Consultancy Report number 7, by Ophélie Ratel – November 2020

Description of the final version of the statistical model and first results analysis

Objective

Express the quantity of biomass recovered over time in primary exploited and secondary forest plots, by including co-variables (landscape, climate, topography, soil) and developing our model in a Bayesian framework, which is particularly adapted when there is little data, thanks to the addition of information that could be described as "non-pure data" or "priors". These priors can be based on previous studies or expert knowledge: they are a way to make sure that our predictions are within the range of acceptable values given our prior knowledge on similar processes, and are thus especially important when data is scarce. Moreover, the Bayesian approach allows a rigorous estimation of parameters correlation and uncertainty.

Hypothesis

- Landscape structure (configuration and composition) have an impact on the regeneration potential of forests (proximity to crops or to old-growth forest patches).
- The more complex the structure of the landscape, the more negatively it influences the rate of forest regeneration.
- The increase in the number of interfaces between forests and intensive agriculture reduces the regeneration potential of forests.
- The presence of large patches of forest near the plots studied will have a positive effect on the regeneration of the plots.

Material

After plots data homogenisation and analysis, 47 plots have been retained to put in the model, 4 from secondary forest and 43 from exploited primary forest. We also took into consideration mean biomass value of 49 primary forest plots located in La Selva, in order to use this amount of biomass as a reference value (Figure 1). Inventories go from 1987 to 2019, with 3 to 17 inventories per plot.

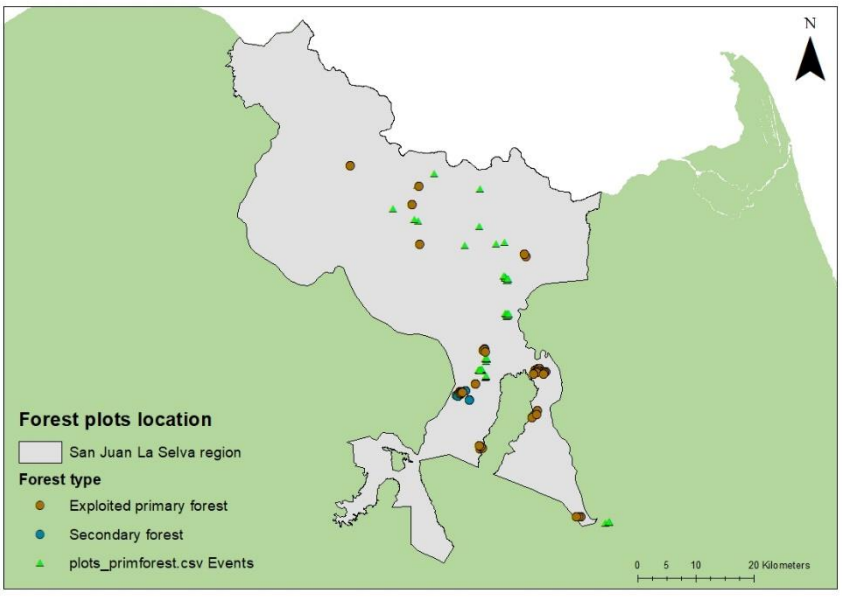

Figure 1: Location of the plots of secondary and exploited forest studied and the primary forests control plots.

In order to calculate the landscape metrics integrated into the model, the previously established land use map was reclassified. An intensive agriculture class includes both banana and pineapple crops, and a forest class includes both primary and secondary forests (Figure 2). This transformation supports the hypothesis that two agricultural land uses, although different, will have the same effect on the structure of the landscape. For the forest class, the two types of forest have been grouped together because in the model the dynamics of both secondary and primary exploited forest plots are studied.

