Title

The Latent Dirichlet Allocation model applied to airborne LiDAR data: A case study on mapping forest degradation associated with fragmentation and fire in the Amazon region

Journal

Methods in Ecology and Evolution, 13(6)

ISSN

2041-210X

Authors

Valle, Denis
Silva, Carlos Alberto
Longo, Marcos
et al.

Publication Date

2022-06-01

DOI

10.1111/2041-210x.13836

Methods in Ecology and Evolution

Valle Denis (Orcid ID: 0000-0002-9830-8876)
Longo Marcos (Orcid ID: 0000-0001-5062-6245)
Brando Paulo (Orcid ID: 0000-0001-8952-7025)

Title: The Latent Dirichlet Allocation model applied to airborne LiDAR data: a case study on mapping forest degradation associated with fragmentation and fire in the Amazon region.

Denis Valle[1*], Carlos Alberto Silva[1], Marcos Longo[2], Paulo Brando[3]

[1] School of Forest, Fisheries, and Geomatics Sciences, University of Florida, PO Box 110410 Gainesville, FL 32611

[2] Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA

[3] Department of Earth System Science, University of California Irvine, California, United States of America.

*Corresponding author: Tel: (352) 392-3806; Email: drvalle@ufl.edu

Abstract:

1. LiDAR data are being increasingly used to provide a detailed characterization of the vertical profile of forests. This characterization enables the generation of new insights on the influence of environmental drivers and anthropogenic disturbances on forest structure as well as on how forest structure influences important ecosystem functions and services. Unfortunately, extracting information from LiDAR data in a way that enables the spatial visualization of forest structure, as well as its temporal changes, is challenging due to the high-dimensionality of these data.
2. We show how the Latent Dirichlet Allocation model applied to LiDAR data (LidarLDA) can be used to identify forest structural types and how the relative abundance of these forest types changes throughout the landscape. The code to fit this model is made available through the open-source R package LidarLDA in github. We illustrate the use of LidarLDA both with simulated data and data from a large-scale fire experiment in the Brazilian Amazon region.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.13836

3. Using simulated data, we demonstrate that LidarLDA accurately identifies the number of forest types as well as their spatial distribution and absorptance probabilities. For the empirical data, we found that LidarLDA detects both landscape-level patterns in forest structure as well as the strong interacting effect of fire and forest fragmentation on forest structure based on the experimental fire plots. More specifically, LidarLDA reveals that proximity to forest edge exacerbates the impact of fires, and that burned forests remain structurally different from unburned areas for at least seven years, even when burned only once. Importantly, LidarLDA generates insights on the 3D structure of forest that cannot be obtained using more standard approaches that just focus on top-of-the-canopy information (e.g., canopy height models based on LiDAR data).
4. By enabling the mapping of forest structure and its temporal changes, we believe that LidarLDA will be of broad utility to the ecological research community.

Resumo:

1. Dados de LiDAR sao cada vez mais usados para caracterizar a estrutura vertical da floresta. Essa caracterização permite a geração de novos insights em relação a influência de fatores ambientais e distúrbios antropogênicos na estrutura da floresta e insights em relação a como a estrutura da floresta influencia importantes funções e serviços ecossistêmicos. Infelizmente, a extração de informações de dados de LiDAR de uma maneira que permita a visualização espacial da estrutura da floresta, assim como as mudanças temporais, tem sido desafiador por conta da alta dimensionalidade destes dados.
2. Nós mostramos que o modelo Latent Dirichlet Allocation aplicado a dados de LiDAR (LidarLDA) pode ser usado para identificar tipologias estruturais e para revelar como que a abundancia relativa destas tipologias mudam ao longo da paisagem. O código usado para ajustar o modelo se encontra disponível no pacote do R chamado LidarLDA no github. Nós ilustramos o uso do

LidarLDA tanto com dados simulados quanto com dados empíricos de um experimento de fogo de grande escala na Amazônia Brasileira.

3. Usando dados simulados, nós demonstramos que o LidaLDA identifica bem o número de tipologias florestais assim como sua distribuição espacial e as probabilidades de absorbância. Em relação aos dados empíricos, nós mostramos que o LidarLDA detecta padrões no nível da paisagem em relação a estrutura da floresta assim como um forte efeito da interação entre fogo e fragmentação florestal na estrutura florestal nas parcelas queimadas experimentalmente. Mais especificamente, LidarLDA revela que a proximidade com a borda da floresta aumenta o impacto do fogo e que áreas queimadas permanecem estruturalmente diferentes das áreas não queimadas por pelo menos sete anos, mesmo se estas áreas foram queimadas apenas uma vez. É importante enfatizar que o LidarLDA gera insights na estrutura 3D da floresta que não são obtidos usando abordagens mais comuns que focam apenas em informação oriunda do topo da copa (e.g., modelos de altura de copa baseados em dados de LiDAR).
4. Nós acreditamos que a habilidade de mapear a estrutura da floresta e suas mudanças temporais fará com que o modelo LidarLDA seja de grande utilidade para a comunidade de pesquisas ecológicas.

Keywords: Latent Dirichlet Allocation (LDA), LiDAR, fire, forest fragmentation, Amazon, tropical forests

1. Introduction

Forests provide a wide range of ecosystem services, such as nutrient cycling, flood control, wildlife habitat, timber and non-timber forest products, and carbon sequestration (Jenkins & Schaap, 2018, Mori *et al.*, 2017). Forest structure is a key determinant of several of these ecosystem services (Felipe-Lucia *et al.*, 2018) and, as a result, there has been a long-standing interest in characterizing forest structure, understanding how forest structure is influenced by environmental drivers and anthropogenic activities, and how it in turn influences key ecosystem functions and services (Jucker *et al.*, 2018, Longo *et al.*, 2020). Importantly, changes in forest structure associated with natural or anthropogenic disturbances such as wind, fire, timber or wood fuel harvest, are widespread. For example, forest degradation can account for a substantial fraction of the carbon emissions, sometimes even exceeding the amount of emissions associated with deforestation (Pearson *et al.*, 2017, Vancutsem *et al.*, 2021). Given that forest degradation is likely to increase even more in the future as climate change interacts to exacerbate the effect of human activities (Alencar *et al.*, 2015, Brando *et al.*, 2020), accurate characterization of forest structure and its temporal changes associated with different types of disturbances will become increasingly important to improve the understanding and modeling of these disturbances and their impacts.

A prominent source of high-resolution data of the three-dimensional structure of forests has been airborne light detection and ranging (LiDAR). Unfortunately, efficiently summarizing and extracting all the information on forest structure from LiDAR 3D point cloud data can be challenging. One approach is to calculate summary statistics for grid cells at a given spatial resolution, such as mean and maximum return height, standard deviation of the return heights, and height percentiles (Almeida *et al.*, 2019a, Andersen *et al.*, 2013, Costa *et al.*, 2021, Jucker *et al.*, 2018, Rex *et al.*, 2020, Silva *et al.*, 2017). Another approach consists of first characterizing the vertical structure of forests by calculating leaf area density (LAD) to then describe the vertical and horizontal heterogeneity in LAD with summary statistics (e.g.,

Shannon and Simpson structural complexity indices and LAD for different height intervals) (Almeida *et al.*, 2019a, Almeida *et al.*, 2019b, Carrasco *et al.*, 2019). These LiDAR-derived metrics are then used for multiple purposes. For example, one of the most common uses of these metrics in tropical forests is to predict above-ground biomass (AGB) (Almeida *et al.*, 2019a, Andersen *et al.*, 2013, Costa *et al.*, 2021, d'Oliveira *et al.*, 2012, Rex *et al.*, 2020, Silva *et al.*, 2017). These AGB predictions can be used, for example, to identify areas subject to selective logging and quantify its impacts (Andersen *et al.*, 2013, d'Oliveira *et al.*, 2012, Rex *et al.*, 2020, Silva *et al.*, 2017). Aside from predicting AGB, LiDAR-derived metrics have also been used for predicting wildlife diversity (Carrasco *et al.*, 2019), generating forest parameters for fire behavior models (Riano *et al.*, 2003), and understanding the synergistic effect of proximity to forest edge, fire, and windstorms on tree mortality (Silverio *et al.*, 2019). Unfortunately, the visualization of spatial and temporal changes in forest structure with this plethora of LiDAR-derived metrics is challenging.

