bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data-intensive multidimensional modeling of forest dynamics

Jean F. Liénard [1], Dominique Gravel [2], Matthew V. Talluto [2], Nikolay S. Strigul [1]*

[1]: Department of Mathematics, Washington State University Vancouver, Washington, USA

[2]: Département de Biologie, Universit du Québec à Rimouski, Québec, Canada

*: corresponding author email: nick.strigul@vancouver.wsu.edu

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Summary:

1. Forest dynamics are highly dimensional phenomena that are poorly understood theoretically. Modeling these dynamics is data-intensive and requires repeated measurements taken with a consistent methodology. Forest inventory datasets offer unprecedented opportunities to model these dynamics, but they are analytically challenging due to high dimensionality and sampling irregularities across years. We develop a methodology for predicting forest stand dynamics using such datasets.

2. Our methodology involves the following steps: 1) computing stand level characteristics from individual tree measurements, 2) reducing the characteristic dimensionality through analyses of their correlations, 3) parameterizing transition matrices for each independent dimension using Gibbs sampling, and 4) deriving predictions of forest developments at different timescales.

3. Applying our methodology to the Quebec forest inventory database, we discovered that four independent dimensions were required to describe the stand structure. We were able to successfully estimate transition matrices for each of these dimensions. The model predicted substantial short-term increases in biomass and longer-term increases in the average age of trees, biodiversity, and shade intolerant species.

4. *Synthesis.* Using highly dimensional and irregularly sampled forest inventory data, our original data-intensive methodology provides both descriptions of the short-term dynamics as well as predictions of forest development on a longer timescale. This method can be applied in other contexts such as conservation and silviculture, and can be delivered as an efficient tool for sustainable forest management.

Key-words: data-intensive model, forest dynamics, Gibbs sampling, Markov chain model, Markov chain Monte Carlo, patch-mosaic concept, plant population and community dynamics

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

Forest ecosystems are complex adaptive systems with hierarchical structures resulting from self-organization at multiple levels (Levin, 1999). These levels, or scales, include genomes, cells, organs, individual organisms, populations, and landscapes. Numerous distinct processes occur simultaneously at all of these levels of forest organization. Gene expression and cellular processes occur at spatial and temporal scales of micrometers and milliseconds, individual tree growth and competition for resources can be studied at scales measured in the ranges of centimeters–meters and months–years, while processes at the ecosystem level, such as forest succession, occur on scales of kilometers and decades. The challenge of ecology is that all these processes are interdependent. Linking processes across scales is difficult because we currently lack an integrated theoretical framework describing self-organization in forest ecosystems as well as the quantitative framework for implementing the theory (Peters et al., 2007, Chave, 2013).

The hierarchical patch-mosaic concept allows for the development of multi-scale quantitative models for forest self-organization. The general idea is to unify two ecological theories, the patch-mosaic concept and the hierarchical organization of ecosystems (Wu and Loucks, 1996). The patch-mosaic concept was actively developed in the second half of the twentieth century after Watt (1947) suggested that ecological systems can be considered a collection of patches at different successional stages. Ecological equilibria arise at the level of the mosaic of patches rather than at the level of one patch. The hierarchical concept treats ecosystems as nested arrangements of units of biological organization (O'Neill et al., 1986). Thus, the hierarchical patch-mosaic concept represents ecosystems as nested collections of patch-mosaic systems that evolve simultaneously at several hierarchical scales (Wu and Loucks, 1996).

One example of this concept applied to forested ecosystems is the Matreshka modeling framework (Strigul, 2012). In this model, landscape-level vegetation dynamics occur on several hierarchical scales, including individual trees, forest stands and landscape units driven by natural and anthropogenic disturbances of small, intermediate or large magnitudes. Small-scale forest disturbances are associated with gap dynamics, where individual tree mortality causes gaps in the canopy of the tree crown size. Intermediate disturbances affect tree neighborhoods and create larger gaps equal to several tree crowns. Finally, large-scale catastrophic disturbances destroy all trees in the stand and cause complete regeneration. Disturbances are caused by natural processes such as mechanical damage by wind and hurricanes, forest fires, diseases and insect outbreaks, or by different logging practices. Predictive models developed within this framework take into account patch-mosaic dynamics occurring at these nested hierarchical scales. In particular, within-stand patch dynamics (gap dynamics) occurring at the level of individual trees can be predicted by an individual-based forest model (for example, SORTIE (Pacala et al., 1996) or LES models (Strigul, 2012)) or by the Perfect Plasticity Approximation (Strigul et al., 2008), and dynamics at the stand level can be modeled using Markov chain models (Strigul et al., 2012).

In this paper, we develop a landscape-scale patch-mosaic model of forest stand dynamics using a Markov chain framework, and develop a method for parameterizing the model using widely available forest inventory data. Parameterizing such a model is difficult, as the patch-mosaic modeling framework requires a quantitative characterization of forest stands reflecting changes caused by succession as well as anthropogenic and natural disturbances, implying the

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

need for large, long-term datasets demonstrating stand evolution over time. Candidate forest inventory datasets are often collected at irregular time intervals that are not synchronized across the focal area, and data collection procedures including spatial plot design and tree measurement methods can be different at various survey times and conducted by different surveyors (Strigul et al., 2012). Thus, it is necessary to develop robust and efficient statistical methods to obtain unbiased and consistent model estimates. To overcome these difficulties, we develop a Gibbs sampling procedure for augmenting the data and estimating rates of change in stand structure.

Model development and validation are illustrated using the Quebec provincial forest inventory data (Perron et al., 2011). This inventory is one of the extended forest inventories that have been established in North America, among others led by the Canadian provincial governments and the USDA Forest Inventories and Analysis program in the USA. These inventories provide a representative sample of vegetation across the landscape through a large number of permanent plots that are measured repeatedly. Although they were originally developed for estimating growth and yield, they were rapidly found to be extremely useful in the contexts of forest ecology, biogeography and landscape dynamics. Each permanent plot consists of individually marked trees that are periodically surveyed and remeasured. Each plot can be considered as a forest stand and then, theoretically, the forest inventories provide empirical data sufficient for parametrization and validations of patch-mosaic models (Strigul et al., 2012). However, practical development of the patch-mosaic forest models including parametrization, validation and prediction is notoriously challenging and data-intensive, assuring the need for development of new modeling approaches and data analysis methods.

Our objective in this study is to develop a data-intensive modeling methodology that allows us to understand and predict forest stand dynamics as complex phenomena while making minimal simplifying modeling assumptions. The idea of a data-intensive modeling approach is to develop and explore a predictive and quantitative theory of forest dynamics using statistical modeling, rather than using hypothesis-driven theoretical modeling. We focus here on the development of the modeling framework and illustrate the application of the framework to a large forest inventory dataset spanning 38 years of observations collected in Quebec. We present in this paper the general methodology and demonstrate each of its steps on the Quebec dataset. In particular, we consider the dimensionality of macroscopic stand characteristics in this dataset and present evidence that much of it is irreducible when considering our goal of predicting landscape dynamics. Finally, we apply the method to predict long-term dynamics of Quebec forests, as represented by a subset of macroscopic properties that best represent the variability in the data, and we validate the model utilizing two independent subsets of the original data.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2 Patch-mosaic modeling framework

The goal of this section is to introduce the modeling of patch-mosaic using Markov chains, which is generalized and employed to understand multidimensional forest dynamics in the next sections. The patch-mosaic concept assumes that the vegetation at the landscape level can be represented as a collection of isolated spatial units - patches - where patch development follows a general trajectory and is subject to disturbances (Watt, 1947, Levin and Paine, 1974). Patch-mosaic models are derived using the conservation law, which takes into account patch aging and other changes to macroscopic variables representing succession, growth of patches in space, and disturbances (Levin and Paine, 1974). The same general idea as well as mathematical derivations are broadly used in population dynamics to describe age- and size-structured population dynamics. Patch-mosaic models can be partial differential equations or discrete models depending on whether time and patch stages are assumed to be continuous or discrete. Classic continuous patch-mosaic models are based on the application of the conservation law to continuously evolving patches that can be destroyed with a certain probability, and can be represented by the advection equation (model developed by Levin and Paine, 1974, for fixed-size patches) or equivalently by the Lotka-McKendick-von Foerster model (Strigul et al., 2008). The continuous patch-mosaic models have been used in forest ecology to describe modeling of individual-based dynamics of canopy trees within the stand or forest gap dynamics (Kohyama et al., 2001, Kohyama, 2006).

In the case of patches changing in discrete time, the derivation of the conservation law leads to discrete-type patch-mosaic models. In particular, the advection-equation model (Levin and Paine, 1974) is essentially equivalent to several independently developed discrete models (Feller, 1971, Van Wagner, 1978, Leslie, 1945). These models consider only large scale catastrophic disturbances (patch "death" process), destroying the patch, which then develops along the selected physiological axis until the next catastrophic disturbance Levin and Paine (1974). However, the existence of small and intermediate disturbances that affect forest stands is a critical phenomenon that calls for a reappraisal of classic patch-mosaic models (Strigul et al., 2012). The stochastic model we are considering here employs a Markov chain framework that is capable of taking into account different disturbances. In this model, the next state of a forest stand depends only on the previous state, and the probabilities of going from one state into another are summarized in what is called a transition matrix, denoted T. We summarize the distribution of states at time t as the row vector X_t, with length equal to the number of discrete classes of patch state and with a sum equal to 1. We can predict $X_{t+\Delta t}$ by multiplying the transition matrix:

$$X_{t+\Delta t} = X_t \times T \tag{1}$$

To project an arbitrary number n time steps into the future, one simply multiplies by T^n instead of T. The Perron-Frobenious Theorem guarantees the existence of the long-term equilibrium that can be practically found as the normalized eigenvector corresponding to the first eigenvalue or by iterative sequence of state vectors. In this paper we employ the iterative method as it allows to observe the transient dynamics, and to deduce the long-term equilibrium we simply chooses an n large enough to satisfy the condition:

$$|X_{t+n\Delta t} - X_{t+(n-1)\Delta t}| < \epsilon \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We illustrate in Fig. 1 the application of the Markov chain methodology to the modeling of forest stand dynamics across three simplified models. All these models have been designed to have the same probability of going forward (probability of 20% to go to the next state) and of going backward (when all earlier states are pooled, the probability to go backward is consistently 10%). However, the long-term distributions are substantially different across models, demonstrating the need for a data-intensive approach that incorporates variable-scale disturbances to quantitatively constrain the transition matrices.