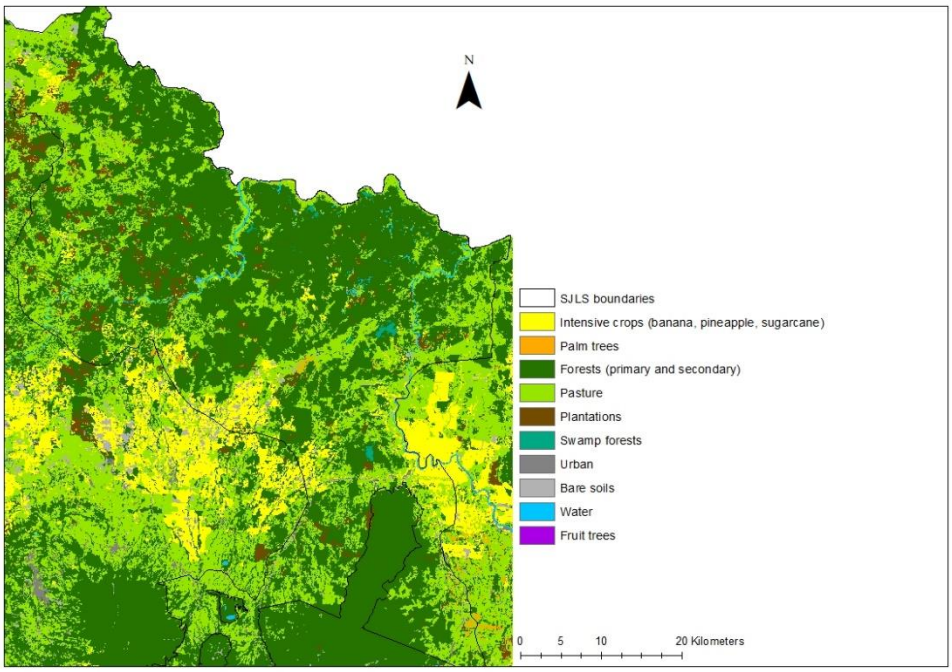

Figure 2: Land use classification used for landscape metrics calculation.

Method

Implementation of the model was done using R Studio and Stan package (Stan Development Team, 2020[1]). Calibration was carried out using Stan's programming language (Carpenter et al., 2017[2]), and was developed in R (RCore Team, 2019[3]). The model was built following a three-parameter exponential function:

$$predAGB_{p,c}max = AGBmax * \left(1 - e^{-\beta_p * (t_c + t0_p)^{\theta}}\right) \quad \textbf{(1)}$$

With p the plot, c the census; $predAGB_{p,c}max$ is the predicted biomass in plot p at census c; $t_c > 0$ the recovery time, i.e. the time since the disturbance (deforestation or logging); $t0_p$ is the initial recovery time; *AGBmax* is the maximum attainable biomass; β_p is the recovery rate to reach AGBmax and θ the shape parameter: when $\theta > 1$ the function is sigmoid.

[1] Stan Development Team, 2020. RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/.

[2] Carpenter et al 2017. Stan: a probabilistic programming language Journal of Statistical software RCore Team 2020.

[3] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

This type of equation was chosen based on the following criteria:

(i) The function only takes positive values (no negative biomass);
(ii) A saturation of total biomass recovery (the recovering forest cannot accumulate biomass forever);
(iii) A null biomass at t = 0 (in the case of secondary forests);
(iv) A flexible function that can take a sigmoid shape.

This last criterion was chosen because the analysis of the dynamics of the observed data showed that the plots seemed to follow a sigmoid growth pattern, with a rather slow start of biomass recovery, then an increase in the recovery rate and finally the arrival at a saturation threshold. This equation therefore seemed to us the most appropriate for predicting biomass recovery in the plots studied (Figure 3).

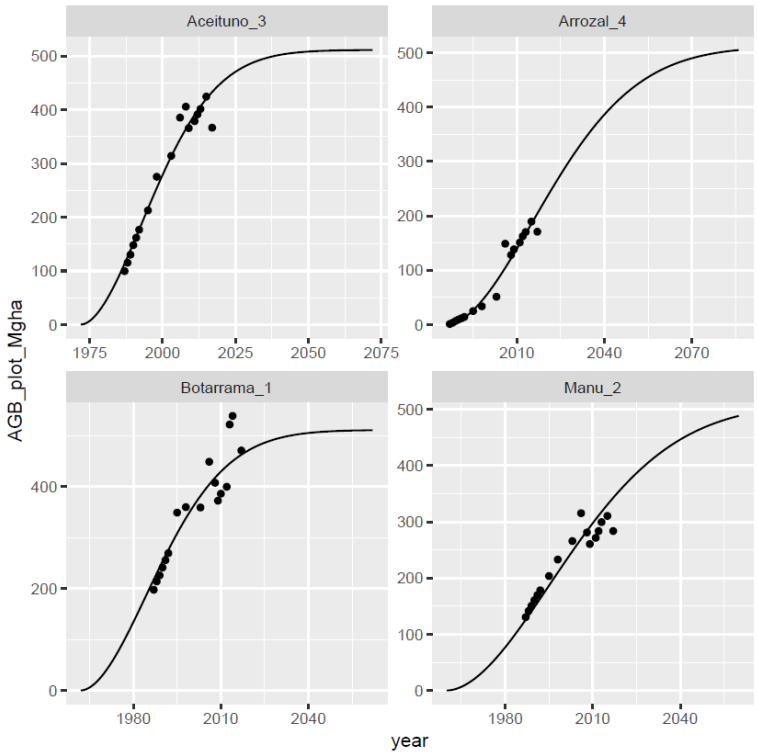

Figure 3: Prediction curve following an exponential function with three parameters. Biomass recovery in secondary forest plots according to time (black dots).