One approach to more concisely characterize forest structure is to create forest types (e.g., floodplain and terra-firme forests) from the LiDAR-derived 3D point cloud. Indeed, given the importance of forest structure for multiple ecosystem services and functions, several studies have attempted to classify forest types to enable the development of tailored forest inventory and management strategies. For example, Moran *et al.* (2018) described an approach where dissimilarity was calculated using a random forest algorithm and, based on this dissimilarity metric, hierarchical clustering was used to create groups. Ultimately, this data-driven classification approach led to the creation of 14 meta-classes across approximately 170 thousand ha, enabling an intuitive comparison and assessment of forest structure. Similarly, Adnan *et al.* (2019) developed a methodology that combined hierarchical clustering and classification trees (CART) to create forest structural types and showed how this methodology can be useful to compare forest structure across bioregions. These forest types can also be used to optimize field data collection. For example, Papa *et al.* (2020) used a clustering approach to stratify the forest,

demonstrating how this stratification can result in substantial reduction of the sampling effort required for forest inventory.

Current approaches to creating forest types rely on hard clustering methods for dimension reduction, resulting in a small set of relatively homogeneous clusters, hence simplifying the visualization and interpretation of results. However, hard clustering methods assume that any given site can only belong to a single forest type, thus neglecting that some forest areas can have characteristics that are intermediary between two (or more) forest types. For example, areas along the slope between floodplain and terra-firme forests in the Amazon region are likely to have intermediate forest structure, species composition, and diversity, which may be quite different from the stereotypical floodplain or terra-firme forest (Salm *et al.*, 2015, Wittmann *et al.*, 2006). However, hard classification schemes might impose one of these classes. Indeed, although these hard-clustering approaches have been extensively used by researchers across multiple environmental science fields, few ecological theories predict the sharp delineations implied by these hard clustering methods (Legendre & Legendre, 2012). Importantly, because each site can only belong to a single cluster, hard clustering approaches often have to create many more groups to accommodate transition areas, limiting its ability to effectively reduce data dimensionality, with important consequences for the visualization and interpretation of results (Valle *et al.*, 2018).

The Latent Dirichlet Allocation (LDA) model is a type of unsupervised mixed-membership model (often called grade of membership model) that enables the characterization of sampling units as comprised of a single forest structural type or as a combination of multiple forest types (hereafter just forest type or cluster). This method was originally developed for text-mining applications (Blei *et al.*, 2003) but has since been used in a wide range of fields, such as fraud detection (Xing & Girolami, 2007), extraction of semantic information from satellite imagery (Vaduva *et al.*, 2013), bioinformatics (Liu *et al.*, 2010), microbiology (Hosoda *et al.*, 2020), and ecology (Christensen *et al.*, 2018, Dietzel *et al.*, 2019,

Knott *et al.*, 2019, Muhlfeld *et al.*, 2020, Sommeria-Klein *et al.*, 2019, Valle *et al.*, 2018, Valle *et al.*, 2014). LDA has also been used to model LiDAR data in the past (e.g., Yang & Kang, 2018, Zhiqing *et al.*, 2020). However, differently from the model described here, this past work relied on a version of LDA that does not account for occlusion (i.e., the partial or complete blockage of LiDAR light pulses by different objects such as leaves and branches), a key characteristic for our task of identifying forest types. Furthermore, in this past work, LDA was used only to extract features to help a subsequent classification algorithm instead of using LDA results as the primary outcomes.

In this article, we present a modified version of LDA, called LidarLDA, and show how it can be used to gain novel insights from LiDAR data regarding the vertical structure of forests while accounting for occlusion. We start this article by providing an overview of the proposed methodology. We then illustrate with simulated data how this model can estimate the true number of clusters and can recover the spatial distribution of these clusters. Finally, we showcase the insights this model can generate by applying it to LiDAR data from an area of approximately 1,000 ha in the Brazilian Amazon, part of which was subject to a large-scale (i.e., 150 ha) fire experiment. We finalize this article by discussing potential applications of this approach, current limitations, and priorities for future development of this approach.

2. Material and Methods

2.1. Structure of the Latent Dirichlet Allocation model applied to LiDAR data (LidarLDA)

The proposed model is based on the LDA model adapted for presence/absence biodiversity data described in Valle *et al.* (2018) and Albuquerque *et al.* (2019). To use this model for LiDAR data, data need to be discretized horizontally and vertically. More specifically, a systematic grid with a particular

spatial resolution is created within the area of interest (e.g., 50 x 50 m grid cells) and the height of the returns is discretized by creating multiple vertical layers of constant depth (e.g., 1-m layers).

The data that LidarLDA relies on consist of the number of LiDAR returns within a vertical layer h (i.e., N_{ih}) and the total number of pulses that reach this vertical layer (i.e., $\widetilde{N}_{ih}$) for each grid cell i. Because airborne LiDAR light pulses originate from above the canopy, if we assume that light pulses are vertically oriented, we can calculate $\widetilde{N}_{ih}$ as all returns in grid cell i between the ground and the top of layer h (i.e., $\widetilde{N}_{ih} = \sum_{h'=1}^{h} N_{ih'}$). These data are stored into two matrices of same size, where rows correspond to different grid cells and columns correspond to different vertical layers.

This model assumes that each light pulse j (j=1,…,$\widetilde{N}_{ih}$) in grid cell i that reaches vertical layer h can either be returned ($x_{ijh} = 1$) or not ($x_{ijh} = 0$). Because x_{ijh} is a binary variable, we relied on a Bernoulli distribution and we assume that

$$x_{ijh} | \omega_{ijh} = k \sim Bernoulli(\phi_{kh})$$

where ω_{ijh} is the corresponding latent cluster assignment variable and ϕ_{kh} is a probability parameter. Notice that $\omega_{ijh} = k$ indicates that this particular light pulse was assigned to cluster k. Therefore, this variable determines the subscript of the probability parameter ϕ_{kh}. The vector of parameters $\boldsymbol{\phi_k} = [\phi_{k1}, \phi_{k2}, \dots]$ characterizes the vertical profile of cluster k and, together with the vectors for the other clusters, form the rows of the $\boldsymbol{\Phi}$ matrix.

Because the latent cluster assignment variable ω_{ijh} has to be an integer between 1 and K (the maximum number of clusters specified by the modeler), we assume a categorical distribution. This distribution is a generalization of the Bernoulli distribution and is similar to a multinomial distribution with just a single trial. The main difference is that the categorical distribution models numerical labels (i.e., the latent cluster assignment) whereas a multinomial distribution models a vector full of zeroes except for a single element which is equal to one. Our categorical distribution is given by

$$\omega_{ijh} \sim Categorical(\boldsymbol{\theta_i})$$

where $\boldsymbol{\theta_i}$ is a vector of probabilities that sum to one. The vector $\boldsymbol{\theta_i}$ characterizes grid cell i with the relative abundances of the different clusters.

Because this model is estimated in a Bayesian framework, we complete the specification of this model by adopting the following semi-conjugate priors:

$$\phi_{kh} \sim Beta(\alpha, \beta)$$

and

$$\boldsymbol{\theta_i} \sim TSB(\gamma)$$

where TSB stands for the Truncated Stick-Breaking prior. This prior is defined indirectly. First, we define

$$V_{ik} \sim Beta(1, \gamma)$$

for k=1,…,K-1 whereas V_{iK} is set to one. The parameters $V_{i1}, \dots, V_{iK}$ are then used to calculate θ_{ik} with the following equations:

$$\theta_{i1} = V_{i1}$$

$$\theta_{ik} = V_{ik} \prod_{p=1}^{k-1} (1 - V_{ip}) \text{ for k>1}$$

As described in detail in Valle *et al.* (2021a), the TSB prior enables the automatic selection of the optimal number of clusters if this number is smaller than K. As a result, the use of the TSB prior avoids the standard approach of having to run the model multiple times with different number of clusters to then select the best number using an information criterion (e.g., AIC or BIC). The approach of using information criterion to select the optimal number of clusters can be computationally expensive and has been shown to often lead to an over-estimation of the number of clusters (Casella *et al.*, 2014, Pohle *et al.*, 2017).