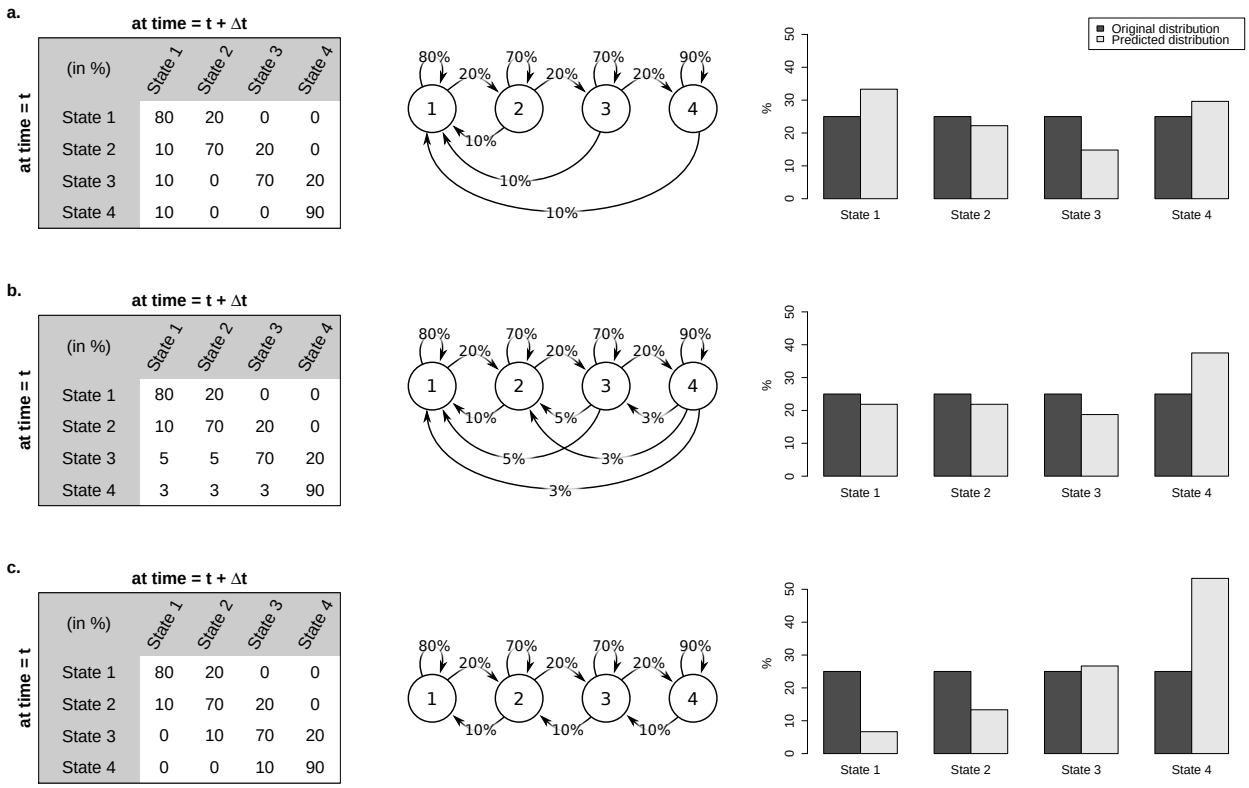


Fig. 1. Hypothetical examples illustrating the Markov chain model for forest stand dynamics. Three different scenarios are considered: birth-and-disaster (a), birth-and-disturbances (b) and birth-and-regression (c). For each scenario, we show the transition matrix rounded to the closest percent (left), the corresponding Markov chain graphs (middle) and the long-term equilibrium when starting from the same uniform distribution (right). In all three scenarios, the probability of going to the next state is 20%, and the probability of going to any earlier state is 10%; however, the resulting long-term distribution of states are very different.

The classic patch-mosaic methodology assumes that patch dynamics can be represented by changes in a single macroscopic variable characterizing the state of the patch as a function of time (Levin and Paine, 1974). Forest disturbances are traditionally associated with a loss of biomass; however, Markov chain models based only on biomass do not capture forest succession comprehensively (Strigul et al., 2012). In this paper, instead of considering a model for one variable, we consider a multidimensional model in which independent transition

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

matrices are obtained for uncorrelated stand characteristics. This can be considered the next stage of generalizing patch-mosaic models for complex adaptive systems, which change in multiple dimensions simultaneously.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 Materials and Methods

3.1 Outline

This section introduces multidimensional Markov chain modeling of forest dynamics and presents an original methodology to construct transition matrices from forest inventory data. The substantial challenge is that these data include irregular sampling intervals and variable numbers of plots sampled in each year. We describe the general concepts of the methodology along with practical guidelines using the inventory led by the provincial Ministry of Natural Resources and Wildlife in Quebec (Appendix 1), from which we extracted individual-level tree measurements taken from repeatedly sampled plots.

Prior to the discussion of the underlying theoretical rationale and implementation details of the methodology, we provide a concise outline of the key steps to use Gibbs sampling to estimate a transition matrix from irregular measurements:

1. Compute stand level characteristics for each plot and for each survey year. Analyze the dimensionality of these characteristics using correlation and principal component analysis.

2. Construct temporal sequences of independent characteristics depending on forest survey dates. Use Gibbs sampling to infer the transition matrix. This algorithm consists of random initialization of missing values followed by iteration of parameter estimation and data augmentation:

   - Parameter estimation: Compute the transition matrix using the (augmented) sequences of plot transitions.
   - Data augmentation: Draw new sequences conditional on the new transition matrix.

The transition matrices for Quebec forests were obtained using this method with a three-year time step. Future and equilibrium landscape characteristics were predicted according to equations 1 and 2 (see the Results section).

3.2 Step 1: stand characteristics and dimensional analysis

This step consists of (a) the selection of a set of stand-level forest characteristics, (b) the dimensional analysis of these characteristics, (c) their decomposition into uncorrelated axes, and (d) the discretization of these uncorrelated axes.

Our modeling method can be applied for the prediction of any forest stand characteristic under the condition that it is computable from every single plot survey. The particular choice of the characteristics depends on available data and research objectives. A general guideline is that these characteristics should summarize data from individual trees into macroscopic indicators of stand structure, which can then be used to compare forests across different ecosystems. We consider six characteristics of Quebec forests according to the rationale presented in Strigul et al. (2012) and Lienard et al. (2014) (under review). We denote S the set of species inside each plot and T the set of trees inside each plot, and compute for each single plot survey the following characteristics:

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

- biomass, estimated from Jenkins et al. (2003), using the formula: $\sum_{i \in T} e^{\mathrm{B1}_i + \mathrm{B2}_i log(d_i)}$ where B1 and B2 are species specific density constants, and d is the trunk diameter at breast height in cm. The resulting biomass is expressed in 10^3 kg/ha, and is an indicator of the overall wood quantity of forest stands.

- basal area, computed as the sums of trunk diameters at breast height d: $\sum_{i \in T} \pi \left(\frac{d_i}{2}\right)^2$. The basal area is expressed in m^2/ha, and indicates the density of trunks in forest stands.

- intra-plot diversity (evenness), computed as the Gini-Simpson index (Hill, 2003), with $\sharp(s)$ referring to the number of trees with species s and $\sharp(T)$ referring to the total number of trees inside each plot: $1 - \sum_{s \in S} \left(\frac{\sharp(s)}{\sharp(T)}\right)^2$. This provides an index in the 0-1 range describing the species heterogeneity at the stand level, with high values indicating a high heterogeneity.

- extra-plot diversity (species richness), computed as the number of species present in a plot: $\sharp(S)$. In the Quebec dataset, this indicator ranges from 1 to 8 species, and is interpreted as another measure of diversity.

- shade tolerance index, a new metric introduced by Strigul and Florescu (2012) and Lienard et al. (2014) (under review) describing the shade tolerance rank of species r: $\sum_{i \in T} \frac{\sharp(s) r_i}{\sharp(T)}$. This index ranges from 0 to 1, with high values denoting forest stands composed of typically late successional species and low values denoting forest stands composed of typically early successional species in Quebec (Lienard et al., 2014, under review).

- average age, computed as the average of tree ages a : $\sum_{i \in T} \frac{a_i}{\sharp(T)}$. This commonly-used indicator approximates the stand age in the forest inventory analysis (see Strigul et al. 2012 for a discussion of this characteristic).

Statistical relations of these stand-level characteristics were analyzed using standard multivariate methods. First, we computed the Pearson correlation coefficients both in the whole dataset and in the dataset broken down in decades (to avoid biases due to their temporal autocorrelation). We then performed a principal component analysis (PCA) to examine (a) the number of components needed to explain most of the variance as well as (b) the projection of characteristics in the space defined by these components.

In general, it is possible for a multidimensional model to operate on the space of principal components. Such a model would (a) project the characteristics into the low-dimensional space given by the principal components, then (b) predict their dynamics in this new space, and finally (c) perform the inverse transformation to obtain predictions on the characteristics. In our application to the Quebec dataset, we discovered that four uncorrelated characteristics approximate well the principal component space (namely biomass, average age of trees, Gini-Simpson and shade tolerance indexes, *cf.* Results). The multidimensional model employs this approximation and is based on transition matrices of these forest characteristics. It substantially simplify interpretation of modeling predictions.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Prior to the computation of transition matrices in the Markov chain framework, it is necessary to discretize continuous variables into distinct states (Strigul et al., 2012). The general approach is to subdivide data into uniformly spaced states, with a precision that is small enough to capture the details of the distribution but large enough to be insensitive to statistical noise in the dataset. In addition, the computational effort needed to infer transition matrices is proportional to the square of the number of states, and available computational power may constitute a practical limitation to the number of states. In the Quebec dataset, the stand-level characteristics span different ranges (see Fig. 1 in Appendices), with the biomass distribution in particular showing a long tail for the highest values. In order to capture enough details of the distributions of the Quebec characteristics, we opted to remove plots in the long tail of the biomass (those with a biomass higher than 50,000 kg/ha, representing roughly 4% of the total dataset) and then subdivided the remaining plots into 25 biomass states. An alternative approach would be to merge the rarely occurring high-biomass states into the last state as was implemented in Strigul et al. (2012). We conducted a comparison of these two approaches and found no significant differences. For the other characteristics investigated (i.e. the internal diversity, shade tolerance index, and average age), we found that 10 states were enough to capture their distributions with sufficient detail.

3.3 Step 2: Gibbs sampling methodology

Inferring a Markov Chain model for characteristics computed with field data sampled at irregular intervals is a challenging problem. Indeed, the usual direct approach of establishing the n-year transition matrix by simply counting the number of times each state changes to another after n years can not be employed in most forest inventories, as successive measurements on the same plot are not made with constant time intervals. This irregularity in sampling results in states of the forest plots that are not observed, and can be modeled as missing data. Two classes of algorithms can be used to model a transition matrix describing the dynamics of both observed and missing data: expectation-maximization (EM) and Markov chain Monte Carlo (MCMC), of which Gibbs sampling is a specific implementation. Both classes of algorithms are iterative and can be used to find the transition matrix that best fits the observed data. EM algorithms consist of the iteration of two steps: in the expectation step the likelihood of transition matrices is explicitly computed given the distribution of the missing data inferred from the previous transition matrix estimate, and in the maximization step a new transition matrix maximizing this likelihood is chosen as the new estimate (Dempster et al., 1977). MCMC algorithms can be seen as the Bayesian counterpart of EM algorithms, as at each iteration a new transition matrix is stochastically drawn with the prior information of estimated missing data, and in turn new estimates for the missing data are stochastically drawn from the new transition matrix (Robert and Casella, 2004). EMs are deterministic algorithms, and as such they will always converge to the same transition matrix with the same starting conditions; conversely, MCMCs are stochastic and are not guaranteed to converge toward the same estimate with different random seeds. While both algorithms are arguably usable in our context, the ease of implementation and lower computational cost of MCMC algorithms led us to prefer them over EM (Deltour et al., 1999). We selected Gibbs sampling as a flexible MCMC implementation (Geman and Geman, 1984). We provide in the following a brief presentation of Gibbs sampling. Additional

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

implementation details are in Appendix 1.2, and we refer to Robert and Casella (2004) for the general principles underlying MCMC algorithms and to Pasanisi et al. (2012) for an extended description of Gibbs sampling to infer transition probabilities in temporal sequences.

