with economic productivity, social institutions, accessibility and exploration to have little relationship with metrics. RF was better at explaining variations in rates of deforestation, remaining forest and patch patterns, than the multiple linear regression approach. We conclude that RF provides a promising methodology for elucidating the relationships between land use and cover changes with potential drivers.

Keywords: Land use and land cover change, Socio-economic and bio-geophysical factors, Machine-learning technique, Stepwise Multiple Regression, Minas Gerais State, Tropical forests.

Acknowledgements

The authors would like to thank the Federal University of Lavras (UFLA) for providing the data. L. Zanella would like to acknowledge support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), who provided a Ph.D. scholarship, and R. Solar and T. S. de Carvalho for assistance with the statistical analysis.

1 Introduction

A large proportion of the Earth's surface has been transformed by anthropogenic land use activities in recent centuries. Land use and land cover change (hereafter, LUCC) was once considered a local environmental issue, but is becoming globally important due to its increasingly widespread effects upon natural environments (Foley 2005; Lambin and Geist 2006). Comprehending these effects requires, in part, the understanding of relationships between variations in socio-economic (hereafter, S-E) and bio-geophysical (hereafter, BGP) factors associated with the LUCC with which they co-occur (Geist and Lambin 2001; Geist and Lambin 2002). However, understanding these relationships is difficult because LUCC is a result of complex interactions among social, economic, and environmental factors acting across different scales of space and time (Geist and Lambin 2001; Geist and Lambin 2002; Caldas et al. 2013). Therefore, it is necessary to design studies carefully so that inferences are reliable. Unreliable conclusions can lead to distorted management recommendations, resulting in missed conservation opportunities, and a waste of resources and time (Oliveira et al. 2017).

Several studies have investigated relationships between LUCC and a wide variety of S-E and environmental factors. LUCC are commonly expressed in terms of deforestation rates (DF) and forest fragmentation metrics (FF). Examples of these multiscale and multifactor dynamics influencing LUCC patterns are: the increasing demand for food and other commodities (Aide and Grau 2004; DeFries et al. 2004; Barbier et al. 2010; Caldas et al. 2013), shift in regional economies household level conditions (Perz 2004; Richards et al. 2008; Wright and Samaniego 2008; Gaughan et al. 2009), indirect effect of tourism (Gaughan et al. 2009), globalization of markets (Hecht et al. 2006; Parés-Ramos et al. 2008), and presence and effectiveness of social institutions (Hecht et al. 2006; Richards et al. 2008).

The studies addressing the impacts of LUCC upon tropical systems has also improved significantly in recent decades (Malhi et al. 2014). Those impacts have been separated into two different types: underlying (or indirect) and proximate (or immediate) causes (Geist and Lambin 2002). Proximate causes are human actions that directly affect these changes, while underlying causes affect these changes indirectly (Geist and Lambin 2002). The main recognized proximate causes of LUCC in tropical countries are: agricultural expansion (e.g., shifting cultivation and permanent cultivation), cattle ranching, and infrastructure expansion (e.g., transportation infrastructure) (Pfaff 1999; Geist and Lambin 2001; Perz et al. 2007). Furthermore, LUCC is also influenced by the underlying drivers, especially demographic dynamics (e.g. population growth) and economic factors (e.g. local or international demand for commodities) (Geist and Lambin 2001; Caldas et al. 2013). In many regions, there is a clear relationship between population change and LUCC (Geist and Lambin 2002). However, other studies have shown that LUCC can be modified by socio-economic and environmental factors (Geist and Lambin 2002).

A few studies have attempted to investigate drivers and associated factors of land use and cover changes in the Brazilian Atlantic Forest. Silva et al. (2007) conducted a local scale study and found an indirect influence of topographic relief on forest cover. (Teixeira et al. 2009) showed that proximate causes influence the dynamics of deforestation and forest re-growth. They identified that losses in young secondary vegetation and forest were far from rivers, on gentle slopes and near urban areas, while higher forest re-growth rates were near rivers, on steep slopes and far from dirt roads. Freitas et al. (2010) analysed the effects of roads, topography, and land use on forest cover dynamics and demonstrated that forest dynamics were directly related to past road density, past land use (buildings and agriculture expansion), and slope variation. Lira et al. (2012) described LUCC in three Atlantic Forest fragmented landscapes (in São Paulo state) over time and found that LUCC deviated from a random trajectory. Their results also suggested a forest transition in some Atlantic Forest regions. Freitas et al. (2013) used a combination of statistical approaches – multivariate data analysis (CCA), linear regression models (OLS), local spatial regression models (GWR) and spatial clustering procedures (SKATER) – to investigate relationships between LUCC processes and environmental and S-E factors in an Atlantic Forest region with an area of ~12,000 km^2 in the state of Rio Grande do Sul. Their findings revealed a competitive and inter-related set of LUCC processes, due to the landscape complexity. More recently, Ferreira et al. (2015) investigated how forest cover and agricultural land use varied in an area of Atlantic Forest in São Paulo state, emphasizing sugarcane expansion. Besides, a general trend of decline followed by stabilization of forest remnants in this biome may be assumed due to different deforestation rates in the Brazilian states (SOS Mata Atlântica/INPE 2014). However, there are discrepancies between data sets provided by different organizations, which is necessary to understand the landscape dynamics (Farinaci and Batistella 2012).

LUCC studies also used a range of statistical techniques. Some studies have used relatively simplistic approaches, such as Mann-Whitney and Kruskal-Wallis tests (Quezada et al. 2013), or correlation analyses (Beilin et al. 2014). Others have applied more robust approaches, combining or comparing different methods, such as statistical redundancy analyses (RDA) (Parcerisas et al. 2012); ordinary least squares regression (OLS) and geographically weighted regression (GWR) (Jaimes et al. 2010; Gao and Li 2011); canonical correspondence analysis (CCA), OLS, GWR and spatial clustering procedures (Freitas et al. 2013); stepwise multiple regression models (Gong et al. 2013). Most of these studies considered a limited number of potential independent factors that had normal distributions, as this is the basic requirement for using parametric techniques. Therefore, modelling approaches must be further evaluated in terms of the choice of independent and metrics, as well as the selection and interpretation of appropriate statistical methods. There is also a need for further studies that include a large number of factors encompassing, as much as possible, all aspects of the S-E and BGP context within which LUCC is taking place.