Finally, the parameters $\alpha > 0$ and $\beta > 0$ are specified by the modeler and describe the prior beliefs regarding the absorptance probabilities. For example, $\alpha = \beta = 1$ is a common choice because it describes a uniform prior distribution for ϕ_{kh}. Similarly, the parameter $0 < \gamma < 1$ is also specified by

the modeler and controls the amount of sparseness that is a priori expected (i.e., smaller γ values encourage the model to find fewer clusters) (Valle *et al.*, 2021a).

2.2. Data decomposition implied by LidarLDA

One way to better understand this model is to realize that these assumptions are equivalent to:

$$N_{ih} \sim Binomial(\widetilde{N}_{ih}, p_{ih})$$

Notice that p_{ih} is the conditional probability of a return within vertical layer h given that the light pulse has reached this layer. As a result, p_{ih} accounts for the occlusion of the LiDAR light pulses as they pass through the canopy. It is important to also note that we exclude the vertical layer that is closest to the ground because, by definition, $N_{i1} = \widetilde{N}_{i1}$ for this layer and therefore p_{i1} is always equal to one.

The probability p_{ih} is sometimes referred to as the absorptance probability (not to be confused with absorbance) and is similar to the Leaf Area Density (LAD) definition used in Hosoi and Omasa (2006), the vegetation density index used by d'Oliveira *et al.* (2012) to detect logging infrastructure, and the canopy density metric described in Moran *et al.* (2018). We also note that $p_{ih} = 1 - GF_i(h)$, where $GF_i(h)$ is the gap fraction from the top of the canopy to the top of vertical layer h. Therefore, one can calculate the leaf area density at height h as $LAD_i(h) = -\frac{\ln(1-p_{ih})}{k \times \Delta z}$, where Δz is the height of each vertical layer, assumed to be constant, and k is the extinction coefficient (Bouvier *et al.*, 2015).

As explained in Albuquerque *et al.* (2019), LidarLDA decomposes p_{ih} with the following expression:

$$p_{ih} = \boldsymbol{\theta}_i^T \widetilde{\boldsymbol{\Phi}}_h$$

Recall that $\boldsymbol{\theta}_i^T$ is a size K vector that characterizes grid cell i by containing probabilities that sum to one, representing the relative abundances of each of the K clusters. The vector $\widetilde{\boldsymbol{\Phi}}_h = [\phi_{1h}, \dots, \phi_{Kh}]$, corresponding to a column of the $\boldsymbol{\Phi}$ matrix, is also a size K vector that contains the absorptance probabilities associated with each of the K clusters for vertical layer h.

To illustrate this decomposition, consider the following results for two hypothetical grid cells. The first grid cell has a higher absorptance probability in the shorter vertical layers, suggesting a forest with relatively open canopy and considerable amount of short vegetation (Fig. 1, "Data" panel). The second grid cell has relatively high absorptance probabilities across several vertical layers, suggesting a forest with vegetation of various heights. Based on these data, LidarLDA might identify clusters with relatively distinct vertical profiles. This is captured by the vector $\boldsymbol{\phi}_{\mathbf{k}}$ for each cluster (Fig. 1, "$\boldsymbol{\phi}_{\mathbf{k}}$" panel). For example, cluster 1 could be characterized by low absorptance probabilities across all vertical layers, indicating areas with bare soil. On the other hand, clusters 2 through 4 might be characterized by probabilities that are increasingly concentrated on taller vertical layers, indicating increasingly taller vegetation types.

Because of the characteristics of each cluster, LidarLDA might determine that cluster 1 is much more common in grid cell 1 whereas clusters 3 and 4 are more common in grid cell 2. This is captured by the vector $\boldsymbol{\theta}_{\boldsymbol{i}}$ for each grid cell (Fig. 1, "$\boldsymbol{\theta}_{\mathbf{i}}$" panel). Finally, the inner product of $\boldsymbol{\theta}_{\boldsymbol{i}}$ and $\widetilde{\boldsymbol{\phi}}_{h}$ can be calculated to recover the original LiDAR data, clarifying why LidarLDA can be viewed as a decomposition approach for these data (Fig. 1, "Decomposition" panel).

Fig. 1. Schematic representation of how LidarLDA decomposes LiDAR data into cluster with distinct vertical profiles. Panel A shows the original data together with the corresponding empirical absorptance probabilities, calculated as $\frac{N_{ih}}{\widetilde{N}_{ih}}$. Panels in B shows the $\boldsymbol{\theta}_{\mathbf{i}}$ and $\boldsymbol{\phi}_{\mathbf{k}}$ parameter vectors estimated by LidarLDA. Finally, panel C shows how multiplying $\boldsymbol{\theta}_{\mathbf{i}}$ and $\widetilde{\boldsymbol{\phi}}_{\boldsymbol{h}}$ can recover the original vertical profiles.

2.3. LidarLDA algorithm implementation

We fit this LidarLDA using the Gibbs sampler algorithm originally described in Valle *et al.* (2018) and Albuquerque *et al.* (2019). This algorithm iteratively samples each parameter from its full

conditional distribution (FCD). These FCDs are all available in closed form and are described below.

We start by defining two key quantities that will be used throughout this section. The quantity n_{ihk1} is the number of returns in grid cell i and vertical layer h that were assigned to cluster k. This quantity is calculated as $n_{ihk1} = \sum_j I(\omega_{ijh} = k, x_{ijh} = 1)$. Similarly, let n_{ihk0} be the number of light pulses that are not returned, which can be calculated as $n_{ihk0} = \sum_j I(\omega_{ijh} = k, x_{ijh} = 0)$.

The FCD for $\mathrm{V_{ik}}$ (the parameter that implicitly defines the probability of each cluster in grid cell i $\boldsymbol{\theta_i}$), is given by

$$p(\mathrm{V_{ik}}| \dots) \propto \left[\prod_j \prod_h Cat(\omega_{ijh}|\boldsymbol{\theta_i})\right] \times Beta(V_{ik}|1, \gamma)$$

$$= Beta(n_{i.k.} + 1, n_{i.(>k).} + \gamma),$$

where $n_{i.k.} = \sum_h n_{ihk0} + n_{ihk1}$ and $n_{i.(>k).} = \sum_{k'=k+1}^{K} \sum_h n_{ihk'0} + n_{ihk'1}$.

Recall that the absorptance probability of cluster k in vertical layer h is given by ϕ_{kh}. The FCD for this parameter is given by:

$$p(\phi_{\mathrm{kh}}| \dots) \propto \left[\prod_j \prod_i Bern(x_{ijh}|\phi_{kh})^{I(\omega_{ijh}=k)}\right] \times Beta(\phi_{kh}|\alpha, \beta)$$

$$= Beta(n_{.hk1} + \alpha, n_{.hk0} + \beta),$$

where $n_{.hk0} = \sum_i n_{ihk0}$ and $n_{.hk1} = \sum_i n_{ihk1}$.

Finally, the FCD for the vector containing $n_{ih11}, \dots, n_{ihK1}$ is given by

$$[n_{ih11}, \dots, n_{ihK1}] \sim Multinom(N_{ih}, \boldsymbol{p_{ih1}})$$

where $\boldsymbol{p_{ih1}} = \frac{1}{\sum_q \theta_{iq}\phi_{qh}}[\theta_{i1}\phi_{1h}, \dots, \theta_{iK}\phi_{Kh}]$ and N_{ih} is the number of returns in grid cell i at vertical layer h. Similarly, the FCD for the vector containing $n_{ih10}, \dots, n_{ihK0}$ is given by

$$[n_{ih10}, \dots, n_{ihK0}] \sim Multinom(\widetilde{N}_{ih} - N_{ih}, \boldsymbol{p_{ih0}}).$$

where $\boldsymbol{p_{ih0}} = \frac{1}{\sum_q \theta_{iq}(1-\phi_{qh})}[\theta_{i1}(1-\phi_{1h}), \dots, \theta_{iK}(1-\phi_{Kh})]$ and $\widetilde{N}_{ih}$ is the total number of light pulses that reach grid cell i at vertical layer h. The detailed derivation of these FCDs is provided in Appendix 1.

The Gibbs sampler algorithm was implemented in R (R Core Team, 2020) and C++ (invoked from R using the "Rcpp" package; Eddelbuettel, 2013, Eddelbuettel & Francois, 2011). To run this model, the user has to specify the maximum number of groups K and the model, through a stick-breaking prior that imposes sparsity (Valle *et al.*, 2021a), will often find that only a subset of the specified groups are needed to adequately represent most of the observations. We provided the code as an R package called LidarLDA, freely available in github (https://github.com/drvalle1/LidarLDA) and archived in Zenodo (DOI 10.5281/zenodo.5781482, https://zenodo.org/badge/latestdoi/390455503). This package comes with a detailed tutorial explaining how to format LiDAR data for LidarLDA as well as how to fit the model, interpret, and visualize its results based both on simulated and empirical data.