To apply Gibbs sampling for the estimation of the transition matrices, it is required to include macroscopic plot characteristics in a set of temporal sequences. For each plot p, this is done by inserting each characteristic $s_{(p,i)}$ measured in the i-th year at position i of a row vector S_p representing the temporal sequence of this plot. For example, if a plot p was sampled only at years 1 and 3 during a 5-year inventory, allowing for the computation of characteristics $s_{(p,1)}$ and $s_{(p,3)}$, then its sequence would be the row vector $S_p = [s_{(p,1)}, \bullet, s_{(p,3)}, \bullet, \bullet]$, where $\bullet$ denotes a missing value. The sequences are mostly composed of unknown values as only a fraction of the forest plots were surveyed each year. In the application to the Quebec dataset, a reduction of the size of these temporal sequences was performed. However, it is not a pre-requisite for the application of Gibbs sampling (see Appendix 1.2 for a detailed description of this reduction and an illustrative example). Let further Y be the matrix constructed using all the sequences S, with rows corresponding to successive measures of different plots and columns corresponding to different years. The preliminary step of Gibbs sampling consists of replacing the missing values $\bullet$ in Y at random, resulting in so-called augmented data $Z^{[0]}$. Then, the two following steps are iterated a fixed number of times H:

1. in the **parameter estimation** step, we draw a new transition matrix $\Phi_i^{[h]}$ conditional on the augmented data Z^{h-1}:

$$\Phi_i^{[h]} | Z^{[h-1]} \sim Dir(\gamma_{i,1} + w_{i,1}^{[h-1]}, ..., \gamma_{i,r} + w_{i,r}^{[h-1]}) \tag{3}$$

   with Dir is the Dirichlet distribution, γ are biasing factors set here uniformly to 1 as we include no prior knowledge on the shape of the transition matrix (Pasanisi et al., 2012), and $w_{i,j}$ are the sufficient statistics defined as

$$w_{i,j} = \sum_{t \,\in\, \text{years}} \quad \sum_{k \,\in\, \text{plots}} \mathbb{1}_{\{y_{k,t-1}=s_i \;\&\; y_{k,t}=s_j\}} \tag{4}$$

2. in the **data augmentation** step, we draw new values for the missing states:

$$\text{for the earliest data } t = 1, \quad \mathbb{P}(z_{k,1}^{[h]} = s_j | z_{k,2}^{[h-1]} = s_i, \Phi^{[h]}) \;\propto \Phi_{j,i}^{[h]} \tag{5}$$

$$\text{for the latest data } t = T, \quad \mathbb{P}(z_{k,T}^{[h]} = s_j | z_{k,T-1}^{[h]} = s_i, \Phi^{[h]}) \;\propto \Phi_{i,j}^{[h]} \tag{6}$$

$$\text{otherwise}, \quad \mathbb{P}(z_{k,1}^{[h]} = s_j | z_{k,t-1}^{[h]} = s_{i_1}, z_{k,t+1}^{[h-1]} = s_{i_2}, \Phi^{[h]}) \;\propto \Phi_{i_1,j}^{[h]} \times \Phi_{j,i_2}^{[h]} \tag{7}$$

As Gibbs sampling is initialized by completing the missing values at random, the first iterations will likely result in transition matrices far away from the optimal. The usual workaround is to ignore the first B transition matrices corresponding to so-called "burn-in" period, leaving only $H - B$ matrices. Furthermore, as Gibbs sampling relies on a stochastic

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

exploration of the search space, a good practice to ensure that Gibbs sampling converged to the optimal transition matrix is to run the whole algorithm R times. There are no general guidelines for setting the H, B and R parameters (Robert and Casella, 2004). We empirically settled with $H = 1000$, $B = 100$ and $R = 50$ in order to ensure that the transition matrices were reproducible for the Quebec dataset, which required $R \times H = 50000$ iterations of parameter estimation and data augmentation steps, which resulted in $R \times (H - B) = 45000$ transition matrices. This process was repeated independently for each plot characteristic. The algorithm was implemented in R and took a total runtime of 4 days on a personal computer to compute the transition matrices for all 4 characteristics studied here.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4 Results

4.1 Multivariate analysis of stand characteristics

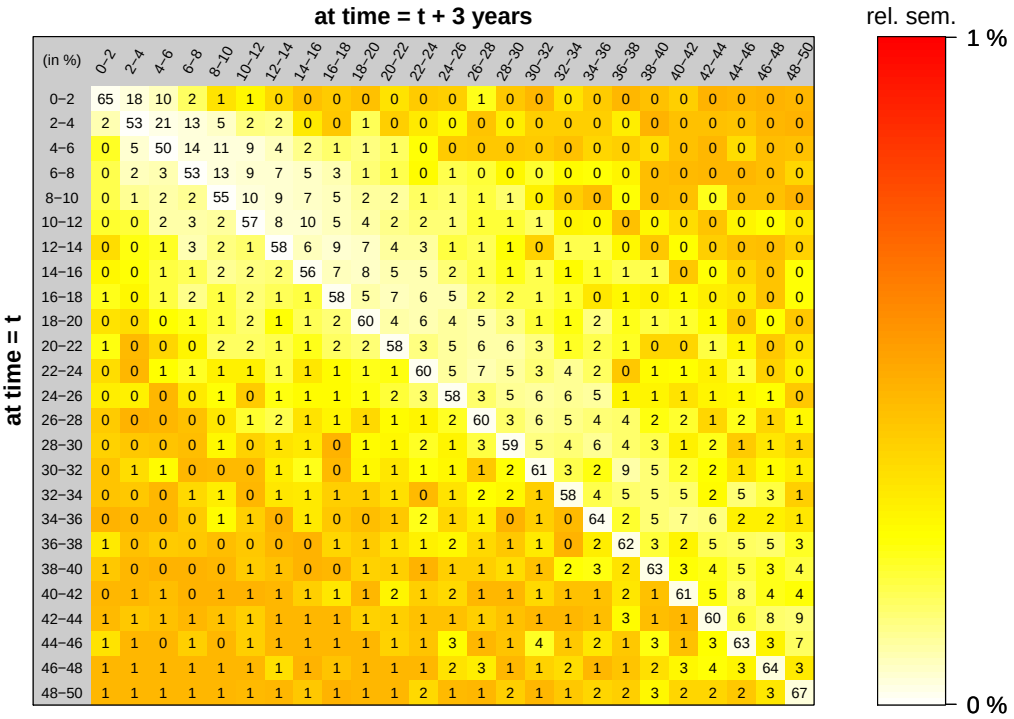
The correlation analysis performed on the Quebec forest inventory (Perron et al., 2011, Appendix 1.1) revealed that biomass and basal area were highly correlated ($r = 0.96$), as well as the external and internal diversity indices ($r = 0.90$, see Appendix 1.3 for the other coefficients). These correlations are further preserved when the correlation analysis is done separately on each decade, from the 1970s until the 2000s (*cf.* tables in Appendix 1.3), confirming the presence of time-independent strong correlations between these two pairs of characteristics.

A PCA applied to the dataset further confirmed that the biomass and basal area on one hand, as well as the external and internal diversity on the other hand, have nearly identical vectors in the principal components space (*cf.* Appendix 1.4). Furthermore, this analysis showed that 4 principal components are required to adequately explain variance in the data; using 3 components accounts for only 87 % of the variance, while 4 components explain up to 98 % of the variance. The PCA revealed that biomass, the internal diversity index, the shade tolerance index, and the average age are close approximations of the different principal components and explain most of the variance. Therefore, these variables have been employed in the following analysis.

4.2 Interpretation of the transition matrices

We present here in detail the transition matrix for biomass with a 3-year time interval, shown in Fig. 2 (the other characteristics are to be found in Appendices, in Figs. 4 and 5). In this matrix, each value at row i and column j corresponds to the probability of transition from state i into state j after 3 years. By definition, rows sum to 100%. This transition matrix, as with the others in Appendix, is dominated by its diagonal elements, which is expected because few plots show large changes in a given 3-year period. The values below the diagonal correspond to transitions to a lower state (hence, they can be interpreted as the probabilities of disturbance), while values above the diagonal correspond to transitions to a higher state (i.e., growth). The transitions in the first column of the matrix correspond to major disturbances, where the stand transitions to a very low biomass condition. As the probabilities above the diagonal are larger than below the diagonal, the overall 3-year prediction is of an increase in biomass. This matrix also shows that plots with a biomass larger than 40,000 kg/ha have a roughly uniform 10% probability of ending with a biomass of less than 20 000 kg/ha 3 years later, which is interpreted as the probability of high-biomass stand to go through a moderate to high disturbance.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(in %)	0–2	2–4	4–6	6–8	8–10	10–12	12–14	14–16	16–18	18–20	20–22	22–24	24–26	26–28	28–30	30–32	32–34	34–36	36–38	38–40	40–42	42–44	44–46	46–48	48–50
0–2	65	18	10	2	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
2–4	2	53	21	13	5	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4–6	0	5	50	14	11	9	4	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6–8	0	2	3	53	13	9	7	5	3	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8–10	0	1	2	2	55	10	9	7	5	2	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10–12	0	0	2	3	2	57	8	10	5	4	2	2	1	1	1	1	0	0	0	0	0	0	0	0	0
12–14	0	0	1	3	2	1	58	6	9	7	4	3	1	1	1	0	1	1	0	0	0	0	0	0	0
14–16	0	0	1	1	2	2	2	56	7	8	5	5	2	1	1	1	1	1	1	1	0	0	0	0	0
16–18	1	0	1	2	1	2	1	1	58	5	7	6	5	2	2	1	1	0	1	0	1	0	0	0	0
18–20	0	0	0	1	1	2	1	1	2	60	4	6	4	5	3	1	1	2	1	1	1	1	0	0	0
20–22	1	0	0	0	2	2	1	1	2	2	58	3	5	6	6	3	1	2	1	0	0	1	1	0	0
22–24	0	0	1	1	1	1	1	1	1	1	1	60	5	7	5	3	4	2	0	1	1	1	1	0	0
24–26	0	0	0	0	1	0	1	1	1	1	2	3	58	3	5	6	6	5	1	1	1	1	1	1	0
26–28	0	0	0	0	0	1	2	1	1	1	1	1	2	60	3	6	5	4	4	2	2	1	2	1	1
28–30	0	0	0	0	1	0	1	1	0	1	1	2	1	3	59	5	4	6	4	3	1	2	1	1	1
30–32	0	1	1	0	0	0	1	1	0	1	1	1	1	1	2	61	3	2	9	5	2	2	1	1	1
32–34	0	0	0	1	1	0	1	1	1	1	1	0	1	2	2	1	58	4	5	5	5	2	5	3	1
34–36	0	0	0	0	1	1	0	1	0	0	1	2	1	1	0	1	0	64	2	5	7	6	2	2	1
36–38	1	0	0	0	0	0	0	0	1	1	1	1	2	1	1	1	0	2	62	3	2	5	5	5	3
38–40	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	2	3	2	63	3	4	5	3	4
40–42	0	1	1	0	1	1	1	1	1	1	2	1	2	1	1	1	1	1	2	1	61	5	8	4	4
42–44	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	60	6	8	9
44–46	1	1	0	1	0	1	1	1	1	1	1	1	3	1	1	4	1	2	1	3	1	3	63	3	7
46–48	1	1	1	1	1	1	1	1	1	1	1	1	2	3	1	1	2	1	1	2	3	4	3	64	3
48–50	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	2	2	3	2	2	2	3	67

Fig. 2. 3-year transition matrix for the biomass. The states are the biomass ranges in 10^3 kg/ha, spanning from $0-2$ to $48-50$ 10^3 kg/ha, and represented here on the left and on top of the matrix. The values $M(i,j)$ inside the matrix correspond to the rounded probability of transition from state i to state j. The color represents the relative standard error of the mean and indicates the robustness of the stochastic search, as explained in Section 4.3. Lighter colors thus indicate a better confidence in the transition value; all relative standard errors of the mean (RSEM) are below 1%, corresponding to a very high confidence, and furthermore the smallest errors are found for the higher transition probabilities close to the diagonal.