Despite all these improvement in our understanding of the impacts of LUCC on tropical environments, there is still no optimal tool for understanding relationships between deforestation/forest fragmentation and S-E or BGP factors. Random Forest analysis (RF; Breiman (2001) is a variable selection technique and has great potential in this respect. RF is capable of identifying complex interactive and non-linear response-predictor relationships, and has excellent predictive performance (Prasad et al. 2006, Smith et al. 2011). Thus, application of RF analysis to disentangle these sorts of relationships may be particularly useful. RF is used widely in bioinformatics (Cutler and Stevens 2006), for land cover classification (Gislason et al. 2006) and analysis of medical experiments for example, with few ecological applications (Prasad et al. 2006). It has recently gained popularity in ecology (Fu et al. 2010; Gilbert and Chakraborty 2011; Bonilla-Moheno et al. 2012; Ellis et al. 2012; Leal et al. 2016).

In this study, we investigate RF regression applied to the task of identifying relationships between a large set of S-E and BGP candidate independent variables (factors), and metrics which quantify the current patterns of deforestation (DF) and forest fragmentation (FF) of the Brazilian Atlantic Forest in the state of Minas Gerais, Brazil. This study considers an unusually large set of more than 300 S-E and BGP factors. Our main objective is to measure the RF ability to identify relationships with variables that describe patterns of forest fragmentation, S-E/BGP and compared its results with those derived from application of stepwise multiple linear regression, a classical statistical approach, to the same datasets. Our hypotheses are: 1. RF is better than STEP at elucidating relationships between S-E and BGP factors and FF/DF metrics. Because of RF capability of identifying complex interactive and non-linear response-predictor relationships, we believe that this analysis address the relationships between factor and metrics more accurately than the classical approach we considered here; 2. RF and STEP identify broadly the same S-E and BGP factors as being most important in explaining variation in FF/DF metrics. Based on the LUCC literature, we expect that certain factors will be identified by the analyses as most important, regardless of the methodological approach used, such as population and roads densities, and topographic measurements (e.g. Geist and Lambin 2002; Silva et al. 2007; Freitas et al. 2010).

2 Methods

2.1 Study area

The study area is located within the state of Minas Gerais, in South-eastern Brazil and comprises the 518 municipalities which fall entirely within the largest contiguous area of the Atlantic Forest biome, and encompasses 34% (19,904,146 ha) of Minas Gerais (IBGE 2017, Fig 1). This study site has a wide variability across the municipalities in the magnitude of DF/FF metrics and in the S-E/BGP factor values.

The study region characterized by rolling hills which rise from 200 m to a medium altitude of 1600 m. It is a very rugged area with a large proportion of highlands as well as plateaus and plains. There are several climates types linked to the different relieves: warmer climate in the north and cooler in the south. The distance from the ocean also has a climatic effect (maritime vs. inland climate, etc) upon the study area. The region is, on average, relatively sparsely populated, with a tendency in higher concentrations of populations towards the south, which has smallest municipality areas too. The south part of the study area is also relatively richer and developed when compared to the other part of the study area and to the Brazilian average. The main industries and sources of employment are the bovine cattle herd - which corresponds to 10% of the Brazilian total -, coffee production and the extraction of iron ore. The location of Belo Horizonte, the largest city and the capital of Minas Gerais, also plays an important role in the establishment of many industries, especially automobile and steel mill industries in its vicinities. This makes Minas Gerais the second largest automotive and metal foundry hub in Brazil. All these information on the study area patterns and more can be found in the ecologic-economical zoning of Minas Gerais, ZEE-MG (Scolforo et al. 2008).

#Fig 1 approximately here

2.2 Variable selection

This work used large datasets provided by two broader-scale projects carried out in Minas Gerais State, Brazil. The DF and FF metrics were derived from the vegetation monitoring system dataset (Scolforo and Carvalho 2006; Carvalho and Scolforo 2008, Carvalho and Scolforo - *unpublished data*), which comprises land cover maps from 2003 to 2011.

A deforestation metric, the growth rate of deforestation (GRD, percentage) was calculated for each municipality using digital change detection applied to Landsat images from the vegetation monitoring system dataset (Scolforo and Carvalho 2006; Carvalho and Scolforo 2008, Carvalho and Scolforo - *unpublished data*). GRD was normalized by the amount of remaining forest area within each municipality.

To quantify forest fragmentation, we used the 2011 land cover map from the vegetation monitoring system dataset (Scolforo and Carvalho 2006; Carvalho and Scolforo 2008, Carvalho and Scolforo - *unpublished data*). A set of 225 landscape metrics from class and landscape levels from all of the different categories available in FragStats 4.0 (McGarigal et al. 2012) were calculated for each of the 518 municipalities considering the forest cover configuration in 2011. These were then passed through a three-stage filtering process to provide a tractable set of metrics for use in our analysis of statistical approaches. Firstly, noting that metrics in datasets such as this can be highly correlated (Riitters et al. 1995), we selected a subset of uncorrelated metrics based on *Pearson* correlation analyses conducted using the Pairs-panel analyses in R. We discarded those metrics which were strongly correlated (defined for these purposes as having correlation coefficients for which $p \leq 0.01$) with selected variables, and therefore deemed to be redundant. When two or more variables were significantly correlated, the selection criteria to choose one of them were mathematical simplicity and an intuitive judgment of their explanatory power in terms of ecological meaning. Secondly, we chose metrics from the remaining subset that were commonly used in literature (those which were repeatedly found in the papers consulted) found via a search on the Web of Knowledge website (http://wok.mimas.ac.uk/). The search was carried out from 2011 to June 2013, using the key-words "landscape metrics" and/or "landscape indices". This search yielded 48 papers, of which four were found, on inspection, to be out of scope, and we had no access to another five. The papers consulted in the review can be seen in the Supplementary material (List S1 – ESM1). Finally, we verified the normality of the residuals from linear models (see the section *Stepwise multiple linear regression* for more details) and those metrics which had non-normally distributed residuals were discarded to enable comparative analysis of the random forest method with classical, parametric multiple regression, which requires normally distributed variables most of the times. At the end of the three-stage filtering process , three landscape metrics representing forest fragmentation at municipality scale were selected: the total remaining forest (CA), a measure of forest cover; the mean Euclidean nearest-neighbour distance (ENN), a measure of patch's isolation from each other; and the patch density (PD), a measure of forest spatial structure (Table 1).

The S-E and BGP factors were derived from the ecologic-economical zoning of Minas Gerais, ZEE-MG (Scolforo et al. 2008). Almost all available factors were derived within political administrative units at the scale of municipalities, the smallest administrative units in Brazil. To avoid bias, we chose to use only the metrics that would allow us to analyse them at the municipality scale.

S-E and BGP factors were obtained from the ZEE-MG database, which collates data from different national agencies. The years for which these variables were collected were limited by the availability of information from national agencies, and ranged from 2003 to 2006. Based on data availability, socio-economic factors from four categories – production, exploration, human and institutional – were used. Variables from further four categories of BGP factors – topography, distance, accessibility, and geographical location – were also selected. This gave an initial list of more than 300 candidate independent factors. Descriptions of how these variables were calculated can be found in Scolforo et al. 2008. From this list, a tractable sub-set of factors was derived using the first step from the filtering process described above for the FF metrics. As a result, a total of 34 S-E and BGP factors were selected as factors for use in our comparative analysis of statistical approaches (see Table S2, in the supplementary material, for a complete description of all factors).