2.4. Assessing algorithm convergence

In relation to assessing the convergence of our Markov Chain Monte Carlo (MCMC) algorithm, it is important to note that this is a very large model given the large number of parameters that are being estimated. To be precise, focusing only on the top-most parameters, there are I x (K-1) parameters in the $\boldsymbol{\Theta}_{\mathrm{I\times K}}$ matrix (I is the number of grid cells, K is the number of clusters) and K x H parameters in the $\boldsymbol{\Phi}_{\mathrm{K\times H}}$ matrix (H is the number of vertical layers). For example, if there are 100,000 grid cells, 10 clusters, and 35 vertical layers, on total there will be 100,000 x 9 =900,000 parameters and 10 x 35 = 350 parameters in the $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ matrices, respectively. For this reason, just storing posterior samples for each parameter in these matrices can be a substantial challenge, particularly if many iterations are used and no thinning is done, and it is not feasible to evaluate convergence by examining each parameter individually. Therefore, we assess convergence solely based on trace-plots of the log-likelihood and

running the Heidelberger and Welch's diagnostic test and Geweke's statistic on the MCMC samples of the log-likelihood.

2.5. Simulated forest structure data

To evaluate the proposed methodology, we relied on simulated data of forest structure in a landscape with an elevational gradient. We assumed that forest structure was strongly influenced by slope and altitude such that forest types gradually changed with elevation, with increasingly shorter trees as elevation increased. This landscape was divided into 2,601 grid cells and we assumed 30 1-m vertical layers, resulting in 78,030 voxels. Furthermore, we assumed that 100 light pulses reached each grid cell and vertical layer. Finally, two simulated datasets were created, one with three and the other with five forest structural types. The parameters used for the $\mathbf{\Phi}$ matrix are given in Appendix 2 while the parameters for the $\mathbf{\Theta}$ matrix are depicted in Fig. 3.

We fitted LidarLDA to these simulated datasets to determine if it was able to correctly determine the true number of forest types, the spatial distribution of these forest types, and their vertical profiles. We assumed a maximum of 10 clusters and we relied on the following prior parameters $\alpha = \beta = 1$ (i.e., a uniform prior for ϕ_{kh}) and $\gamma = 0.1$. We also compared the results from LidarLDA with those from hierarchical clustering (HC), a commonly used approach to identify forest structural types (Adnan *et al.*, 2019, Moran *et al.*, 2018, Papa *et al.*, 2020). To this end, we relied on the function "agnes" from the R package "cluster" (Maechler *et al.*, 2021) to perform agglomerative hierarchical clustering and we used the Kelley-Gardner-Sutcliffe penalty function (implemented using the function "kgs", also from the "cluster" package) to determine the optimal number of groups.

2.6. Empirical data

2.5.1 Fire experiment

We were interested in understanding the joint effect of fire and forest fragmentation on the vertical structure of forests. For this reason, we focus on an area subjected to experimental fire, located in a transitional forest in Mato Grosso, Brazil, in the southern part of the Amazon Basin (13°04'S,52°23'W). In this experiment, four 50 ha (50 x 1000 m) plots bordering a crop field were established in 2004 (red plots in Fig. 2A). As shown in the timeline in Fig. 2B, one of these plots was left unburned (i.e., control plot "C"), one plot was burned once in 2007 (i.e., "1x"), one plot was burned thrice (2004, 2007, and 2010; hereafter "3x") and the remaining plot was burned yearly from 2004 to 2010, except in 2008 (hereafter "6x"). In the "C", "3x", and "6x" plots, transects of 500 m in length and 20 m in width were created at 0, 10, 30, 100, 250, 500, and 750 m from the forest edge and all trees with diameter at breast height (i.e., 1.3 m from the ground; dbh) greater than 20 cm were measured in 2014 within these transects. Additional details regarding this experiment are available in Balch et al. (2011).

Fig. 2. Study area. In this figure, panel A shows a false-color Landsat 5 image of the study region from June 27, 2011. Panel B shows the timeline of the experimental fires and LiDAR data collection for each plot. The control plot is denoted by "C", the plot burned once in 2007 is denoted by "1x", the plot burned 3 times between 2004-2010 (fire interval of 3 years) is denoted by "3x", and the plot burned 6 times between 2004-2010 (i.e., fire interval of 1 year, except for 2008) is denoted by "6x".

2.5.2 LiDAR data and pre-processing

Data were obtained from the Sustainable Landscapes Brazil project and are freely available online at dos-Santos *et al.* (2019). We relied on LiDAR data for 2014 and 2018 from the Tanguro ranch in Mato Grosso, Brazil, covering a landscape of approximately 1,000 hectares. LiDAR data were pre-processed by subtracting the terrain elevation from the return height to account for topography. Returns with negative height were relatively infrequent (i.e., the median percentage of negative heights per grid cell

was equal to 4.7%) and were assigned to a height of 0. The return data were then grouped spatially into 50m x 50m grid cells and 1-m vertical layers. More than 99.9% of the returns were below 35 m, thus our last vertical layer included all returns with height equal or greater than 35 m. Because the calculation of absorptance probabilities p_{ih} assume approximately vertical light pulses, we eliminated all returns with an absolute angle greater than 5 degrees off-nadir. Finally, to reduce data density while also ensuring an adequate amount of data for each vertical layer, we subsampled the data so that there were at most 100 light pulses reaching each voxel (i.e., $\max(\widetilde{N}_{ih}) = 100$). Ultimately, all these pre-processing steps resulted in approximately 800,000 returns spread throughout ~110,000 - 135,000 voxels for each year.

2.5.3 Fitting the model and post-processing the results

We fit LidarLDA to data from 2014 to estimate the vectors $\boldsymbol{\theta}_{i,\mathbf{2014}}$ and $\boldsymbol{\phi}_{\mathbf{k}}$. Similar to the settings for the simulated data, to fit this model, we assumed a maximum of 10 clusters and we relied on the following prior parameters $\alpha = \beta = 1$ (i.e., a uniform prior for ϕ_{kh}) and $\gamma = 0.1$. We ran the algorithm for 200,000 iterations and assessed convergence by examining trace-plots of the log-likelihood. To determine how the relative abundance of each cluster has changed with time, we relied on the folding-in operation. In this operation, the characteristics of each cluster are kept fixed (i.e., $\boldsymbol{\phi}_{\mathbf{k}}$ is not re-estimated) and only the relative abundance of each cluster in each location is re-estimated (i.e., $\boldsymbol{\theta}_{i,\mathbf{2018}}$ is estimated). A comparison between $\boldsymbol{\theta}_{i,\mathbf{2014}}$ and $\boldsymbol{\theta}_{i,\mathbf{2018}}$ enables the determination of how the spatial distribution of these clusters have changed through time.

Because of changes in data acquisition strategies to reduce costs, the LiDAR data for 2018 had considerably fewer returns with absolute off-nadir angle less than 5 degrees, and approximately 18% of the grid cells did not have any return with these characteristics. Because it is hard to visualize the spatial patterns of the clusters identified by LidarLDA if there are gaps in the resulting maps, we interpolated the LidarLDA results for 2018 for each group using inverse distance weighting (idw function within the R

package "gstat") (Pebesma, 2004). Finally, all maps were created using the R package "ggplot2" (Wickham, 2009). All scripts and files required to reproduce our results were archived in Zenodo (DOI 10.5281/zenodo.5781488, https://zenodo.org/badge/latestdoi/433446658).