4.3 Model validation

Two main types of error should be considered when designing a model with a parameter search based on real data. The first error relates to the robustness and efficiency of the estimation of the optimal transition matrix, which was performed with Gibbs sampling in our case. The second type of error encompasses more broadly the capacity of the chosen theoretical framework to predict the system beyond the range of the dataset. In our case, the theoretical framework we relied on is patch-mosaic concept, implemented with the Markov chain framework, to describe the dynamics of our four characteristics.

To estimate the errors of the parameter search, we used the $R(H-B)$ transition matrices to compute for each transition the standard error of the mean (SEM) and the relative standard error of the mean (RSEM, defined as the ratio of the SEM over the transition probability, and expressed as a percentage). The SEM were below 1% throughout the matrices, with the highest errors occurring for very low transition probabilities (i.e., far from the diag-

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

onals). Furthermore, the RSEM were very low, and particularly so for the transitions with the highest probability (Fig. 2 in main text as well as Figs. 4 and 5 in Appendix). We finally computed the SEM in the long-term predicted equilibriums and found values below 0.01%, strengthening the conclusion that negligible errors are to be attributed to the stochastic fit procedure.

To estimate the more general errors in the ability of a Markov-Chain model to predict future forest characteristics based on a dataset such as the Quebec dataset, we performed two additional validations of our methodology. In the first, we ran the Gibbs sampler with only the first 18 years of records (from 1970 to 1988). We then used the model to predict forest state for the period corresponding to the second half of the dataset (i.e., 1989 to 2007), and we compared the predicted dynamics with the aggregated distribution of the second half of the dataset (Fig. 3). Overall, the predictions were highly accurate, with R^2 between observation and prediction ranging from 0.8 to 0.95, indicating that the second half of the dataset is predictable with a Markov chain model based solely on the first half. In the second validation, we randomly split the data into two sets, regardless of year. We then computed the transition matrix and corresponding equilibrium conditions for each half (Fig. 6 in Appendix). Here again, the predictions match closely with values of R^2 higher than 0.98 for the internal diversity, shade tolerance index and average age. The R^2 was near 0.6 for the biomass, mainly because as it has been discretized into 25 states instead of 10 for the other characteristics, making the R^2 metric more sensitive to small differences in the long term predictions of individual states (typically around 1%). This second validation overall showed that the data contained in the inventory is redundant, and that half of it is enough to provide highly accurate long-term estimates for the internal diversity, shade tolerance index and average age. Considering only half of the data at random would likely result in errors of around 1% in the long-term estimates of the biomass.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

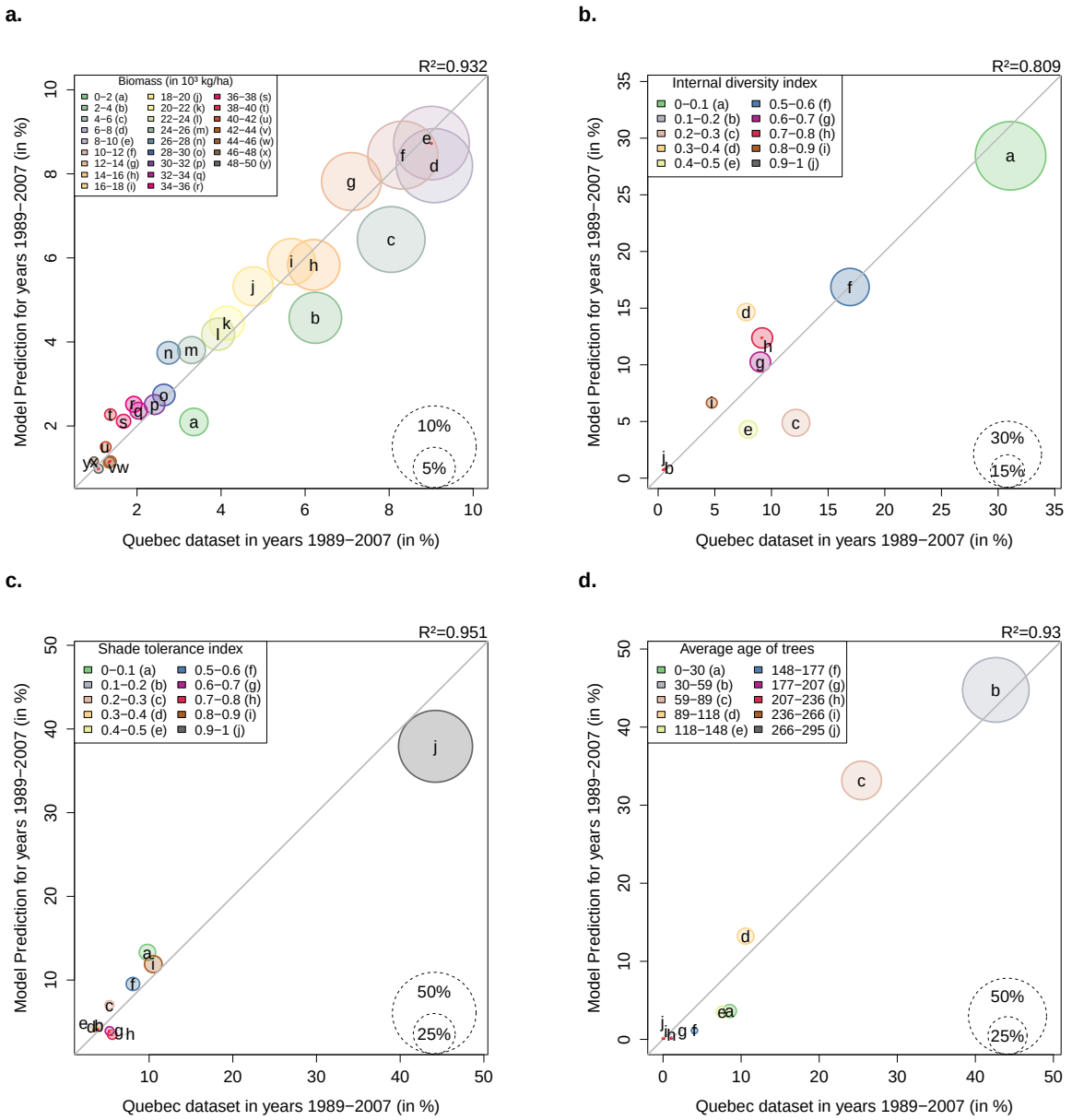


Fig. 3. Results of model validation, showing the second half of the dataset *vs* the model prediction for the classes of each characteristic. For each class, the circle size denotes the number of stands belonging to it in the real dataset. The R^2 measure is indicated on the top right of each plot. The model used to make the prediction was computed using only the first half of the dataset, corresponding to years 1970 to 1988 (see Materials & Methods for details).

4.4 Predictions of temporal dynamics and long-term equilibrium

We applied the inferred transition matrix to predict the state of forest in 2010s, 2020s and 2030s based on their distribution in 2000s. We also predicted the long-term dynamics of the forest stands, by computing the equilibrium states of the transition matrices. Overall,

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the predictions showed an increase in biomass and stand age (Fig. 4 e and h), along with a slight increase in biodiversity (Fig. 4 f) and a slight decrease of the prevalence of late successional species accompanied by a slight increase of early successional species (Fig. 4 g). These predictions are obvious for the biomass and average age of trees by looking at their distributions in the existing dataset (Fig. 4 a and d), while they are less clearly seen when looking at the average distributions of the biodiversity and shade tolerance index (Fig. 4 b and c).