2.3 Random forest analysis (RF)

Random forest analysis is a machine-learning technique that may be used for predictive modelling of multiple outcomes from large input datasets. In short, RF uses an ensemble of decision trees with binary divisions, each capable of producing an outcome when presented with a set of input values (Cutler et al. 2007). For regression modelling problems the tree response is an estimate of dependent (outcome) variable values derived from the given values of a set of independent (input) variables. RF uses a regression tree approach (also known as "CART"; Breiman et al. 1984), to build a number of decision tree models from randomly selected subsets of training samples and factors (Cutler et al. 2007). Model fitness is examined using validation data that is not in the training sub-sample; hence, cross-validation with external data is not necessary. The validation sample is also used to calculate measures of variable relative importance (Ellis et al. 2012). The outcomes from all of the trees are then averaged, which provides predictive accuracy and low bias (Breiman 2001).

We used the R "extendedForest" library provided by the Gradient Forest project (Smith et al. 2011; Ellis et al. 2012) to carry out RF analysis. This package was developed for use in ecological studies of species distributions. It integrates results from RF analyses for a number of individual species distributions into results that enable prediction of multiple species distributions (Smith et al. 2011; Ellis et al. 2012). In addition, it is able to analyse large numbers of potential factors and to reduce bias when predictors are correlated (Smith et al. 2011). In our study, we extended the application of extendedForest by using the DF and FF metrics described above (i.e. GDR, ENN, CA and PD) in place of the species distributions used in the application for which it was originally developed. We build partial dependence plots using the variable relative importance values. Models were fitted with 10,000 trees. In each split, we used one-third of the factors randomly sampled as independent candidates. We excluded from final models the variables with negative relative importance values, which do not contribute to the overall explanation. In order to test our first hypothesis, we also calculated the R2 in RF approach to compare it with outcomes from the stepwise multiple linear regression.

2.4 Stepwise Multiple Linear Regression

From a wide range of possible approaches, we selected stepwise multiple linear regression (hereafter, STEP) as a comparator method against which to assess the performance of RF. This type of technique is arguably the most common approach to data-based prediction and simulation tasks (Whittingham et al. 2006). For situations in which the number of variables is high, as is the case here, it is appropriate to incorporate into the modelling process a method for selecting only those factors that contribute most strongly to the predictive model delivered. The STEP approach to multiple regression is a routine technique for achieving this (see, for example, Efroymson 1960; Hocking 1976; Furundzic 1998). Despite having a number of weaknesses, notably bias in parameter estimation, inconsistencies among model selection algorithms, and an inappropriate focus on a single best model (Burnham and Anderson 2002; Kadane and Lazar 2004; Whittingham et al. 2006), it is used widely within ecology and landscape studies (Whittingham et al. 2006).

The stepwise method combines forward selection and backward elimination procedures (Venables and Ripley 2002; James et al. 2013). It proceeded by first setting up an initial model incorporating a subset of the candidate independent variables (factors). Then, this model was iteratively altered by adding significant factors and/or removing insignificant ones, in a process called the stepping procedure. A variable that enters at an early stage may become superfluous at later stages because of its relationship with other factors subsequently added to the model (Kleinbaum et al. 1998). To check this possibility, at each step a partial F test is carried out for each factors currently in the model, regardless of the stage at which it was entered. The whole process is repeated until no more factors can be added or removed, which means that the model is optimized, or when a specified maximum number of steps is reached. Many statistical methods are available to test the stability and validity of the final regression model. We used the adjusted square of the correlation coefficient (adjusted R^2) and the AIC (Akaike Information Criteria) to assess our final model. The AIC was also used to calculate relative variable importance. Implementation was based on the dredge function for automated model selection, which is available as the R “MuMIn” package (Barton 2014). It calculates AIC values for models with all possible combinations of factors and ranks the models based on the calculated values. MuMin is also highly demanding in terms of computational time and resource requirements. We determined the relative importance of each independent variable selected in the models from STEP approach based on AIC weights (importance function in MuMIn; Burnham and Anderson 2002). The relative importance values were converted to percentages for comparison with the equivalent outcomes from RF.

2.5 Final models

We used specific acronyms for the models we have tested to make it easier for readers to understand them. For this, we use the acronyms of each of the metrics tested, which reflect deforestation (DF): GRD; and forest fragmentation (FF): CA, ENN and PD and we add the acronym of the two analysis approaches that we used: RF and STEP. The results were four models selected using RF approach and four other using STEP approach, respectively: the growth rate of deforestation – RF-GRD and STEP-GRD; the total remaining forest – RF-CA and STEP-CA; the mean Euclidean nearest-neighbour distance – RF-ENN and STEP-ENN; and the patch density – RF-PD and STEP-PD models.

3 Results

3.1 Random forest analysis

The RF analysis provides evidence of the effect SE and BGP factors (see Table S2 in the supplementary material ESM2) on the metrics, explaining high amounts of the observed variance (up to 99%) of some of them, and lower amounts of the observed variance of others (less than ~ 40%) (Fig 2 - see also Table S3 in the supplementary material ESM3). In the latter cases, the outcomes imply that there is restricted explanatory power in the factors, and that variability in some of the models across the municipalities is not explained by the factors considered here. The relative importance of each factor was quantified as its partial contribution to explaining the variability of each of the four metrics tested by both statistical approaches, expressed as a percentage. Although, these values are not quantitatively comparable between the metrics, they allow us to rank the factors in terms of their relative importance in each metric model.

#Fig 2 approximately here

Of the four models using RF approach, RF-GRD performed best, with a very high value (99.40% - Fig 2) of its variance explained by the factors. Distance variables (longitude and the minimum distance of forest patches to the nearest reservoir and the nearest protected area) and geographical location were the most important factors in this respect. Among the many factors selected in GRD model selected by RF, those related to topography and crop production were also relatively important. Longitude (POINT_X) explained a greatest part of the variance in RF-GRD model (Fig 3.a)

#Fig 3 approximately here

The selected patch density model (RF-PD), had the second highest amount of its variation explained (61.52%, Fig 2). A large number of factors were identified as having some role in explaining RF-PD variations between municipalities; those with the highest importance were associated with the road network or were topographic. Roads density was the factor which most explained the variance in this model (Fig 3.b).