3. Results

3.1. Simulated forest structure data

We simulated data with 3 and 5 clusters in which the abundance of each cluster was a function of elevation (Figs. 3A and 3B). Trace-plots and convergence tests suggest that our algorithm applied to these simulated data sets has successfully converged (see details in Appendix 3). We found that LidarLDA estimated well the number of groups given that the first 3 clusters (for the simulated data with 3 clusters) and the first 5 clusters (for the simulated data with 5 clusters) identified by the algorithm accounted for >99% of all the returns on average (Appendix 2). Furthermore, we found that the estimated spatial distribution of each cluster along the elevation gradient (captured by the matrix $\boldsymbol{\Theta}$; Figs. 3C and 3D) closely followed the true distribution of these clusters. Finally, a comparison between the estimated and true absorptance probabilities of each cluster reveals that LidarLDA estimated well the $\boldsymbol{\Phi}$ matrix, with a Pearson correlation coefficient greater than 0.99 (Appendix 2). Taken together, these results reveal that LidarLDA did an excellent job grouping areas with similar 3D profiles and characterizing transition areas comprised of more than one cluster.

Differently from LidarLDA, the agglomerative hierarchical clustering (HC) approach yields hard clustering results (i.e., HC can only assign a single cluster to each grid cell). This is an important limitation. For example, as illustrated in Fig. 3F, HC captures well the overall spatial pattern of the simulated dataset with 5 clusters but fails to capture the transition areas between clusters that are present in Fig. 3B. Importantly, HC found the optimal number of clusters to be 5 even for the data that

was simulated with only 3 clusters (Fig. 3E). As discussed in Valle *et al.* (2018), the reason for this is that hard clustering methods will often yield more clusters than are necessary, often representing transition areas as additional clusters (e.g., yellow and grey clusters in Fig. 3E).

Fig. 3. The true spatial distribution of clusters based on the simulated data (panels A and B) is compared with the spatial distribution estimated based on LidarLDA (panels E and F) and agglomerative hierarchical clustering (HC; panels G and H) based on the simulated data with 3 and 5 clusters (left and right panels, respectively). In these panels, each color represents a different cluster and opacity levels depict the relative abundance of each cluster (transparent = 0 and completely opaque = 1). Elevation is depicted with blue contour lines.

3.2. Empirical data

3.2.1 Number of clusters and their characteristics

Both visual assessment of the trace plot of the log-likelihood and diagnostic test results suggest that our Gibbs sampler algorithm has converged (see details in Appendix 3). By examining the results in the vectors $\boldsymbol{\theta}_\mathbf{i}$, we find that the first 4 clusters together represent, on average, over 99% of all points (Fig. 4A). As a result, from here onwards, we focus on these 4 main clusters. When examining the height distribution of each of these cluster, we find relatively distinct vertical profiles despite significant overlap between clusters (Fig. 4B). For example, cluster 1 has very low absorptance probabilities across almost all vertical layers, suggesting that this cluster represents bare soil, grass or areas with very short vegetation. On the other hand, clusters 2 to 4 represent a gradient from shorter to increasingly taller vertical profiles, respectively. To simplify the reference to these clusters, we label clusters 1 to 4 as "near surface", "short", "intermediate", and "tall", respectively. A schematic representation of these clusters is provided in Fig. 4C.

Fig. 4. Characteristics of the identified clusters. In this figure, panel A displays the distribution of relative abundances of each cluster, as captured by the vector $\boldsymbol{\theta_i}$. Panel B shows the vertical profile of each of the four most important clusters, as captured by the vector $\boldsymbol{\phi_k}$. Panel C provides a schematic representation of these clusters. Figures within panels B and C are ordered from shortest (top) to tallest (bottom) clusters.

Corroborating the schematic representation in Fig. 4C, we found a strong relationship between the different clusters identified by LidarLDA based on the 2014 LiDAR data and the tree diameter distribution for the same year. For example, as shown in Appendix 4, for the plot that was burned 3 times (3x), there is a clear pattern of relatively few and small trees for the transects that are closest to the forest edge and greater abundance and bigger trees as one moves towards the interior of the forest. The LidarLDA-based clusters capture well this pattern since the relative abundance of the near surface cluster (i.e., cluster 1) decreases sharply from the forest edge to the forest interior whereas the relative abundance of clusters 2-4 steadily increases along this gradient (Appendix 4). Similar patterns can be seen for the other areas (i.e., the control area and the 6x plot; see Appendix 4). Furthermore, a comparison of the spatial distribution of the near surface cluster with a map of grass invasion, created based on field observations, supports cluster 4 representing bare soil, grasses and short vegetation (see Appendix 5).

3.2.2 Spatial distribution of LidarLDA clusters in 2014

We found that the spatial distribution of clusters in 2014 was strongly linked to both landscape features and disturbance history. For example, the near surface, short and intermediate clusters (clusters 1 to 3, respectively) were much more common close to the river, whereas the tall cluster 1 was

rare in this area (Fig. 5). In relation to the fire experiments, we can also see in Fig. 5 that the areas burned multiple times had a high proportion of the near surface and short clusters (clusters 1 and 2) whereas the intermediate and tall clusters (clusters 3 and 4) were relatively rare in these areas in 2014. Importantly, the tall cluster was more common in the 6x area when compared to the 3x area, probably a consequence of higher fire intensity in the 3x area due to the fuel buildup enabled by the lower fire frequency (Balch *et al.*, 2015). Furthermore, the area burned once (1x) was more similar to the control area than the areas burned multiple times.

Fig. 5. Heatmaps showing the spatial distribution of each cluster in 2014. Relative abundance of each cluster varies from 0 (cyan) to 1 (purple). Results are only shown for forested areas covered by LiDAR but there is an agricultural field adjacent to the plots. Location of the river is highlighted with blue line while experimental fire plots are outlined in black. The control plot is denoted by "C", the plot burned once in 2007 is denoted by "1x", the plot burned 3 times between 2004-2010 (fire interval of 3 years) is denoted by "3x", and the plot burned 6 times between 2004-2010 (i.e., fire interval of 1 year, except for 2008) is denoted by "6x". Top to bottom panels show the results for individual clusters (numbers in the top left of each panel) and are ordered from low to high stature clusters.

3.2.4 Temporal changes

Assuming 4 main clusters, we use the folding-in operation to compare how the relative abundance of each cluster changed through time by estimating $\boldsymbol{\theta}_{i,2018}$ and calculating $\boldsymbol{\theta}_{i,2018} - \boldsymbol{\theta}_{i,2014}$. This analysis reveals that there is substantial change between 2014 and 2018 at the landscape level, even in areas that were not subject to experimental fire (Fig. 6A). The results for the experimental fire plots, however, are substantially different from those at the landscape level. For instance, at the edge of the forest in the areas burned multiple times, the relative abundance of the near surface cluster (cluster 1)

decreased dramatically with a concurrent increase of the short cluster (cluster 2). In contrast, in the interior of the forest for the areas that were burned multiple times, the short cluster (cluster 2) declined but there was a strong increase in the intermediate cluster (cluster 3).

Another way of visualizing the recovery of the forest after fire at the forest edge and forest interior is using barycentric coordinates, in which we simultaneously display the relative abundance of clusters 1, 2, 3+4 (Fig. 6B). In this figure, points closer to a particular vertex have higher relative abundance of the corresponding cluster and arrows start in 2014 results and point to 2018 results. This figure reveals that the areas burned multiple times (i.e., 3x and 6x) have a much larger fraction of the near surface cluster (cluster 1) at the edge of the forest (i.e., grid cells within 500 m of the forest edge) when compared to the control and 1x plots in 2014. On the other hand, these burned areas tended to have a larger fraction of the short cluster (cluster 2) at the interior of the forest (i.e., grid cells >500 m away from the forest edge). Importantly, only at the interior of the forest have these differences decreased substantially in 2018, revealing a convergence to approximately the same forest structure, whereas there is much less convergence at the forest edge even 8 years after the last fire. Interestingly, the length of these arrows reveals that all the burned areas, including the area burned only once in 2007 (i.e., 1x), are still undergoing large changes in forest structure while the control area has had comparatively smaller changes during the same time period. Taken together, these results illustrate the partial recovery of forest structure after fires stopped (2007 for the 1x plot; 2010 for the 3x and 6x plots) and how distance to forest edge influences this recovery process.

Fig. 6. Recovery process of forest structure between 2014 and 2018 displayed with difference maps (panel A) and barycentric coordinates (panel B). The difference maps were calculated as the relative abundance in 2018 minus the relative abundance in 2014 for each cluster. Increases and decreases are depicted in blue and red, respectively. For the barycentric coordinate figures, each arrow starts at the

coordinates for 2014 and ends at the coordinates for 2018. Each color represents a different experimental plot. Top and bottom panels display the results for the forest edge (defined as all grid cells within 500 m of forest edge) and forest interior (defined as all grid cells more than 500 m away from the forest edge), respectively.