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

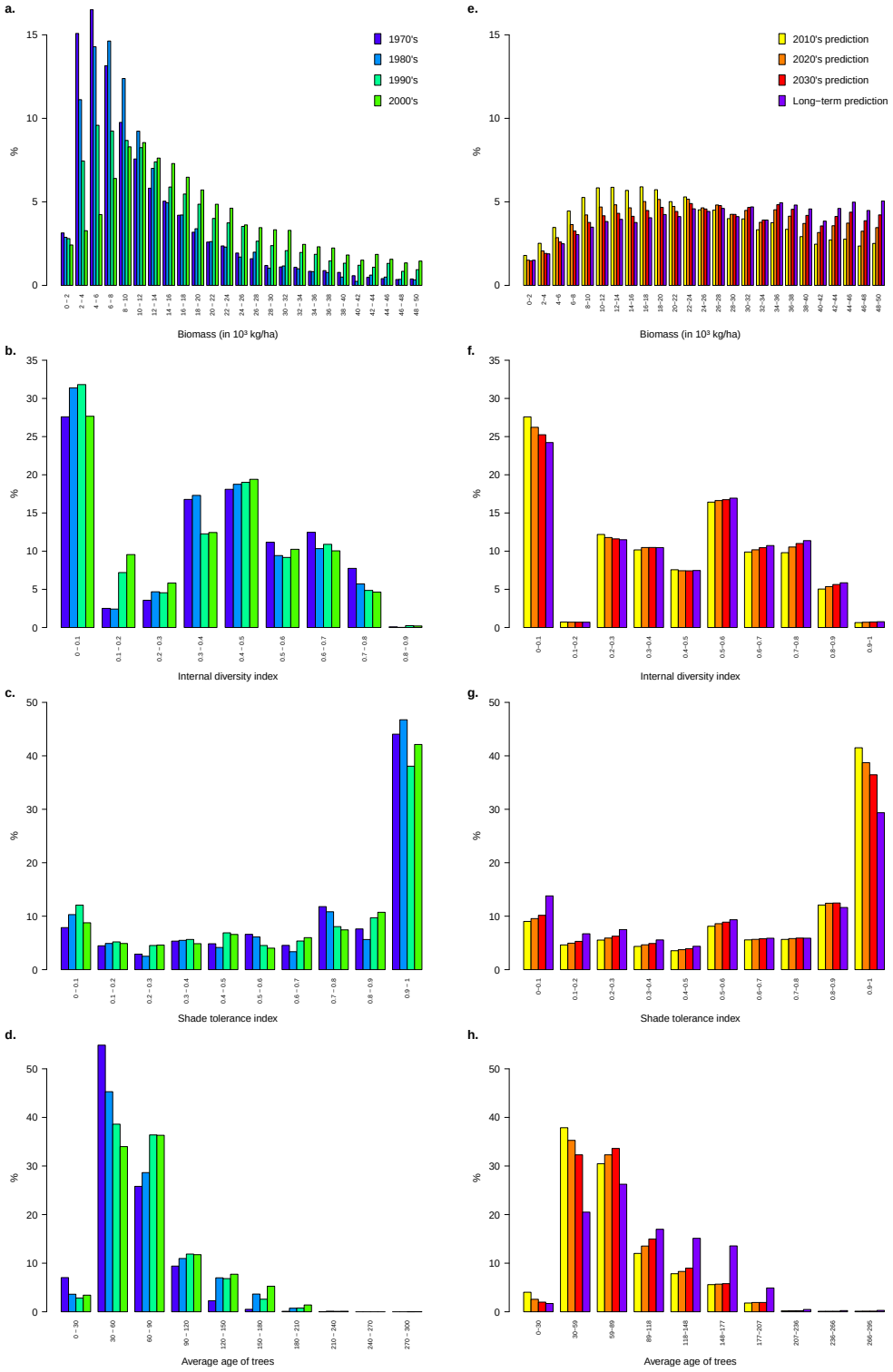


Fig. 4. Current distribution of relevant characteristics from the database, along with the long-term predictions of our models.

These long-term predictions were reached at different timescales depending on the char-

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

acteristics. For biomass, equilibrium was reached by approximately year 2030, but the other characteristics, and in particular the average age of trees in plots, showed much slower dynamics to reach their equilibria (Fig. 4 e to h). The model predicted average relative changes of +38.9% and +14.2% by the 2030s for biomass and stand age, and +44.0% and +37.9% by the time they reach their long-term equilibrium state. Relative changes for the Gini-Simpson diversity index were +5.2% by the 2030s and +7.1% in the long term, and early successional species will become slightly more abundant with a change of -4.7% of the shade tolerance index by the 2030s and -13.6% in the long term.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5 Discussion

We developed a data-intensive approach to multiple-dimensional modeling of forest dynamics. The modeling steps include 1) dimensional analysis of forest inventory data, 2) extraction of non-correlated dimensions, and 3) the application of stochastic optimization to compute probability transition matrices for each dimension. We applied this approach to the Quebec forest inventory dataset and validated the model using two independent subsets of data. Our study demonstrates that at least four uncorrelated dimensions are required to describe Quebec forests: the biomass, biodiversity, shade tolerance index and averaged age of trees. The most pronounced changes predicted for Quebec forests are increases in biomass and stand age. Our model also predicted smaller increases in biodiversity in the prevalence of early successional species. Our results demonstrate the utility of this methodology in predicting long-term forest dynamics given highly dimensional, irregularly sampled data; the model was computationally efficient and highly predictive. Therefore, the framework will be useful both in applied contexts (e.g., conservation, silviculture) as well as in developing our conceptual understanding of how forested ecosystems are organized.

5.1 Predictions for forest dynamics in Quebec

Our model made several notable predictions about future forest dynamics in Quebec. The most pronounced predicted changes are substantial short-term increase in biomass and a longer-term increase in average age of trees (Fig. 4). The model also predicted smaller changes in biodiversity and the shade tolerance index. The increase in biomass is intuitively consistent with the increase in stand age, demonstrating a progression toward more mature stands. Additionally, the increase in late successional species is consistent with increasing biodiversity. To understand stand maturation occurring with the small increase in the prevalence of late successional species, we must recall that neither biomass nor stand age are significantly correlated with shade tolerance index in the dataset (e.g., $r = -0.02$ with 95% confidence interval [-0.03,-0.01] for biomass and shade tolerance). Thus, it is unsurprising that the predictions are not correlated. Further, the predicted changes happen with different temporal dynamics and have different magnitudes, and have probably distinct mechanisms. In particular, while biomass and stand age are affected by both individual tree growth (leading to an increase) and disturbances (leading to a decrease), the shade tolerance index is affected only by disturbance. Mechanistically, small disturbances (e.g., individual tree mortality) will typically promote the recruitment of early successional species into the canopy through gap dynamics (Lienard et al., 2014, under review). On the other hand, intermediate and large-scale disturbances will facilitate early successional species via the development of large canopy openings (Lienard et al., 2014, under review). Thus, increase of intermediate and large-scale disturbances may promote early successional species, while the overall increase in biomass and stand age would result largely from individual tree growth.

The accurate prediction of the second half of the dataset obtained using only the first half of the dataset demonstrate that the natural disturbance regime in the forest plots sampled in the Quebec inventory did not change substantially over the last 30 years. In the context of global warming, this could mean either that (a) there is no substantial consequence yet on the macroscopic dynamics of Quebec forests or that (b) the climatic change consequences

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were already present in the first half of the dataset. In both cases, the inclusion in the transition matrices of future disturbances induced by climatic change (e.g. the increase of forest fire reviewed in Flannigan et al., 2009) could be a promising follow-up of our work by providing quantitative insights on the consequences of global warming on forests. However, the study of changes in disturbances was not the focus of the current study, and we have to be careful in generalizing this conclusion as the gradual non-stationary disturbance regimes might take from 50 to 100 years to show significant departures.

5.2 A data-intensive approach to understand forest dynamics

Modeling complex adaptive systems such as forest ecosystems requires capturing the dynamics of biological units at multiple scales and in multiple dimensions (Levin, 1998, 2003). Ideally, a model based on the physiological processes and interactions of individual organisms should simulate the observed forest structure and predict forest dynamics over different time horizons and environmental variables. However, such mechanistic modeling is very challenging as interactions between individual organisms within the forest stand result in new properties at the stand level, where essential mechanisms, spatial dimensions of variables, and functional relationships between variables are largely unknown. Given these unknowns, a data-intensive approach can be useful for gaining insight into ecosystem structure and functioning provided that sufficient amounts of relevant data are available. The core idea is then to use data-mining tools for scientific discovery in contrast with the traditional hypothesis-driven discovery process (Kelling et al., 2009, Michener and Jones, 2012).

Despite being largely empirical, the data-intensive approach we developed allows us to contribute to a mechanistic understanding of natural disturbances and forest self-organization. The transition matrices obtained with our approach should coincide with transition matrices obtained with a more mechanistic approach. However, such a model would require comprehensive knowledge of magnitudes and frequencies of all natural and anthropogenic disturbances as well as quantitative estimates of how these disturbances affect forest stand characteristics. However, mechanistic interpretations of the transition matrices obtained by our approach could improve our understanding of forest dynamics. The transition rates are estimated based on all recorded transitions and therefore account for all the complexity associated with succession. The matrices we estimated (see Fig. 2 in main text and Figs. 4 and 5 in Appendices) are obviously more complex than mechanisms given in classic theories, which include birth-and-disaster Markov chains (Feller, 1971), forest fire models (Van Wagner, 1978), and advection-reaction equations for patch dynamics (Levin and Paine, 1974). This suggests that these classic theories are not sufficient to describe the complexity of real forests. A potential limitation to a mechanistic interpretation of the transition matrices arises from the Markovian assumption that the transition toward the next state depends solely on the current state. This same assumption is also present in other frameworks such as the continuous patch-mosaic models represented by first-order partial differential equations (Strigul, 2012). If this assumption is not valid, it could bias mechanistically constructed models. Therefore, this assumption warrants further attention in future studies as it has not been yet comprehensively evaluated in forest modeling.

In this study we have analyzed the Quebec forest inventories without explicitly taking into account geographical location of plots, environmental and climatic variables. We have

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

obtained transition matrices covering temperate to boreal forests, with a disturbance regime varying from canopy gaps to disastrous fires. We have repeated the developed approach after subdividing the Quebec dataset into the major ecological domains and have not observed substantial differences between the resulting transition matrices and the general matrices presented in this study (Liénard et al. unpublished data). In addition to this, the biomass transition matrices computed for the Lake States in the US (see Strigul et al. (2012) Tables 2 and 3) and the shade tolerance index transition matrices computed in northeastern parts of the US (Lienard et al., 2014, under review) using are quite similar to the ones presented in this study. It is quite amazing in fact that we could represent macroscopic properties given the neglection of geography. We hypothesize that the macroscopic forest stand dynamics as well as disturbance regimes have substantial similarities across a large number of boreal and temperate forest types, and this will be specifically addressed in our future studies. We believe that the ability to make broad predictions on the forest stand dynamics without going into the fine details of geography is one of the major strengths of our approach.

The application of MCMC procedures allows to compute transition matrices for datasets with irregular sampling intervals and sample sizes. While Gibbs sampling has been introduced 30 years ago (Geman and Geman, 1984), its application to handle missing data in ecology has been mostly limited to stochastic patch occupancy models with a low number of free parameters (5-6) and either artificially simulated data or relatively restricted datasets (e.g. 72-228 resampled locations in ter Braak and Etienne, 2003, Harrison et al., 2011, Risk et al., 2011). From the technical point of view, our application of MCMC differs by taking advantage of the absolute time independence of Markov chains (allowing us to align subsequences starting with a known observation, see Methods and Appendix 1.1). This makes the use of MCMC possible in a data-intensive context, in which both the number of free parameters (600 for the biomass matrix, 90 for each of the biodiversity, shade tolerance and stand age matrices) and the number of samples (32,552) constitute increases of several orders of magnitude. Similar irregularity problems are quite common in ecological datasets, and our technique may have numerous applications beyond the statistical analysis of forest inventories. This methodology can also be applied to other datasets, even with regular samplings, and the same methodology can be applied to deduce transitions with a finer temporal scale.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements

This work was partially supported by a grant from the Simons Foundation (♯283770 to N.S.) and a Washington State University New Faculty SEED grant. D.G. also acknowledges the financial support from a Strategic Grant of NSERC.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Chave, J. (2013). The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecology Letters*, 16:4–16.

Deltour, I., Richardson, S., and Hesran, J.-Y. L. (1999). Stochastic algorithms for markov models estimation with intermittent missing data. *Biometrics*, 55(2):565–573.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. Wiley.

Flannigan, M. D., Krawchuk, M. A., de Groot, W. J., Wotton, B. M., and Gowman, L. M. (2009). Implications of changing climate for global wildland fire. *International Journal of Wildland Fire*, 18(5):483–507.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721–741.

Harrison, P. J., Hanski, I., and Ovaskainen, O. (2011). Bayesian state-space modeling of metapopulation dynamics in the glanville fritillary butterfly. *Ecological Monographs*, 81(4):581–598.