The selected models of total remaining forest (RF-CA) and of the mean Euclidean nearest-neighbour distance between forest patches (RF-ENN) also had relatively-high amounts of their variation explained (40.67 and 39.38%, respectively, Fig 2). The factors with the highest importance for predicting these models were the mean slope of each municipality (Fig 3.c) for the selected RF-CA model and the mean altitude of each municipality (Fig 3.d) for the selected RF-ENN model. Other topographic factors (the mean altitude across each municipality for RF-CA, and the mean slope across each whole municipality, and the mean slope within deforested areas, for RF-ENN) were also relatively important, as were geographical location, distances to the nearest protected area and nearest steel mill, and longitude.

Overall, factors from the geographical location, distance, topography, institutional and accessibility categories appeared among the most important factors in all the four selected models from RF approach, namely: the latitude of municipalities; the minimum distance from forest patches to the nearest steel mill and the longitude of municipalities; mean slope, mean slope within deforested areas and mean altitude; the amount of protected area in each municipality; and the density of roads.

3.2 Comparisons of RF with STEP

Outcomes from the STEP approach are shown alongside those for RF, in as comparable a form as possible (Fig 2). Note that, although "percentage importance" values are quoted for models from both analysis approaches, these values are not quantitatively comparable between these two methods' outcomes or between different metrics addressed in models. Rather, these values allow us to rank the factors in terms of their relative importance for explaining the variability of each model. The percentages of variance explained by the two analysis approaches are, however, comparable. Both approaches provided evidence of relevant relationships, but models from RF approach surpassed the capacity of the classical approach in explain models' variance. However, the results are mixed in terms of the factors selected as being most important by each approach.

The selected STEP-CA model performed best of all models from STEP approach. It explained an amount (39.80% c.f. 40.67% for RF-CA) of CA variation between municipalities similar to that explained by RF. There was also a strong similarity between the most important factors selected by the models from both approaches, since all of the factors selected by STEP were also selected by RF, except soil types and employability. The mean slope was the most important factor explaining the selected models from both approaches. Other important factors were latitude, longitude and mean altitude. The amount of protected area in each municipality and the number of rural family farms were also important in STEP-CA.

STEP-ENN had the second highest value of ENN explained variance f (30.91% by STEP-ENN, 39.38% by RF-ENN). Factors were less similar between ENN models than in the CA models. While the mean altitude was the most important factor found by RF-ENN, four factors were important in the STEP-ENN selected model, namely: the mean slope, soil type, density of roads and latitude.

The selected PD model from STEP approach (STEP-PD) also had a relatively high amount of its variance explained compared to the other models from STEP, but much less than the selected RF-PD model (29.40% c.f. 61.52% for RF-PD). Some of the factors were found in the selected models from both approaches. However, only one of the most important factors appeared in both of these models: the mean slope of deforestation patches, a topographic factor. The density of roads was the factor identified as being most important by RF-PD, while a similar factor, the minimum distance to the nearest road had the highest importance in STEP-PD. Another topographic factor important in the STEP-PD was the minimum mean slope within each municipality, while in RF-PD the mean altitude, and latitude were also important.

There was a strong contrast between the amounts of variance explained for the growth rate of deforestation by STEP (17.36%) and RF (99.4%) approaches. In STEP-GRD, the minimum distances to the nearest protected area and nearest steel mill were the most important factor explaining GRD variance, followed by the mean slope and the amount of protected area. In RF-GRD, the longitude and, secondarily, the latitude and minimum distances to the nearest steel mill and nearest reservoir were also important.

4 Discussion

4.1 Random Forest analysis

In the RF approach' outcomes, we observed that there are some strong relationships between the S-E and BGP factors and DF and FF metrics. RF performed best for the growth rate of deforestation (RF-GRD) and secondarily for patch density (RF-PD) selected models, explaining around 99% and 60% of their variances, respectively – high values for ecological studies. It also performed relatively well for the total remaining forest (RF-CA) and patch isolation mean Euclidean nearest neighbour distance (RF-ENN) selected models, explaining 40.67% and 39.38% of their variances, respectively. In terms of model performance, this may suggest that the random forest approach is good at identifying parameters that describe some macro-scale factors (rate of deforestation and the overall remaining forest) and the distribution of patches within a landscape (their density and mean isolation from each other. Alternatively, these results could be interpreted as indicating that the rate of deforestation, remaining forest and patch-distribution scale variables (GRD, PD, CA and ENN) are closely linked to the factors we have considered here. In other words, RF is particularly good at identifying links for the types of parameters we analyse, since it performs better providing a higher amount of metrics variance explanation. It is important to note that, even using a very large dataset comprising many factors, much of the variance in some of the four metrics was not accounted by our selected models. In addition, the question of whether it is primarily the nature of the model or the nature of the factors that has led to this finding is not answerable by this first application of RF to this type of data, and remains to be addressed by further investigation.

Turning now to consideration of the factors, we found that some of them were particularly strongly related to some of the metrics, for example longitude (which explained 20.7% of GRD variance), road density (which explained 20.4% of PD), and mean altitude (which explained 18.5% of ENN). However, neither the nature of, nor the reason (i.e. whether they are causatively-linked or simply co-vary) for these links are elucidated by RF. Despite these cases of strong individual-variable links, no single independent variable was found to be related to all of the metrics. Geist and Lambin (2002), who investigated the causes of deforestation of tropical forests, also did not find a single important factor. They concluded that forest loss is due to a combination of factors that vary with historical and geographical context. We conclude from the present study that we can expect the same for forest fragmentation.

At the level of independent variable categories and considering only the three variables in each model which made the strongest contributions the explain metrics variance, we found that those from the Geographical location, Topography, Distance and Accessibility categories contributed most to explaining variance in the RF outcomes. On the other hand, variables from the Exploration, Institutional, and Productivity categories made hardly any contribution. Additionally, we found that factors from the Geographical location and Topography categories made up the majority of the most-important independent variable explaining each dependent variable in the models from RF approach. This suggests that the physical environment is more important for determining variations in DF and FF metrics between municipalities, than social or economic issues. Other studies conducted in the Atlantic Forest agree with our results, showing that physical environment factors play a significant role on deforestation and forest fragmentation (Silva et al. 2007; Teixeira et al. 2009; Freitas et al. 2010). In other countries of Latin America, a similar pattern can be also observed, with physical environment being more important than socioeconomic or demographic factors to explain land-cover change (Bonilla-Moheno et al. 2012; Redo et al. 2012). In addition, specifically in our case, geographical location is important considering the discrepancies between the north and south parts of the study area, mainly in terms of development, what also could work as a proxy of some socioeconomic and demographic factors. However, these findings do not exclude the contribution of socioeconomic or demographic factors upon deforestation and forest fragmentation, since they might be indirectly linked to the physical environment factors. For example, deforestation is more likely to be located in lower and less steep terrain, where transport and mechanical agriculture are easier (Apan and Peterson 1998). They are more likely to have occurred in sites more suitable for agriculture (Flamenco-Sandoval et al. 2007; Killeen et al. 2007; Fearnside 2016). This finding has important implications for management policies aimed at conserving the Atlantic forest, and possibly other biomes that are fragmenting under anthropogenic pressures, although it requires further evidence to be confirmed. This points out the importance of valuing biodiversity in impacted sites (lower and less steep terrain) when selecting areas for conservation, for example (Margules and Pressey 2000; Metzger and Casatti 2006). Also, although this ordering of importance of the different types of factors is quite coherent across the RF approach' outcomes, the question remains as to whether it is "true". Claims to this effect are supported by noting that factors that random forest-type methods have identified as most important for classification have been found to coincide with ecological expectations in the literature (Cutler et al. 2007; Wei et al. 2010; Ellis et al. 2012).