Importantly, as described in detail in Appendix 6, differently from the LidarLDA results, Canopy Height Model (CHM) results fail to identify differences in the forest interior between the burned plots (1x, 3x, and 6x) and the control plot. Furthermore, in contrast to the results shown in Fig. 6, a temporal comparison of CHM results suggest minimal change in canopy height in the interior of burned plots and the control plot from 2014 to 2018. Taken together, these results suggest that LidarLDA can reveal much more information regarding forest structure than CHMs.

4. Discussion

In this article, we have shown how a modified LDA model, called LidarLDA, can be used to generate novel insights on forest structure based on LiDAR data. A key feature of this dimension reduction approach is that it enables the spatial and temporal visualization of changes in forest structure while at the same time appropriately accounting for occlusion of LiDAR light pulses. Using simulated data, we illustrate how this model can recover the true number of clusters and the spatial distribution of these clusters as a function of elevation. Furthermore, through our case study in the Amazon region, we reveal landscape-level differences in forest structure associated with proximity to the river as well as the long-term effects of fire and forest fragmentation on forest structure. Importantly, a comparison with other types of LiDAR products that just focus on top-of-canopy information, such as a Canopy Height Model, reveals how much more information can be extracted using LidarLDA regarding the impact of fires and forest fragmentation on forest structure.

Due to its unsupervised nature, our LidarLDA model is well suited for exploratory analysis, potentially revealing novel spatial and temporal patterns of forest structure. Importantly, differently from standard hard clustering approaches used to create forest structural types, the LidarLDA model is able to capture gradual spatial and/or temporal changes in forest structure. For example, the analysis of our simulated data reveals that LidarLDA can accurately capture the gradual spatial changes in forest structure associated with elevation (Fig. 3). On the other hand, hard clustering approaches commonly used to determine forest structural types (Adnan *et al.*, 2019, Moran *et al.*, 2018, Papa *et al.*, 2020) cannot capture these gradual changes because each grid cell can only be assigned to a single cluster. Similarly, characterizing the gradual temporal changes (e.g., as depicted in Fig. 6) would be very challenging with hard clustering approaches. Another important limitation associated with hard clustering approaches is that they often have to create more clusters than warranted to be able to fit the data well and represent these transition zones. This is illustrated with our simulations with 3 clusters and is corroborated by past studies on the ability of hard clustering approaches in describing transition zones (Valle *et al.*, 2018).

In our case study, we characterized approximately 1,000 ha of this landscape and identified the strong influence of distance to the river on forest structure. Furthermore, a comparison of field data and LidarLDA results revealed that LidarLDA could capture well the gradual changes in the diameter distribution of trees resulting from the synergistic effects of fire and distance to forest edge, providing confidence that LidarLDA can be used over large areas to detect spatial and temporal changes in forest structure. The comparison of LidarLDA results for the burned and control plots largely corroborated the results from previous studies based on field measurements at the same site, an important result given the unsupervised nature of LidarLDA. For example, the effect of fire on forest structure is strongest near the forest edge and more pronounced in the 3x plot than on the 6x plot, probably due to the fuel build up between years in the 3x plot (Balch *et al.*, 2015). More fuel in drier conditions favors high-intensity

fires, which can lead to increased postfire tree mortality, greater loss of aboveground live biomass (Brando *et al.*, 2014) and increased grass invasion (Silverio *et al.*, 2013), with substantial change in species composition (Valle *et al.*, 2021b). Interestingly, the temporal comparison revealed substantial changes in forest structure even after almost a decade after fires have ceased, capturing the ongoing process of post-fire forest recovery. In contrast, the results from the canopy height model do not reveal major differences between the forest interior of the control plots and the burned plots and fail to capture the large temporal changes in forest structure (Appendix 6). Ultimately, by relying only on information from the highest trees, CHMs miss other changes in the 3D structure of the forest. Finally, we note that past studies focused on the Tanguro ranch have ignored the area that was burned just once (1x) because no field data were collected for this site. LidarLDA results reveal that the short cluster (cluster 2) is decreasing in this plot while the tall cluster (cluster 4) is increasing (Fig. 6), indicating substantial change in forest structure between 2014 and 2018, even though this plot was burned just once in 2007.

An important limitation of our methodology is the speed of our algorithm. Although our algorithm leverages C++ within R to perform the most computationally intensive tasks, our model can still be computationally intensive to fit because we rely on an iterative Markov Chain Monte Carlo (MCMC) approach. This was not a problem when data were spatially discretized into 50 x 50 m grid cells for a single region; it took approximately 1.6 hours to run LidarLDA for 20,000 iterations on the 2014 dataset containing approximately 3,900 grid cells and 35 height classes on an Intel Core i7 desktop with 3.4 GHz processor and 16 GB of RAM. This was done assuming a maximum of 10 clusters. On the other hand, when the number of grid cells was increased by 10-fold while keeping all of the other characteristics constant, our algorithm took 13.7 hours. Monitoring larger landscapes and/or using smaller grid cells would likely require high performance computing. Exploring approaches to speeding

up the fitting of LidarLDA (e.g., using variational Bayes methods; Blei *et al.*, 2017) is likely to be a very important topic for future research if larger datasets are to be analyzed.

To prepare the data for LidarLDA, we adopted 50 x 50 m grid cells and discretized height into vertical layers of 1-m in width. These settings are relatively standard in LiDAR studies focused on forest structure in this region (e.g., Andersen *et al.*, 2013, Carrasco *et al.*, 2019, Papa *et al.*, 2020, Silva *et al.*, 2017) but it is important to acknowledge the tradeoffs associated with these choices. For example, while choosing smaller grid cells can potentially represent spatial variation at a finer scale, two important drawbacks of relying on smaller grid cells are that the number of light pulses per grid cell within the pre-specified angle range can be relatively small, hampering inference, and the model is likely to take longer to fit. Furthermore, a finer spatial scale may or may not be ecologically relevant depending on the size of individual trees and their canopies. As a result, the decision regarding which grid cell size to adopt requires one to consider the trade-off between algorithm speed and data availability versus the ecological importance of fine scale spatial variation. A related concern is that of over-fitting the data given that LidarLDA already contains a large number of parameters and the number of parameters increases with the number of grid cells. The standard approach to determining if the data are being over-fitted is to evaluate if out-of-sample predictions deteriorate as the number of parameters increases. Unfortunately, this straight-forward approach does not work for LidarLDA because, like many other LDA-type models, it does not include predictor variables and therefore predictions cannot be made. While the use of the truncated stick-breaking prior helps in ensuring parsimony by limiting the number of clusters, additional research is still needed to determine when overfitting is likely to be an issue for models like LidarLDA. Finally, it is not clear what the minimum number of light pulses per grid cell and vertical layer should be for LidarLDA to estimate well the absorptance probability of the different clusters. We believe this is an important topic for future research.

We have shown that LidarLDA enables the visualization of spatial and temporal patterns of forest structure in a way that provides much more information than standard canopy height models. As a result, we believe that LidarLDA will become an indispensable tool for scientists interested on how large-scale phenomena (e.g., selective logging, climate change, and fire) and biophysical characteristics (e.g., topography, soil fertility, and rainfall) influence forest structure and/or how forest structure influences ecosystem services (e.g., erosion control, recreation, wildlife habitat, water supply and/or regulation). For example, it is possible that LidarLDA could be used in the future to monitor forest concessions, assessing the short-term structural damage associated with logging as well as how long it takes for the forest to recover most of its structure after logging. Similarly, it is possible that LidarLDA could be used to better determine emissions associated with understory fire by assessing changes in structural biomass and the required time for forests to regain their original structure. Despite our focus on forests, it is important to emphasize that LidarLDA is likely to also be useful to characterize the structural complexity and answer similar questions for other types of vegetation. Given the increasing availability of LiDAR data, collected from unmanned aerial vehicles (UAVs), planes (e.g., data from the National Ecological Observatory Network [NEON]), or satellites (e.g., data collected by the NASA's Global Ecosystem Dynamics Investigation [GEDI] mission), the time is ripe for ecological applications to use the full potential of these high-dimensional datasets. We hope that LidarLDA can contribute to this effort.