Hill, M. O. (2003). Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432.

Jenkins, J., Chojnacky, D., Heath, L., and Birdsey, R. (2003). National-scale biomass estimators for united states tree species. *Forest Science*, 49(1):12–35.

Kelling, S., Hochachka, W., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience*, 59(7):613.

Kohyama, T. (2006). The effect of patch demography on the community structure of forest trees. *Ecological Research*, 21(3):346–355.

Kohyama, T., Suzuki, E., Partomihardjo, T., and Yamada, T. (2001). Dynamic steady state of patch-mosaic tree size structure of a mixed dipterocarp forest regulated by local crowding. *Ecological Research*, 16(1):85–98.

Leslie, P. H. (1945). On the use of matrices in certain population mathematics. *Biometrika*, 33:183–212.

Levin, S. A. (1998). Ecosystems and the biosphere as complex adaptive systems. *Ecosystems*, 1(5):pp. 431–436.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Levin, S. A. (1999). *Fragile dominion: complexity and the commons.* Perseus Publishing, Cambridge, MA.

Levin, S. A. (2003). Complex adaptive systems: Exploring the known, the unknown and the unknowable. *American Mathematical Society*, 40:3–19.

Levin, S. A. and Paine, R. T. (1974). Disturbance, patch formation, and community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 71(7):2744–2747.

Lienard, J., Florescu, I., and Strigul, N. (2014). An appraisal of the classic forest succession paradigm with the shade tolerance index. *Under review.*

Michener, W. K. and Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, 27(2):85 – 93.

O'Neill, R. V., DeAngelis, D., Waide, J., and Allen, T. F. H. (1986). *A hierarchical concept of ecosystems.* Princeton University Press.

Pacala, S. W., Canham, C. D., Saponara, J., Silander Jr., J. A., Kobe, R. K., and Ribbens, E. (1996). Forest models defined by field measurements: estimation, error analysis and dynamics. *Ecological Monographs*, 66(1):1–43.

Pasanisi, A., Fu, S., and Bousquet, N. (2012). Estimating discrete markov models from various incomplete data schemes. *Computational Statistics & Data Analysis*, 56(9):2609–2625.

Perron, J., Morin, P., et al. (2011). Normes d´inventaire forestier: Placettes-échantillons permanentes.

Peters, D. P. C., Bestelmeyer, B. T., and Turner, M. G. (2007). Cross-scale interactions and changing pattern-process relationships: Consequences for system dynamics. *Ecosystems*, 10(5):790–796.

Risk, B. B., De Valpine, P., and Beissinger, S. R. (2011). A robust-design formulation of the incidence function model of metapopulation dynamics applied to two species of rails. *Ecology*, 92(2):462–474.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*, volume 319. Citeseer.

Strigul, N. (2012). Individual-based models and scaling methods for ecological forestry: implications of tree phenotypic plasticity. In Garcia, J. and Casero, J., editors, *Sustainable Forest Management*, pages 359–384. InTech, Rijeka, Croatia.

Strigul, N. and Florescu, I. (2012). Statistical characteristics of forest succession. In *Abstracts of 97th ESA Annual Meeting*. Ecological Society of America.

Strigul, N., Florescu, I., Welden, A. R., and Michalczewski, F. (2012). Modelling of forest stand dynamics using markov chains. *Environmental Modelling and Software*, 31:64 – 75.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Strigul, N., Pristinski, D., Purves, D., Dushoff, J., and Pacala, S. (2008). Scaling from trees to forests: Tractable macroscopic equations for forest dynamics. *Ecological Monographs*, 78(4):523–545.

ter Braak, C. J. and Etienne, R. S. (2003). Improved bayesian analysis of metapopulation data with an application to a tree frog metapopulation. *Ecology*, 84(1):231–241.

Van Wagner, C. E. (1978). Age-class distribution and the forest fire cycle. *Canadian Journal of Forest Research*, 8(2):220–227.

Watt, A. S. (1947). Pattern and process in the plant community. *J.Ecol.*, 35:1–22.

Wu, J. and Loucks, O. L. (1996). From balance of nature to hierarchical patch dynamics: A paradigm shift in ecology. *Quarterly Review of Biology*, 70(4):439–466.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix to "Data-intensive multidimensional modeling of forest dynamics"

Jean F. Liénard [1], Dominique Gravel [2], Matthew V. Talluto [2], Nikolay S. Strigul [1]*

[1]: Department of Mathematics, Washington State University, Washington

[2]: Département de Biologie, Universit du Québec à Rimouski, Québec

*: corresponding author email: nick.strigul@vancouver.wsu.edu

Appendix 1.1 Description of the Quebec Dataset

The Quebec forest inventory program was started in 1970s and is still ongoing nowadays, using a constant methodology across the years (Perron et al., 2011). In this study, we considered only permanent plots that were sampled at least twice since the beginning of the inventory, resulting in 32552 plot measurements taken from 11660 different locations throughout Quebec. These permanent plots are circular of area $400m^2$. At the time of establishment as well as at each successive measurements, every tree with a diameter greater than 90 mm was numbered with paint, measured and finally recorded in the database (Perron et al., 2011, Duchesne and Ouimet, 2009). For the computation of the plot characteristics, we specifically relied on the species determined by the forester, the D.B.H. measured using a diameter tape to the nearest millimeter and the age measured from various sources.

The version of the Quebec dataset used in our study spans from 1970 until 2007 included. Figure 1 shows the distributions of the measurements across the years in the database, as well as the repartition of the intervals between two successive measurements.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

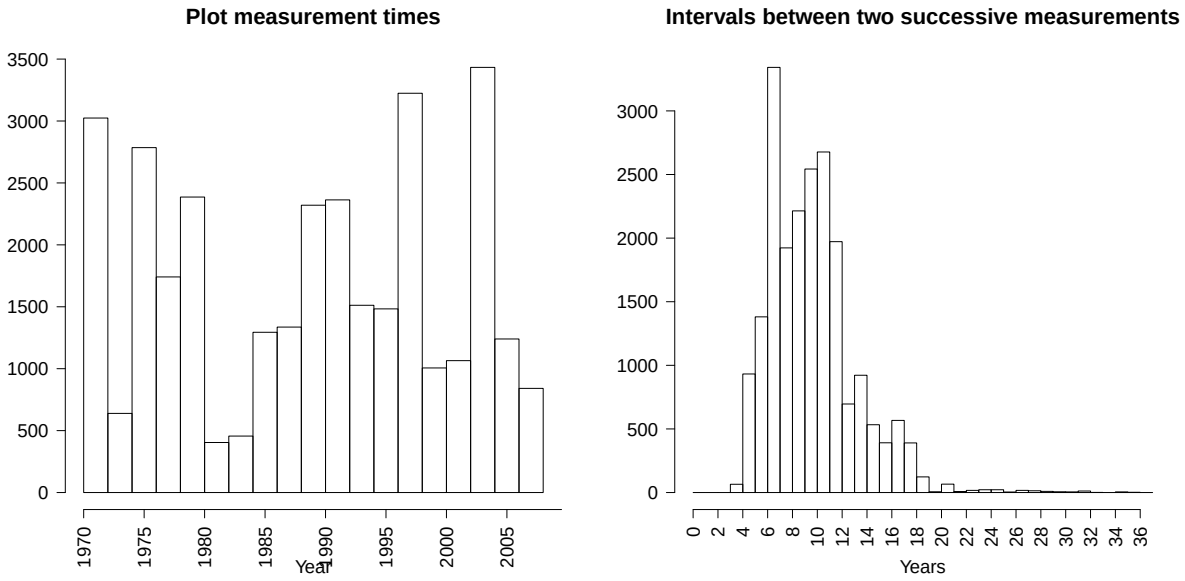


Figure 1: Distribution of the times of plot measurement (left) and distribution of intervals between successive measurements of the same plot (right) in the Quebec dataset.

The distributions of stand characteristics across all the database records is plotted in Figure 2. Different distribution patterns are evident. In particular, the biomass, basal area and stand age distributions reveal the presence of long tails for the high values.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

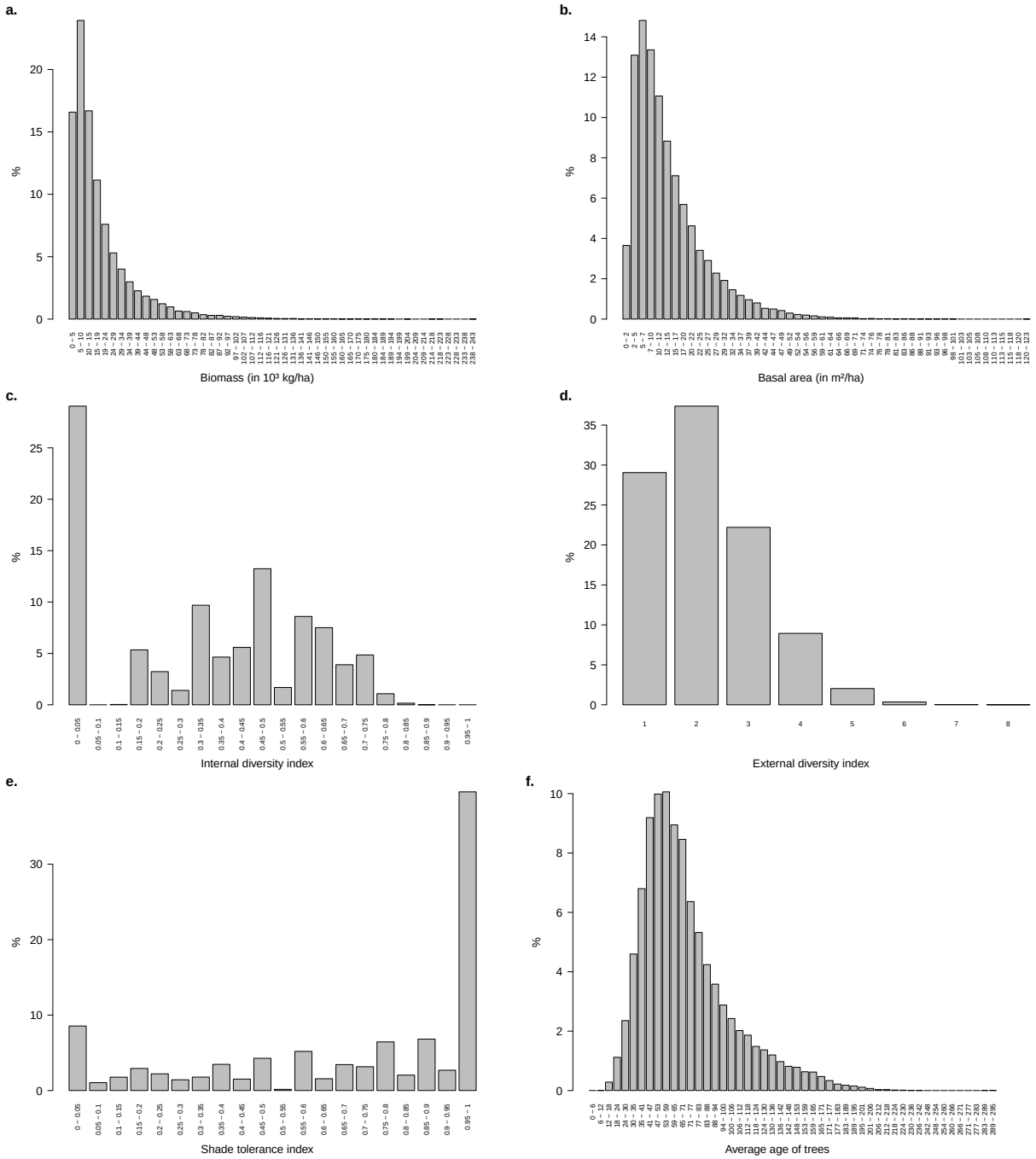


Figure 2: Distribution of stand characteristics

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix 1.2 Gibbs sampling implementation

Forest inventory protocols constitute so-called ignorable data collection mechanisms (Little and Rubin, 1987), as the plots are sampled according to a pre-defined policy that depends mainly on organizational constraints on the foresters' activity, and not on the actual composition of the plots (Perron et al., 2011). This allows for the use of Gibbs sampling as described in main text, which follows the implementation of Pasanisi et al. (2012).