4.2 Comparisons of RF with STEP

Like RF, the STEP approach found some strong relationships between the S-E/BGP factors and DF/FF metrics. Unlike RF, STEP selected models found the most explained-variance and strongest relationships for the amount of forest, followed by the isolation of forest patches. Unlike RF, however, there was less difference in the performances of models from STEP approach: while the explained variances from RF ranged from ~40% to 99%, STEP explained between ~18 and 40% of the variance of all four metrics, confirming our first hypothesis, that RF addresses the relationships between factor and metrics more accurately than STEP approach.

Contrary to our second hypothesis, there was more disagreement than agreement, overall, in terms of the selection and importance of the factors between the two approaches. A low number of factors was selected as important and shared by them. Considering the categories of factors, both approaches found that factors from the Topography category were of higher importance in all selected models, while the Geographical location was more important in the selected models from RF than from STEP approach. Variables from the Distances and Accessibility categories were of intermediate importance, and variables from the Exploration, Institutional, and Production categories were of little importance. In the selected models from STEP approach, we found that the most-important independent variable explaining each dependent variable model also belonged to the Distances and Topographic categories.

The most important factors of selected models in RF approach were subtly different than those selected in STEP approach. Considering the selected rate of deforestation model from RF, the most important factor influencing it is longitude of municipalities, which represents a measure of the distance from the ocean (climate) and also to socioeconomic longitudinal gradient. We expected that deforestation increases towards a socioeconomic gradient, which may reflect a higher degree of developed, and consequently, higher exploration of natural resources, for example. On the other hand, the most important factors in the selected model from STEP were the minimum distance to the protected area. In a similar way, we expected that deforestation decreases when forest patches are closer to natural reserves. The smaller the distance, the closer the forest patches are to a natural reserve. This may mean that there is a greater amount of forest in the municipalities where the forest patches are closer to the natural reserves, whereas in those municipalities where the reserves are more distant, there is possibly a smaller amount of forest, and therefore, deforestation rate is also smaller. Although different, these two factors may be ecologically linked to deforestation rates.

Turning to isolation of forest patches, two different factors from the Topography category appeared as most important factors in the selected models from RF and STEP, respectively, the mean altitude of each municipality and the mean slope across each whole municipality. Although different measurements, these factors are related to the relief of the study area, that plays an important role influencing deforestation (Silva et al. 2007) in The Atlantic Forest Biome. Also, due to an intense exploration in the last 500 year, the Atlantic forest remnants are currently restrict to the higher elevations and steeper reliefs (Dean 1996; Oliveira-Filho and Fontes 2000; Ribeiro et al. 2009; Kauano et al. 2012). The most important factor was the same for the amount of forest in both selected models from RF and STEP approaches: mean slope; also related to the study area relief.

The density of forest patches was mostly affected by two similar factors: the density of roads in the selected model from RF; and the minimum distance to the nearest road in the selected model from STEP. These findings are consistent, since roads serve as fragmenting features (Forman and Alexander 1998; Butler et al. 2004), subdividing forests, increasing the number of forest patches, and reducing forest connectance. Roads have few positive, neutral and numerous negative environmental impacts. Positive impacts include increasing accessibility (Leinbach 1995), which can also be negative since this facilitates deforestation (Laurance et al. 2001). Negative impacts include habitat loss, degradation, and fragmentation, direct wildlife mortality, and road avoidance behaviours by wildlife (Forman and Alexander 1998). Therefore, density of roads plays an effective role in forest fragmentation, and the minimum distance to the nearest road also reflects this role.

Notwithstanding a few similarities between the outcomes of the two modelling approaches, differences between them are strongly evident. However, the reasons for these differences are not clear from our results, and require further investigation. Nonetheless, in theory, one would expect the RF approach' outcomes to identify more reliably than STEP the factors that have greatest influence over models. This expectation arises from the greater robustness of random-forest type methods compared to traditional regression approaches. Unlike traditional regression, which has well known weaknesses, despite still being widely used in ecology (Whittingham et al. 2006), random forest methods make no assumptions about the distributions of variables and are robust to outliers in factors. They can also handle situations where the number of factors exceeds the number of observations and have a novel variable importance measure, which does not suffer the shortcomings of traditional variable selection methods, such as selecting only one or two variables among a group of equally good but highly correlated predictors (Cutler et al. 2007). Thus, the greater range of values of explained variance in the RF outcomes compared to the STEP outcomes may be indicative of their greater robustness and ability to distinguish meaningfulness relationships. Furthermore, many studies that have applied classical regression approaches to understand the drivers of forest cover changes (e.g. Jaimes et al. 2010; Gao and Li 2011; Freitas et al. 2013; Gong et al. 2013) may have had to use a restricted number of factors to be able to satisfy requirements of normality, which could have hindered the analyses, whereas the flexibility and robustness of RF overcomes such limitations.

Despite its advantages, RF used to be one main limitation. Unlike traditional regression methods, RF did not produce relationships between independent and metrics that have simple representations (such as linear equations), making ecological interpretation difficult (Cutler et al. 2007). Nevertheless, the R "extendedForest" library has overcome this issue. This package allows us to generate partial plots, which indicate the direction and form of the independent response of a variable. Therefore, we can now convert the RF outcomes into equations for quantitatively predicting changes in DF and FF metrics that might arise from changes in the BGP and S-E factors considered here. Additionally, RF has exploited structure in our high-dimensional data set not "visible" to STEP in the GRD and PD selected models to provide an apparently clearer picture of these metrics' relationships to the factors.