5. Acknowledgements

We would like to thank Michael Keller, Ekena Pinage, Divino Silverio, and Lucas Paolucci for providing helpful comments and suggestions on an earlier draft of this article. This work was partly supported by the US Department of Agriculture National Institute of Food and Agriculture McIntire–Stennis project 1005163 and by the US National Science Foundation award 2040819 to DV. M. L. was

supported in part by the Next Generation Ecosystem Experiments-Tropics, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research.

6. Conflict of Interest

The authors declare that there is no conflict of interest.

7. Author contributions

Denis Valle: Conceptualized the article, developed the software, created figures and ran the analysis, wrote the original draft. **Carlos Alberto Silva**: contributed with ideas that improved the proposed methodology, reviewed and edited the manuscript**. Marcos Longo:** contributed with ideas that improved the proposed methodology, reviewed and edited the manuscript. **Paulo Brando**: Curated the data, reviewed and edited the manuscript.

8. Data availability statement

The aggregate data for the fire experiments are stored and publicly available at Dryad (Valle, 2022a). The R package to run LidarLDA is freely available at https://github.com/drvalle1/LidarLDA and is archived in Zenodo (Valle, 2022c). The raw LiDAR data are freely available through the website https://www.paisagenslidar.cnptia.embrapa.br/webgis/. The edited LiDAR data, together with all the scripts and files required to reproduce our results, are archived in Zenodo (Valle, 2022b).

9. References

Adnan S, Maltamo M, Coomes DA *et al.* (2019) A simple approach to forest structure classification using airborne laser scanning that can be adopted across bioregions. Forest Ecology and Management, **433**, 111-121.

Albuquerque P, Valle D, Li D (2019) Bayesian LDA for mixed-membership clustering analysis: the Rlda package. Knowledge-Based Systems.

Alencar AA, Brando PM, Asner GP, Putz FE (2015) Landscape fragmentation, severe drought, and the new Amazon forest fire regime. Ecological Applications, **25**, 1493-1505.

Almeida CT, Galvao LS, Aragao LEOC *et al.* (2019a) Combining LiDAR and hyperspectral data for aboveground biomass modeling in the Brazilian Amazon using different regression algorithms. Remote Sensing of Environment, **232**.

Almeida DRA, Stark SC, Shao G *et al.* (2019b) Optimizing the Remote Detection of Tropical Rainforest Structure with Airborne Lidar: Leaf Area Profile Sensitivity to Pulse Density and Spatial Sampling. Remote Sensing, **11**, https://doi.org/10.3390/rs11010092.

Andersen H-E, Reutebuch SE, Mcgaughey RJ, D'oliveira MVN, Keller M (2013) Monitoring selective logging in western Amazonia with repeat lidar flights. Remote Sensing of Environment.

Balch JK, Brando PM, Nepstad DC *et al.* (2015) The susceptibility of southeastern Amazon forests to fire: insights from a large-scale burn experiment. Bioscience, **65**, 893-905.

Balch JK, Nepstad D, Curran LM *et al.* (2011) Size, species, and fire behavior predict tree and liana mortality from experimental burns in the Brazilian Amazon. Forest Ecology and Management, **261**, 68-77.

Blei DM, Kucukelbir A, Mcauliffe JD (2017) Variational inference: a review for statisticians. Journal of the American Statistical Association, **112**, 859-877.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research, **3**, 993-1022.

Bouvier M, Durrieu S, Fournier RA, Renaud J-P (2015) Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. Remote Sensing of Environment, **156**, 322-334.

Brando PM, Balch JK, Nepstad DC *et al.* (2014) Abrupt increases in Amazonian tree mortality due to drought-fire interactions. Proceedings of the National Academy of Science, **111**, 6347-6352.

Brando PM, Soares Filho BS, Rodrigues L *et al.* (2020) The gathering firestorm in southern Amazonia. Science Advances, **6**.

Carrasco L, Giam X, Papes M, Sheldon KS (2019) Metrics of lidar-derived 3D vegetation structure reveal contrasting effects of horizontal and vertical forest heterogeneity on bird species richness. Remote Sensing, **11**.

Casella G, Moreno E, Giron FJ (2014) Cluster analysis, model selection, and prior distributions on models. Bayesian Analysis, **9**, 613-658.

Christensen EM, Harris DJ, Ernest SKM (2018) Long-term community change through multiple rapid transitions in a desert rodent community. Ecology, **99**, 1523-1529.

Costa MBT, Silva CA, Broadbent EN *et al.* (2021) Beyond trees: mapping total aboveground biomass density in the Brazilian savanna using high-density UAV-lidar data. Forest Ecology and Management, **491**.

D'oliveira MVN, Reutebuch SE, Mcgaughey RJ, Andersen H-E (2012) Estimating forest biomass and identifying low-intensity logging areas using airborne scanning lidar in Antimary State Forest, Acre State, Western Brazilian Amazon. Remote Sensing of Environment, **124**, 479-491.

Dietzel K, Valle D, Fierer N, U'ren JM, Barberan A (2019) Geographical distribution of fungal plant pathogens in dust across the United States. Frontiers in Ecology and Evolution, **7**.

Dos-Santos MN, Keller MM, Morton DC (2019) LiDAR surveys over selected forest research sites, Brazilian Amazon, 2008-2018. ORNL DAAC, Oak RIdge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1644; https://www.paisagenslidar.cnptia.embrapa.br/webgis/.

Eddelbuettel D (2013) *Seamless R and C++ integration with Rcpp,* New York, Springer.

Eddelbuettel D, Francois R (2011) Rcpp: seamless R and C++ integration. Journal of Statistical Software, **40**, 1-18.

Felipe-Lucia MR, Soliveres S, Penone C *et al.* (2018) Multiple forest attributes underpin the supply of multiple ecosystem services. Nature Communications, **9**.

Hosoda S, Nishijima S, Fukunaga T, Hattori M, Hamada M (2020) Revealing the microbial assemblage structure in the human gut microbiome using Latent Dirichlet Allocation. Microbiome, **8**.

Hosoi F, Omasa K (2006) Voxel-based 3-D modeling of individual trees for estimating leaf area density using high-resolution portable scanning Lidar. IEEE Transactions on Geoscience and Remote Sensing, **44**, 3610-3618.

Jenkins MA, Schaap B (2018) Background Analytical Study 1: Forest Ecosystem Services. In: *United Nations Forum on Forests.* https://www.un.org/esa/forests/wp-content/uploads/2018/05/UNFF13_BkgdStudy_ForestsEcoServices.pdf

Jucker T, Hardwick SR, Both S *et al.* (2018) Canopy structure and topography jointly constrain the microclimate of human-modified tropical landscapes. Global Change Biology, **24**, 5243-5258.

Knott JA, Jenkins MA, Oswalt CM, Fei S (2019) Community-level responses to climate change in forests of the eastern United States. Global Ecology and Biogeography, 1-16.

Legendre P, Legendre L (2012) *Numerical ecology,* Amsterdam, Elsevier Science.

Liu B, Liu L, Tsykin A *et al.* (2010) Identifying functional miRNA-mRNA regulatory modules with correspondence Latent Dirichlet Allocation. . Bioinformatics, **26**, 3105-3111.

Longo M, Saatchi S, Keller M *et al.* (2020) Impacts of degradation on water, energy, and carbon cycling of the Amazon tropical forests. Journal of Geophysical Research: Biogeosciences, **125**, e2020JG005677.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2021) cluster: Cluster Analysis Basics and Extensions. R package version 2.1.2. https://CRAN.R-project.org/package=cluster

Moran CJ, Rowell EM, Seielstad CA (2018) A data-driven framework to identify and compare forest structure classes using LiDAR. Remote Sensing of Environment, **211**, 154-166.

Mori AS, Lertzman KP, Gustafsson L (2017) Biodiversity and ecosystem services in forest ecosystems: a research agenda for applied forest ecology. Journal of Applied Ecology, **54**, 12-27.

Muhlfeld CC, Cline TJ, Giersh JJ, Peitzsch E, Florentine C, Jacobsen D, Hotaling S (2020) Specialized meltwater biodiversity persists despite widespread deglaciation. Proceedings of the National Academy of Science, 1-7.

Papa DA, Almeida DRA, Silva CA *et al.* (2020) Evaluating tropical forest classification and field sampling stratification from lidar to reduce effort and enable landscape monitoring. Forest Ecology and Management, **457**.