Several steps of data preparation are essential to apply MCMC algorithms to any dataset. We present here two improvements that are specifically relevant for the structure of the Quebec inventory, and which have been used for both the validation and predictions on these data. These are presented and justified briefly for their example values; applying our methodology to different databases could call for alternative reductions, or for no reduction at all if computational resources are not a limiting factor to perform the inference with Gibbs sampling.

First, because no plots were resampled faster than every 3 years in the inventory, we lowered the computational cost by keeping only one state every 3 years. In addition, to further lower the length of sequences, we extracted all sub-sequences of length 4 (hence corresponding to $4 * 3 = 12$ years) containing at least two measurements. These two manipulations shorten the sequences dramatically, from sequences originally spanning over 47 years and composed mostly of missing data, to sequences of length 4 and containing at least 2 measurements.

See the following box for a concrete example encompassing all the steps used to include yearly characteristics into temporal sequences.

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We provide as an example the data preparation with biomass modified from the Quebec inventory for one plot:

- 1975: biomass = 8.2 10^3 kg/ha
- 1981: biomass = 4.6
- 1988: biomass = 3.9
- 1992: biomass = 13.7
- 2000: biomass = 14.3

The preliminary step is to express these numerical values into discrete states. In our study, we subdivided biomass into 25 states from 0 to 50 10^3 kg/ha, corresponding to the following description of the same plot's biomass:

- 1975: biomass $\in 8 - 10$ 10^3 kg/ha
- 1981: biomass $\in 4 - 6$
- 1988: biomass $\in 2 - 4$
- 1992: biomass $\in 12 - 14$
- 2000: biomass $\in 14 - 16$

The first proposed step to simplify this temporal sequence is discretization into 3-year intervals:

70-72	73-75	76-78	79-81	82-84	85-87	88-90	91-93	94-96	97-99	00-02	03-05	06-08
?	8-10	?	4-6	?	?	2-4	12-14	?	?	14-16	?	?

The second step of discretization is to further split the sequence into sub-sequences starting with a known observation and containing at least one additional observation, as well as dismissing the absolute value of the year in order to retain only the relative dynamics:

	T	T+3 years	T+6 years	T+9 years
Sequence 1	8-10	?	4-6	?
Sequence 2	4-6	?	?	2-4
Sequence 3	2-4	12-14	?	?
Sequence 4	12-14	?	?	14-16

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix 1.3 Correlation Analysis

Correlation matrices using the Pearson correlation coefficient were computed for all years (Figure 3), as well as decade-by-decade (see the following tables):

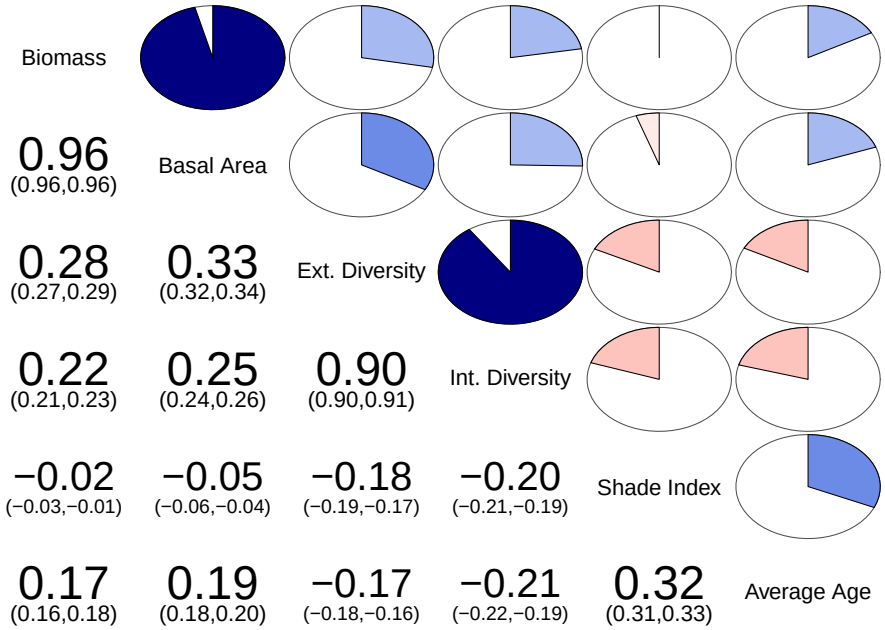


Figure 3: Graphical display of the correlation matrix for each characteristic; the numbers below the diagonal are the Pearson's coefficients (with 95% confidence intervals) and the circles above the diagonal provide visual indications of the correlations, with blue colors denoting positive correlations and pink colors denoting negative correlations (Friendly, 2002).

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Correlation for 2000s

age-supplied dataset (6292 data)	BIOMASS	BA	MEANSUCC	EXTDIV	INTDIV	MEANAGE
BIOMASS	1.00	0.96	−0.11	0.36	0.33	0.08
BA	0.96	1.00	−0.15	0.40	0.37	0.08
MEANSUCC	−0.11	−0.15	1.00	−0.24	−0.27	0.35
EXTDIV	0.36	0.40	−0.24	1.00	0.89	−0.15
INTDIV	0.33	0.37	−0.27	0.89	1.00	−0.16
MEANAGE	0.08	0.08	0.35	−0.15	−0.16	1.00

Correlation for 1990s

age-supplied dataset (9845 data)	BIOMASS	BA	MEANSUCC	EXTDIV	INTDIV	MEANAGE
BIOMASS	1.00	0.96	−0.02	0.31	0.28	0.10
BA	0.96	1.00	−0.05	0.37	0.33	0.10
MEANSUCC	−0.02	−0.05	1.00	−0.15	−0.18	0.35
EXTDIV	0.31	0.37	−0.15	1.00	0.90	−0.22
INTDIV	0.28	0.33	−0.18	0.90	1.00	−0.24
MEANAGE	0.10	0.10	0.35	−0.22	−0.24	1.00

Correlation for 1980s

age-supplied dataset (6097 data)	BIOMASS	BA	MEANSUCC	EXTDIV	INTDIV	MEANAGE
BIOMASS	1.00	0.96	0.06	0.18	0.14	0.11
BA	0.96	1.00	0.06	0.22	0.17	0.17
MEANSUCC	0.06	0.06	1.00	−0.16	−0.17	0.40
EXTDIV	0.18	0.22	−0.16	1.00	0.92	−0.20
INTDIV	0.14	0.17	−0.17	0.92	1.00	−0.21
MEANAGE	0.11	0.17	0.40	−0.20	−0.21	1.00

Correlation for 1970s

age-supplied dataset (9417 data)	BIOMASS	BA	MEANSUCC	EXTDIV	INTDIV	MEANAGE
BIOMASS	1.00	0.95	−0.02	0.22	0.18	0.26
BA	0.95	1.00	−0.03	0.26	0.19	0.29
MEANSUCC	−0.02	−0.03	1.00	−0.20	−0.23	0.26
EXTDIV	0.22	0.26	−0.20	1.00	0.92	−0.16
INTDIV	0.18	0.19	−0.23	0.92	1.00	−0.19
MEANAGE	0.26	0.29	0.26	−0.16	−0.19	1.00

Correlation for all years

age-supplied dataset (31651 data)	BIOMASS	BA	MEANSUCC	EXTDIV	INTDIV	MEANAGE
BIOMASS	1.00	0.96	−0.04	0.29	0.23	0.17
BA	0.96	1.00	−0.07	0.33	0.25	0.19
MEANSUCC	−0.04	−0.07	1.00	−0.18	−0.21	0.32
EXTDIV	0.29	0.33	−0.18	1.00	0.90	−0.17
INTDIV	0.23	0.25	−0.21	0.90	1.00	−0.21
MEANAGE	0.17	0.19	0.32	−0.17	−0.21	1.00

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix 1.4 Principal component analysis

To confirm the correlation analysis results, we applied a principal component analysis to the dataset, summarized here:

principal component analysis	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.589	1.3124	0.9958	0.7927	0.30958	0.19001
Proportion of Variance	0.421	0.2871	0.1653	0.1047	0.01597	0.00602
Cumulative Proportion	0.421	0.7080	0.8733	0.9780	0.99398	1.00000

The loadings are presented in the following table and can be visualized in the biplots (Figure 4).

Loadings	PC1	PC2	PC3	PC4	PC5	PC6
Bionass	0.479	0.43	−0.24	0.200	−0.142	0.683
Basal area	0.495	0.42	−0.23	0.131	0.091	−0.710
Shade tolerance index	−0.171	0.35	0.70	0.593	−0.011	−0.029
External diversity	0.509	−0.31	0.36	−0.082	0.700	0.133
Internal diversity	0.484	−0.36	0.37	−0.102	−0.694	−0.099
Average tree age	−0.042	0.54	0.36	−0.758	−0.012	0.035