5 Conclusion

Understanding spatial relationships between patterns of DF/FF metrics and S-E/BGP factors is important for land use management. The main contribution of this study is the testing of a relatively new application of RF for detecting this kind of relationship, its application to a very large dataset, and its comparison with a traditional multiple linear regression method. We found that RF performs better than multiple regression at explaining metrics describing forest patch patterns (PD and ENN) and broader landscape structures (GRD and CA). Given the well-established advantages of decision-tree-based methods over those of classical multiple regression (Breiman et al. 1984; Breiman 2001; Prasad et al. 2006; Cutler et al. 2007; Cutler et al. 2008; Pitcher et al. 2011; Ellis et al. 2012; Cutler 2013; Smith et al. 2013), we suggest that the reasons for these differences are likely to be because the patch-pattern metrics and broader landscape structures vary in less smooth or monotonic ways (McGarigal et al. 2012) – ways that RF is able to capture, but multiple regression is not. Still, we have shown that RF provides a promising methodology for identifying these relationships, and that it has the potential to be an effective tool for providing essential information for aiding land use management decisions, not only in terms of planning, but also for conservation actions, as proposed by Zanella et al. (2012), in cases of high rates of anthropogenic biodiversity loss, as it is the case of the Atlantic Forest.

The initial investigation reported in the present study is, however, only a first step in exploiting this method's potential. One aspect that requires further consideration is the scale of the study area and the very wide variety of S-E and BGP contexts, which it encompasses. Even in relatively small areas, a multitude of diverse factors are at work (Qasim et al. 2013), and variations in contexts may have influenced model performance in the present study. Landscape pattern is scale-sensitive (Gao and Li 2011) and the unusually large degree of heterogeneity in the Atlantic forest biome is likely only to exacerbate this issue. Policies need to be crafted at appropriate spatial scales and with specific contexts in mind. Thus, an important development of this initial study of RF application to cases of DF and FF would be to repeat it at different spatial scales, to identify more precisely the S-E and BGP factors associated with these processes.

References

Aide TM, Grau HR (2004) Globalization, Migratin, and Latin American Ecosystems. Science (80-) 305:1915–1916. doi: 10.1126/science.1103179

Apan AA, Peterson JA (1998) Probing tropical deforestation: the use of GIS and statistical analysis of georeferenced data. Appl Geogr 18:137–152.

Barbier EB, Burgess JC, Grainger A (2010) The forest transition: Towards a more comprehensive theoretical framework. Land use policy 27:98–107. doi: 10.1016/j.landusepol.2009.02.001

Barton K (2014) Multi-model inference. R Packag MuMIn version 1105 46.

Beilin R, Lindborg R, Stenseke M, et al (2014) Analysing how drivers of agricultural land abandonment affect biodiversity and cultural landscapes using case studies from Scandinavia, Iberia and Oceania. Land use policy 36:60–72. doi: 10.1016/j.landusepol.2013.07.003

Bonilla-Moheno M, Aide TM, Clark ML (2012) The influence of socioeconomic, environmental, and demographic factors on municipality-scale land-cover change in Mexico. Reg Environ Chang 12:543–557. doi: 10.1007/s10113-011-0268-z

Breiman L (2001) Random forests. Mach Learn 45:5–32.

Breiman L, Friedman J, Stone CC, et al (1984) Classification and regression trees. Wadsworth, Belmont Calif.

Burnham K, Anderson D (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. New York

Butler BJ, Swenson JJ, Alig RJ (2004) Forest fragmentation in the Pacific Northwest: Quantification and correlations. For Ecol Manage 189:363–373. doi: 10.1016/j.foreco.2003.09.013

Caldas MM, Goodin D, Sherwood S, et al (2013) Land-cover change in the Paraguayan

Chaco: 2000–2011. J Land Use Sci 4248:1–18. doi: 10.1080/1747423X.2013.807314

Carvalho LMT de, Scolforo JR (2008) Inventário Florestal de Minas Gerais: Monitoramento da Flora Nativa 2005-2007. Editora da UFLA, Lavras

Cutler A (2013) Trees and Random Forests. NIH 1R15AG037392-01.

Cutler A, Cutler DR, Stevens JR (2008) Tree-based methods. In: Xiaochun Li, Xu R (eds) High-Dimensional Data Analysis in Cancer Research. Springer New York, pp 89–109

Cutler A, Stevens J (2006) Random Forests for microarrays. Methods Enzymol 422–432.

Cutler DR, Edwards Jr. TC, Beard KH, et al (2007) Random Forests for Classification in Ecology. Ecology 88:2783–2792. doi: 10.1890/07-0539.1

Dean W (1996) With broadax and firebrand: the destruction of the Brazilian Atlantic Forest. California

DeFries RS, Foley JA, Asner GP (2004) Land-use choices: Balancing human needs and ecosystem function. Front Ecol Environ 2:249–257. doi: 10.1890/1540-9295(2004)002[0249:LCBHNA]2.0.CO;2

Efroymson M (1960) Multiple regression analysis. Math methods Digit Comput 191–203.

Ellis N, Smith SJ, Pitcher CR (2012) Gradient forests: calculating importance gradients on physical predictors. Ecology 93:156–68.

Farinaci JS, Batistella M (2012) Variação na cobertura vegetal nativa em São Paulo: um panorama do conhecimento atual. Rev Árvore 36:695–705. doi: 10.1590/S0100-67622012000400011

Fearnside P (2016) The Roles and Movements of Actors in the Deforestation of Brazilian Amazonia The Roles and Movements of Actors in the Deforestation of Brazilian Amazonia. doi: Artn 23

Ferreira MP, Alves DS, Shimabukuro YE (2015) Forest dynamics and land-use transitions in the Brazilian Atlantic Forest: the case of sugarcane expansion. Reg Environ Chang

15:365–377. doi: 10.1007/s10113-014-0652-6

Flamenco-Sandoval A, Martínez Ramos M, Masera OR (2007) Assessing implications of land-use and land-cover change dynamics for conservation of a highly diverse tropical rain forest. Biol Conserv 138:131–145. doi: 10.1016/j.biocon.2007.04.022

Foley JA (2005) Global Consequences of Land Use. Science (80-) 309:570–574. doi: 10.1126/science.1111772

Forman RTT, Alexander LE (1998) Roads and Their Major Ecological Effects. Annu Rev Ecol Syst 29:207–231. doi: 10.1146/annurev.ecolsys.29.1.207

Freitas MWD, Santos JR Dos, Alves DS (2013) Land-use and land-cover change processes in the Upper Uruguay Basin: linking environmental and socioeconomic variables. Landsc Ecol 28:311–327. doi: 10.1007/s10980-012-9838-9

Freitas SR, Hawbaker TJ, Metzger JP (2010) Effects of roads, topography, and land use on forest cover dynamics in the Brazilian Atlantic Forest. For Ecol Manage 259:410–417. doi: 10.1016/j.foreco.2009.10.036

Fu W, Liu S, Degloria SD, et al (2010) Characterizing the “fragmentation–barrier” effect of road networks on landscape connectivity: A case study in Xishuangbanna, Southwest China. Landsc Urban Plan 95:122–129. doi: 10.1016/j.landurbplan.2009.12.009

Furundzic D (1998) Application example of neural networks for time series analysis : Rainfall — runoff modeling. 64:383–396.