Pearson TRH, Brown S, Murray L, Sidman G (2017) Greenhouse gas emissions from tropical forest degradation: an underestimated source. Carbon Balance and Management, **12**.

Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Computers & Geosciences, **30**, 683-691.

Pohle J, Langrock R, Van Beest FM, Schmidt NM (2017) Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. Journal of Agricultural, Biological, and Environmental Statistics, **22**, 270-293.

R Core Team (2020) R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/

Rex FE, Silva CA, Corte APD *et al.* (2020) Comparison of statistical modelling approaches for estimating tropical forest aboveground biomass stock and reporting their changes in low-intensity logging areas using multi-temporal LiDAR data. Remote Sensing, **12**.

Riano D, Meier E, Allgower B, Chuvieco E, Ustin SL (2003) Modeling airborne laser scanning data for the spatial generation of critical forest parameters in fire behavior modeling. Remote Sensing of Environment, **86**, 177-186.

Salm R, Prates A, Ressye Simoes N, Feder L (2015) Palm community transitions along a topographic gradient from floodplain to terra firme in the eastern Amazon. Acta Amazonica, **45**, 65-74.

Silva CA, Hudak AT, Vierling LA *et al.* (2017) Impacts of airborne lidar pulse density on estimating biomass stock and changes in a selectively logged tropical forest. Remote Sensing, **9**.

Silverio DV, Brando PM, Balch JK, Putz FE, Nepstad DC, Oliveira-Santos C, Bustamante MMC (2013) Testing the AMazon savannization hypothesis: fire effects on invasion of a neotropical forest by native cerrado and exotic pasture grasses. Philosophical Transaction of the Royal Society B: Biological Sciences, **368**, 20120427.

Silverio DV, Brando PM, Bustamante MMC, Putz FE, Marra DM, Levick SR, Trumbore SE (2019) Fire, fragmentation, and windstorms: a recipe for tropical forest degradation. Journal of Ecology, **107**, 656-667.

Sommeria-Klein G, Zinger L, Coissac E, Iribar A, Schimann H, Taberlet P, Chave J (2019) Latent Dirichlet Allocation reveals spatial and taxonomic structure in a DNA-based census of soil biodiversity from a tropical forest. Molecular Ecology Resources, 1-16.

Vaduva C, Gavat I, Datcu M (2013) Latent Dirichlet Allocation for spatial analysis of satellite images. IEEE Transactions on Geoscience and Remote Sensing, **51**, 2770-2786.

Valle D (2022a) Data from: Size, species, and fire behavior predict tree and liana mortality from experimental burns in the Brazilian Amazon. Dryad, Dataset. https://doi.org/10.5061/dryad.vq83bk3s5

Valle D (2022b) LidarLDA analysis scripts. https://zenodo.org/badge/latestdoi/433446658

Valle D (2022c) LidarLDA package source code. https://zenodo.org/badge/latestdoi/390455503

Valle D, Albuquerque P, Zhao Q, Barberan A, Fletcher Jr. RJ (2018) Extending the Latent Dirichlet Allocation model to presence/absence data: a case study on North American breeding birds and biogeographic shifts expected from climate change. Global Change Biology.

Valle D, Baiser B, Woodall CW, Chazdon R (2014) Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. Ecology Letters, **17**, 1591-1601.

Valle D, Jameel Y, Betancourt B, Azeria E, Attias N, Cullen J (2021a) Automatic selection of the number of clusters using Bayesian clustering and sparsity-inducing priors. Ecological Applications.

Valle D, Shimizu G, Izbicki R *et al.* (2021b) The Latent Dirichlet Allocation model with covariates (LDAcov): a case study on the effect of fire on species composition in Amazonian forests. Ecology and Evolution.

Vancutsem C, Achard F, Pekel J-F *et al.* (2021) Long-term (1990–2019) monitoring of forest cover changes in the humid tropics. Science Advances, **7**, eabe1603.

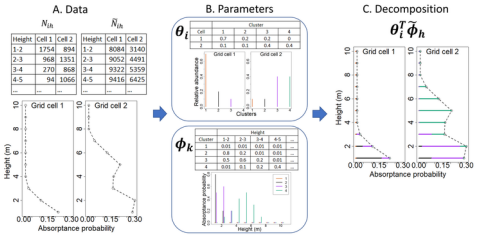
Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis,* New York, Springer-Verlag.

Wittmann F, Schongart J, Montero JC *et al.* (2006) Tree species composition and diversity gradients in white-water forests across the Amazon Basin. Journal of Biogeography, **33**, 1334-1347.

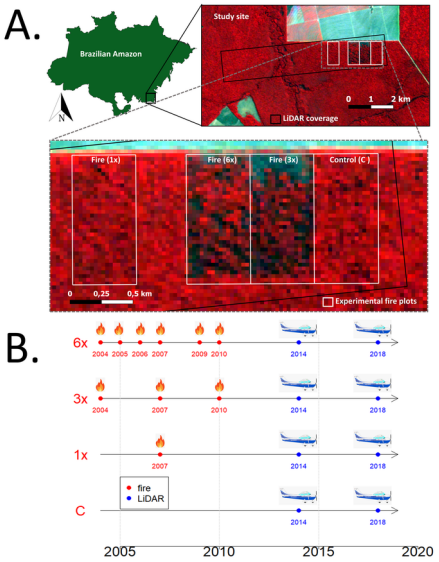
Xing D, Girolami M (2007) Employing Latent Dirichlet Allocation for fraud detection in telecommunications. Pattern Recognition Letters, **28**, 1727-1734.

Yang J, Kang Z (2018) Voxel-based extraction of transmission lines from airborne LiDAR point cloud data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **11**, 3892-3904.

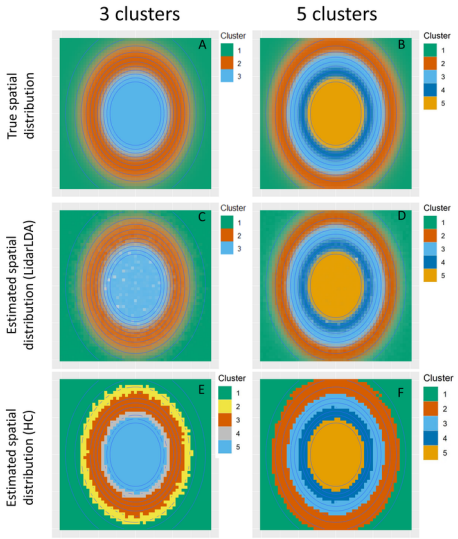
Zhiqing L, Pengcheng L, Qing X, Shuai X, Yang Z (2020) Point-cloud detection of buildings based on a latent Dirichlet allocation model with waveform data. Remote Sensing Letters, **11**, 235-244.



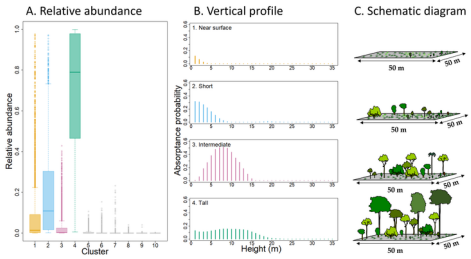
MEE3_13836_figure1.tif



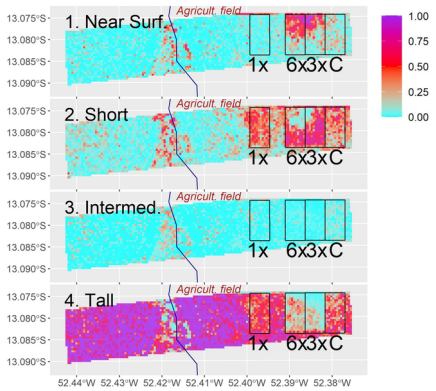
MEE3_13836_figure2.tif



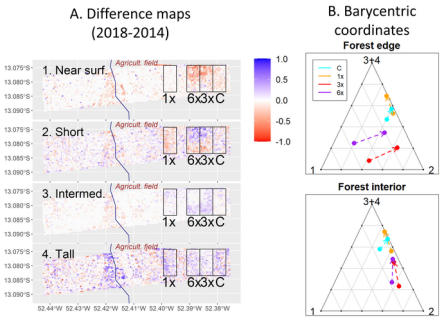
MEE3_13836_figure3.tif



MEE3_13836_figure4.tif



MEE3_13836_figure5.tif



MEE3_13836_figure6.tif