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

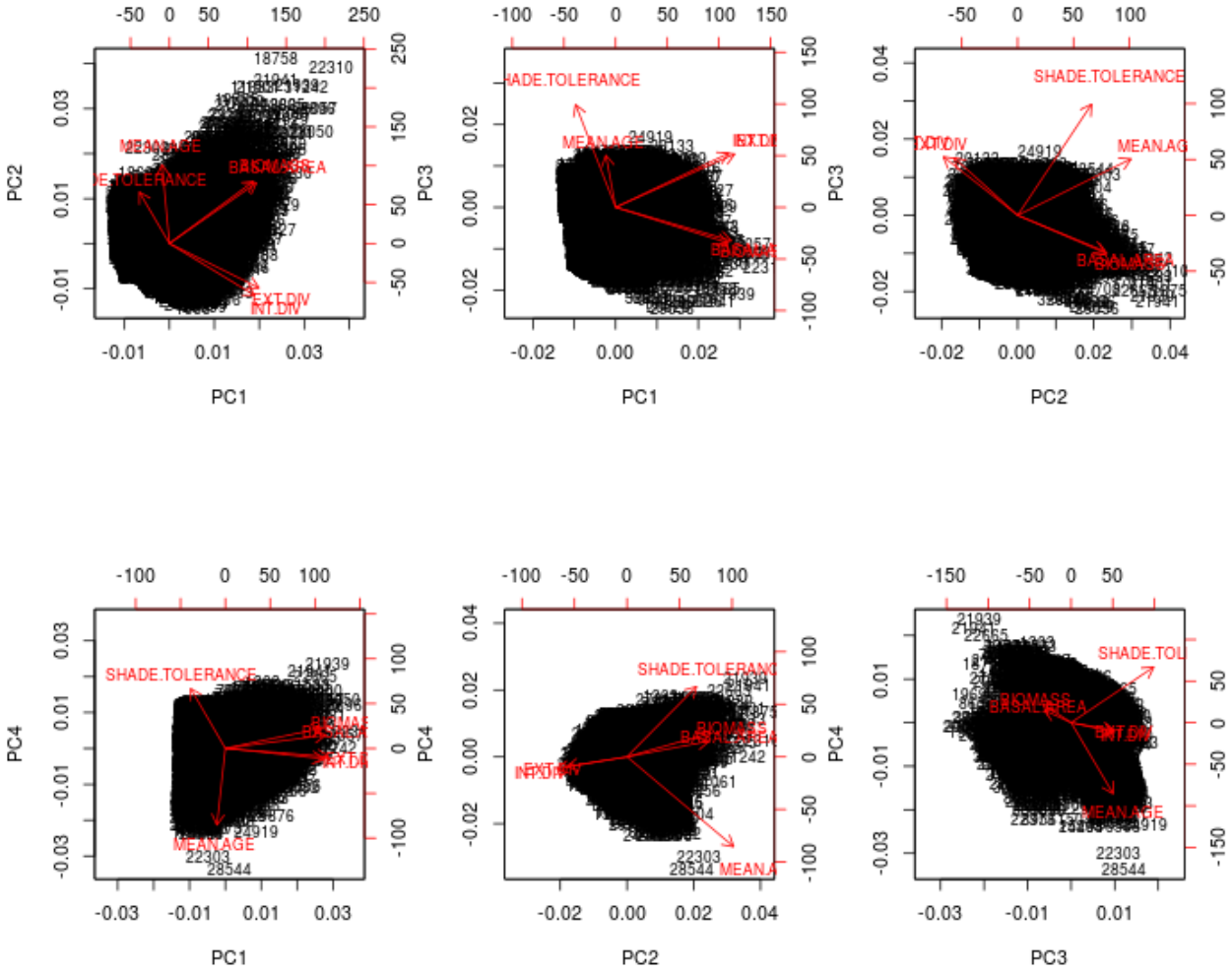


Figure 4: Biplots of all six studied characteristics in the spaces defined by the first four principal components (PC1 to PC4).

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix 1.5 Transition matrices of biodiversity, shade tolerance index and average age

at time = t + 3 years

at time = t (in %)	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1
0–0.1	93	0	2	2	1	1	0	0	0	0
0.1–0.2	4	65	3	16	1	4	3	3	1	1
0.2–0.3	4	1	79	9	4	2	1	0	0	0
0.3–0.4	4	1	11	67	5	7	3	2	0	0
0.4–0.5	3	0	5	8	65	11	5	2	1	0
0.5–0.6	2	0	1	5	5	78	4	3	1	0
0.6–0.7	1	0	1	2	4	7	72	9	3	0
0.7–0.8	1	0	0	1	1	5	11	75	6	0
0.8–0.9	0	0	0	0	1	2	6	13	76	2
0.9–1	0	0	0	1	1	1	1	6	19	70

rel. sem. 1 % – 0 %

Figure 5: 3-year transition matrix of the internal diversity.

at time = t + 3 years

at time = t (in %)	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1
0–0.1	92	3	2	1	1	1	0	0	0	0
0.1–0.2	8	76	9	3	1	1	0	1	0	0
0.2–0.3	3	11	73	4	4	3	1	1	0	0
0.3–0.4	1	2	6	72	8	6	1	2	1	1
0.4–0.5	1	2	10	6	64	13	3	1	0	0
0.5–0.6	1	1	2	4	4	72	6	4	4	1
0.6–0.7	0	0	1	2	2	9	71	7	6	2
0.7–0.8	0	0	0	1	2	4	6	69	13	4
0.8–0.9	1	0	1	0	1	4	3	4	78	8
0.9–1	0	0	0	0	0	1	0	1	3	95

at time = t + 3 years

at time = t (in %)	0–30	30–59	59–89	89–118	118–148	148–177	177–207	207–236	236–266	266–295
0–30	73	27	0	0	0	0	0	0	0	0
30–59	1	91	8	0	0	0	0	0	0	0
59–89	0	3	90	6	0	0	0	0	0	0
89–118	0	1	4	86	9	0	0	0	0	0
118–148	0	0	1	3	86	9	0	0	0	0
148–177	0	1	0	1	4	84	9	0	0	0
177–207	1	1	1	1	1	14	72	5	1	1
207–236	7	8	7	8	8	11	18	16	8	9
236–266	10	11	10	10	10	11	10	10	8	9
266–295	8	9	8	9	8	9	9	10	8	22

rel. sem. 1 % – 0 %

Figure 6: 3-year transition matrices of the shade tolerance index (left) and stand age (right).

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix 1.6 Validation

To estimate the general errors in the ability of a Markov-Chain model to predict future forest characteristics based on a dataset such as the Quebec dataset, we performed two different validations by splitting the dataset using two different partitions.

In the first validation, we ran the Gibbs sampler with only the first 18 years of records from 1970 to 1988. We then used the obtained models to predict what would be the forest state for a period corresponding to the second half of the dataset. In details, we pooled all data from the early years and computed from them an aggregated distribution of the first half of the dataset. We then predicted the aggregated distribution 18 years later by iterating 6 times in our 3-year transition models. We finally compared this predicted average with the aggregated distribution in the dataset of the years spanning from 1989 to 2007 (Figure 3 in main text). As mentionned in the article, the predictions were highly accurate, with R^2 between observation and prediction ranging from 0.8 to 0.95.

In the second validation, we randomly split the data in two different halves, regardless of the year. We then proceeded to the computation of the transition matrix for each half, and computed the corresponding equilibrium states (Figure 7).

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

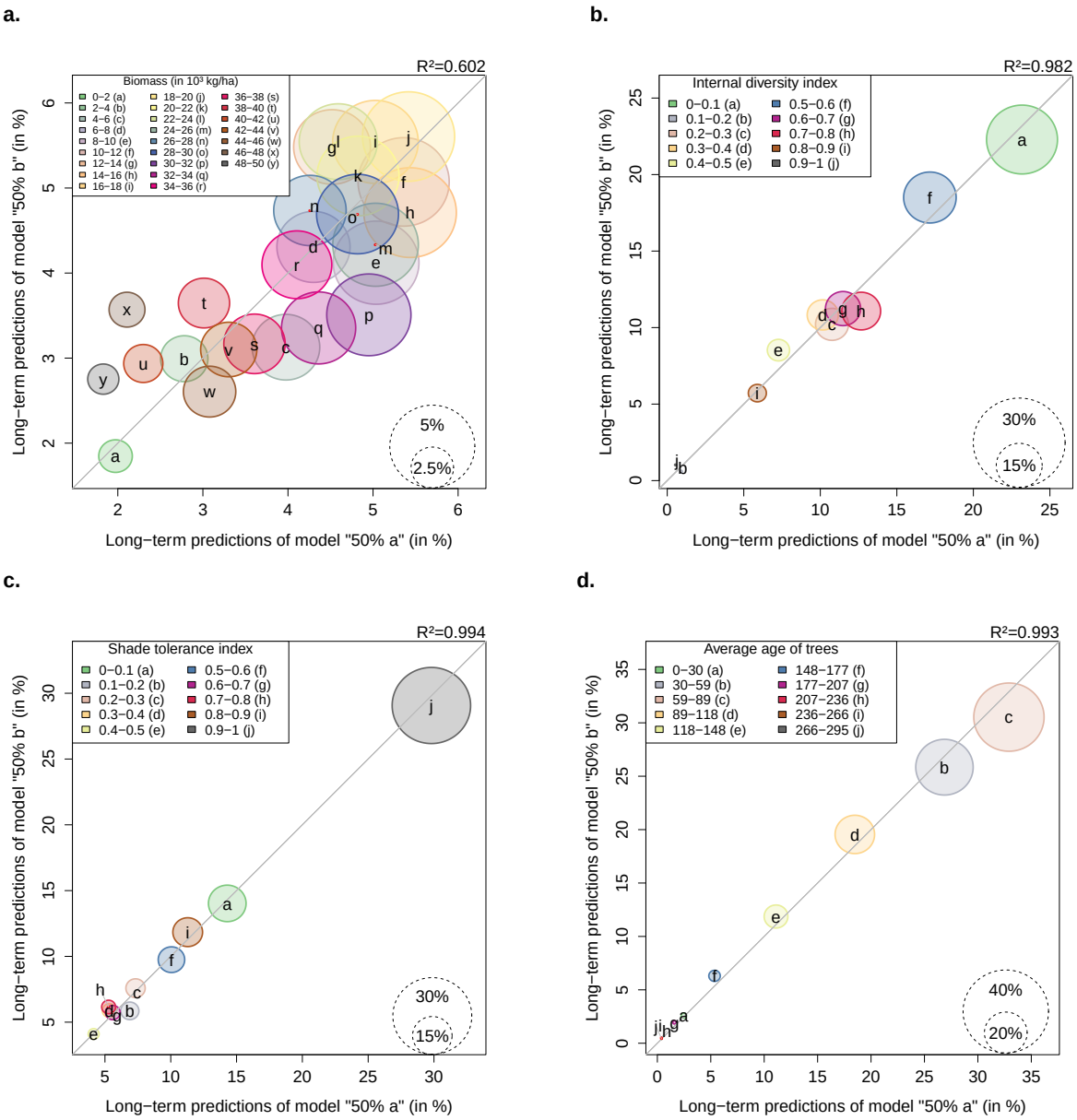


Figure 7: Predictions at equilibrium from two models, each computed with a different half of the dataset (denoted here "50% a" and "50% b") for the states of each characteristic. For each state, the circle size denotes the number of stands belonging to it in the real dataset. The R^2 error measure is indicated on the top right of each plot. In this validation, the two halves of data used to compute the transition matrix correspond to a random split of the dataset, regardless of the years (see text for details).

bioRxiv preprint doi: https://doi.org/10.1101/005009; this version posted May 10, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Duchesne, L. and Ouimet, R. (2009). Relationships between structure, composition, and dynamics of the pristine northern boreal forest and air temperature, precipitation, and soil texture in quebec (canada). *International Journal of Forestry Research*, 2009.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324.

Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*, volume 539. Wiley New York.

Pasanisi, A., Fu, S., and Bousquet, N. (2012). Estimating discrete markov models from various incomplete data schemes. *Computational Statistics & Data Analysis*, 56(9):2609–2625.

Perron, J., Morin, P., et al. (2011). Normes d´inventaire forestier: Placettes-échantillons permanentes.