Gao J, Li S (2011) Detecting spatially non-stationary and scale-dependent relationships between urban landscape fragmentation and related factors using Geographically Weighted Regression. Appl Geogr 31:292–302. doi: 10.1016/j.apgeog.2010.06.003

Gaughan AE, Binford MW, Southworth J (2009) Tourism, forest conversion, and land transformations in the Angkor basin, Cambodia. Appl Geogr 29:212–223. doi: 10.1016/j.apgeog.2008.09.007

Geist HJ, Lambin EF (2001) What drives tropical deforestation? LUCC Report Series No. 4. Louvain-la-Neuve

Geist HJ, Lambin EF (2002) Proximate Causes and Underlying Driving Forces of Tropical Deforestation. Bioscience 52:143–150. doi: 10.1641/0006-3568(2002)052[0143:PCAUDF]2.0.CO;2

Gilbert A, Chakraborty J (2011) Using geographically weighted regression for environmental justice analysis: Cumulative cancer risks from air toxics in Florida. Soc Sci Res 40:273–286. doi: 10.1016/j.ssresearch.2010.08.006

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random Forests for land cover classification. Pattern Recognit Lett 27:294–300. doi: 10.1016/j.patrec.2005.08.011

Gong C, Yu S, Joesting H, Chen J (2013) Determining socioeconomic drivers of urban forest fragmentation with historical remote sensing images. Landsc Urban Plan 117:57–65. doi: 10.1016/j.landurbplan.2013.04.009

Hecht SB, Kandel S, Gomes I, et al (2006) Globalization, forest resurgence, and environmental politics in El Salvador. World Dev 34:308–323. doi: 10.1016/j.worlddev.2005.09.005

Hocking RR, Mar N (1976) A Biometrics Invited Paper . The Analysis and Selection of Variables in Linear Regression. 32:1–49.

IBGE (2017) Estados. Minas Gerais. http://www.ibge.gov.br/estadosat/perfil.php?sigla=mg. Accessed 15 Jun 2017

Jaimes NBP, Bosque Sendra J, Franco R, et al (2010) Exploring the driving forces behind deforestation in the state of Mexico (Mexico) using geographically weighted regression. Appl Geogr 30:576–591. doi: 10.1016/j.apgeog.2010.05.004

James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning: with Applications in R. New York

Kadane JB, Lazar N a (2004) Methods and Criteria for Model Selection. J Am Stat Assoc 99:279–290. doi: 10.1198/016214504000000269

Kauano ÉE, Torezan JMD, Cardoso FCG, Marques MCM (2012) Landscape structure in the northern coast of Paraná state, a hotspot for the brazilian Atlantic Forest conservation. Rev Árvore 36:961–970. doi: 10.1590/S0100-67622012000500018

Killeen TJ, Calderon V, Soria L, et al (2007) Thirty years of land-cover change in Bolivia. Ambio 36:600–6. doi: 10.1579/0044-7447(2007)36[600:TYOLCI]2.0.CO;2

Kleinbaum D, Kupper L, Nizam A, Rosenberg E (1998) Applied regression analysis and other multivariable methods, third. Duxbury Press., Pacific Grove - CA

Lambin EF, Geist HJ (2006) Land-Use and Land-Cover Change: Local Processes and Global Impacts, 1st edn. Springer Verlag, Berlin

Laurance WF, Cochrane MA, Bergen S, et al (2001) The Future of the Brazilian Amazon. Science (80-) 291:438–439. doi: 10.1126/science.291.5503.438

Leal CGCG, Pompeu PS, Gardner TA, et al (2016) Multi-scale assessment of human-induced changes to Amazonian instream habitats. Landsc Ecol 31:1725–1745. doi: 10.1007/s10980-016-0358-x

Leinbach TR (1995) Transport and third world development: review, issues, and prescription. Transp Res Part A Policy Pract 29:337–344. doi: 10.1016/0965-8564(94)00035-9

Lira PK, Ewers RM, Banks-Leite C, et al (2012) Evaluating the legacy of landscape history: extinction debt and species credit in bird and small mammal assemblages in the Brazilian Atlantic Forest. J Appl Ecol 49:1325–1333. doi: 10.1111/j.1365-2664.2012.02214.x

Malhi Y, Gardner T a., Goldsmith GR, et al (2014) Tropical Forests in the Anthropocene. Annu Rev Environ Resour 39:125–159. doi: 10.1146/annurev-environ-030713-155141

Margules CR, Pressey RL (2000) Systematic conservation planning. Nature 405:243–53. doi:

10.1038/35012251

McGarigal K, Cushman S, Ene E (2012) FragStats v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps Computer software program produced by the authors at the University of Massachusetts. http://www.umass.edu/landeco/research/fragstats/fragstats.html. Accessed 1 May 2016

Metzger JP, Casatti L (2006) Do diagnóstico à conservação da biodiversidade : o estado da arte do programa BIOTA / FAPESP. 6:1–23.

Oliveira-Filho A, Fontes M (2000) Patterns of Floristic Differentiation among Atlantic Forests in Southeastern Brazil and the Influence of Climate. Biotropica 32:793–810. doi: 10.1111/j.1744-7429.2000.tb00619.x

Oliveira VHF, Barlow J, Gardner T, Louzada J (2017) Do we select the best metrics for assessing land use effects on biodiversity? Basic Appl Ecol. doi: 10.1016/j.baae.2017.03.002

Parcerisas L, Marull J, Pino J, et al (2012) Land use changes, landscape ecology and their socioeconomic driving forces in the Spanish Mediterranean coast (El Maresme County, 1850–2005). Environ Sci Policy 23:120–132. doi: 10.1016/j.envsci.2012.08.002

Parés-Ramos IK, Gould WA, Aide TM (2008) Agricultural abandonment, suburban growth, and forest expansion in Puerto Rico between 1991 and 2000. Ecol Soc. doi: 1

Perz SG (2004) Are agricultural production and forest conservation compatible? Agricultural diversity, agricultural incomes and primary forest cover among small farm colonists in the Amazon. World Dev 32:957–977. doi: 10.1016/j.worlddev.2003.10.012

Perz SG, Caldas MM, Arima E, Walker RT (2007) Unofficial road building in the Amazon: Socioeconomic and biophysical explanations. Dev Change 38:529–551. doi: 10.1111/j.1467-7660.2007.00422.x

Pfaff ASP (1999) What Drives Deforestation in the Brazilian Amazon? Evidence from

Satellite and Socioeconomic Data*. J Environ Econ Manage 37:26–43. doi: 10.1006/jeem.1998.1056

Pitcher CR, Ellis N, Smith SJ (2011) Example analysis of biodiversity survey data with R package gradientForest Gradient Forest basics. 1–16.

Prasad AM, Iverson LR, Liaw A, et al (2006) Newer Tree Classification and Techniques : Forests Random Prediction Bagging for Ecological Regression. Ecosystems 9:181–199. doi: 10.1007/S10021-005-0054-1

Qasim M, Hubacek K, Termansen M (2013) Underlying and proximate driving causes of land use change in district Swat, Pakistan. Land use policy 34:146–157. doi: 10.1016/j.landusepol.2013.02.008

Quezada ML, Arroyo-Rodríguez V, Pérez-Silva E, Aide TM (2013) Land cover changes in the Lachuá region, Guatemala: patterns, proximate causes, and underlying driving forces over the last 50 years. Reg Environ Chang 14:1139–1149. doi: 10.1007/s10113-013-0548-x

Redo DJ, Aide TM, Clark ML (2012) The Relative Importance of Socioeconomic and Environmental Variables in Explaining Land Change in Bolivia, 2001–2010. Ann Assoc Am Geogr 102:778–807. doi: 10.1080/00045608.2012.678036

Ribeiro MCM, Metzger JPJJP, Martensen AC, et al (2009) The Brazilian Atlantic Forest: How much is left, and how is the remaining forest distributed? Implications for conservation. Biol Conserv 142:1141–1153. doi: 10.1016/j.biocon.2009.02.021

Richards PD, Walkerb RT, Arima EY (2008) NIH Public Access. Glob Env Chang 144:724–732. doi: 10.1038/jid.2014.371

Riitters KH, Neil RVO, Hunsaker CT, et al (1995) A factor analysis of landscape pattern and structure metrics. Landsc Ecol 10:23–39.

Scolforo J, Carvalho LMT de (2006) Mapeamento e inventário da flora nativa e dos reflorestamentos de Minas Gerais, 2nd edn. UFLA

Scolforo JR, Oliveira AD de, Carvalho LMT de (2008) Zoneamento ecológico-econômico do estado de minas gerais: Componente sócioeconômico. UFLA, lAVRAS

Silva WG, Metzger JP, Simões S, Simonetti C (2007) Relief influence on the spatial distribution of the Atlantic Forest cover on the Ibiúna Plateau, SP. Brazilian J Biol 67:403–11.

Smith PF, Ganesh S, Liu P (2013) A comparison of random forest regression and multiple linear regression for prediction in neuroscience. J Neurosci Methods 220:85–91. doi: 10.1016/j.jneumeth.2013.08.024

Smith SJ, Ellis N, Pitcher CR (2011) Conditional variable importance in R package extendedForest.

SOS Mata Atlântica/INPE (2014) Atlas dos remanescentes de Mata Atlântica período 2012–2013. São Paulo, Brazil

Teixeira AMG, Soares-Filho BS, Freitas SR, Metzger JP (2009) Modeling landscape dynamics in an Atlantic Rainforest region: Implications for conservation. For Ecol Manage 257:1219–1230. doi: 10.1016/j.foreco.2008.10.011

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, Fourth edi. Springer-Verlag, New York

Wei C-L, Rowe GT, Escobar-Briones E, et al (2010) Global patterns and predictions of seafloor biomass using random forests. PLoS One 5:e15323. doi: 10.1371/journal.pone.0015323

Whittingham MJ, Stephens P a, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 75:1182–9. doi: 10.1111/j.1365-2656.2006.01141.x

Wright SJ, Samaniego MJ (2008) Historical, demographic, and economic correlates of land-use change in the Republic of Panama.

Zanella L, Borém R, Souza C, et al (2012) Atlantic Forest Fragmentation Analysis and Landscape Restoration Management Scenarios. Nat … 10:57–63.

Table 1 Descriptions of deforestation (DF) and forest fragmentation (FF) metrics (dependent variables)

Metric	**Category**	**Formulae**	**Description (unit)[a]**
Growth rate of deforestation (GRD)	Rate of deforestation	$GRD = \frac{((\mathrm{D}f - \mathrm{D}i) / \mathrm{D}i)}{t}$	Growth rate of deforestation from 2003 to 2011. Di = Total area deforested in 2003. Df = Total area deforested in 2011. t = number of years considered (in our case, eight years). Percentage.
Mean Euclidean Nearest-Neighbour (ENN)	Forest patch isolation	$\mathrm{ENN} = \frac{\sum_{j=1}^{n} h_{ij}}{n_i}$	ENN equals the mean distance to the nearest neighbouring patch of forest, based on shortest edge-to-edge distance. h_{ij} = distance (m) from patch j to nearest neighbouring patch

			of the same type (*i*, in this case forest). n_i = number of patches of cover type *i* (forest).
Total remaining forest (CA)	Remaining forest quantification	$\mathrm{CA} = A\left(\frac{1}{10{,}000}\right)$	CA equals the total area (m^2) of the landscape, divided by 10,000 (to convert to hectares). *A* = total landscape area (m^2). CA is important because it defines the extent of the landscape.
Patch density (PD)	Forest spatial structure	$\mathrm{PD} = \frac{n_i}{A}(10{,}00000)$	Patch density increases with a greater number of patches within a reference area and therefore reflects landscape fragmentation.

[a] Details can be found in Mcgarigal et al. (2012).

Fig 1 Minas Gerais State, BR and the 518 municipalities used in this study. The inset map on the left show the location of Minas Gerais State within Brazil

Fig 2 Relative importance plot for factors from random forest (RF) and stepwise multiple regression (STEP) analysis approaches, in percentage (%). Factors are defined in Table S2 (supplementary material ESM2). The eight selected models from both approaches are: the growth rate of deforestation – RF-GRD and STEP-GRD; the total remaining forest – RF-CA and STEP-CA; the mean Euclidean nearest-neighbour distance – RF-ENN and STEP-ENN; and the patch density – RF-PD and STEP-P. Metrics used in models are defined in Table 1. Note that each model shows only the most important predictors. * Percentage of variance explained in each model

Fig 3 Partial contribution of socio-economic (S-E) and bio-geophysical (BGP) factors to deforestation (DF) and forest fragmentation (FF) in Minas Gerais, Brazil, derived from RF analysis approach. Factors are defined in Table S2 (Supplementary material ESM2). A) The growth rate of deforestation (RF-GRD); B) Patch density (RF-PD); C) The total remaining forest area (RF-CA); and D) The Euclidean nearest-neighbour distance (RF-ENN). Metrics used in models are defined in Table 1