The several steps to build the model are:

1. **Describe every input data as vectors**

- Plots number [P] and inventories number [N];
- Exploited forest plots [L] and primary forest plots, used as reference plots [M];
- Year of inventory [year];
- Year of disturbance (deforestation or first exploitation) [dist];
- AGB value per plot [AGB];
- Mean AGB value in primary forest plots, used as reference for *AGBmax* [AGB_I];
- Covariables [covar].

2. Describe each parameter of the exponential function

- *AGBmax* is the maximum attainable biomass at the end of regeneration;
- β_p is the recovery rate of plot *p*. In order to facilitate the interpretation of the results, we have deduced from β the parameter *tm*, which corresponds to the time needed to recover 50% of *AGBmax*. β and *tm* are linked by the following equation:

$$1/2 * AGBmax = AGBmax * (1 - e^{-\beta * tm^{\theta}})$$

$$\Leftrightarrow 1/2 = 1 - e^{-\beta * tm^{\theta}}$$

$$\Leftrightarrow (\log 1/2) = -\beta * tm^{\theta}$$

$$\Leftrightarrow \log 2 = \beta * tm^{\theta}$$

$$\Leftrightarrow tm = ({}^{\log 2}/_{\beta})^{\frac{1}{\theta}} \quad \text{And } \beta = {}^{\log 2}/_{tm^{\theta}} \quad \textbf{(2)}$$

- We added a random plot effect on parameter *tm*, as well the fixed effect of covariables. All covariables where centred and scaled, and their effect is quantified by parameters lambda (see below);
- One parameter *lambda* was estimated for every covariable added (*e.g. lambda1* with *covar1*, etc.). Because all covariables were centred and scaled, the values of lambdas can be compared to estimate the relative effect of covariables on biomass recovery rates;
- θ is the parameter which give the sigmoid shape to the function, when $\theta > 1$;
- $t0_p$ is the initial recovery time, defined for each plot *p*: it is set to zero for secondary forests (that start with a null biomass) and takes positive values for logged forests.

3. Model estimation

The likelihood of observations was defined as follows:

$$obsAGB_{p,c} \sim N(predAGB_{p,c}, \sigma^2) \quad \textbf{(3)}$$

Where $obsAGB_{p,c}$ is the estimated AGB in plot *p* at census *c*; $predAGB_{p,c}$ is the predicted AGB as defined in equation **(1)**; σ is the standard deviation.

We then added priors on the following parameters: the priori on *AGBmax* is a normal distribution of 400 Mg/ha and a standard deviation of 200 Mg/ha. These values were chosen according to prior literature and expert knowledge (Letcher & Chazdon, 2009[4]). We also took into consideration AGB data available in primary forest plots. In the same way, the parameter tm follows a normal distribution of mean 40 years and standard deviation 20 years, compared to the values obtained in Rozendaal & Chazdon (Rozendaal & Chazdon, 2015[5]).

[4] Letcher S.G. & Chazdon R.L. 2009. Rapid recovery of biomass, species richness, and species composition in a forest chronosequence in Northeastern Costa Rica. Biotropica. 41(5), p. 608-617.

[5] Rozendaal D.M.A. & Chazdon R.L. 2015. Demographic drivers of tree biomass change during secondary succession in northeastern Costa Rica. Ecological Applications. 25(2), p. 506-516.

4. Selection of covariables

- Correlation analysis

The matrix of correlation with all covariables showed high levels of correlation between some of the variables, especially between landscape metrics (Figure 4). For example, the percentage of pixel couples forest-forest (pNC-3.3) is strongly positively correlated with the surface of the biggest forest patch (LPI.class_3), whereas it is strongly negatively correlated to the biggest agriculture patch surface (LPI.class_1), the percentage of pixel couples crops-forest (pNC-1.3) and the Shannon heterogeneity indicator (SHDI). Within the environmental covariables, elevation is highly correlated to mean precipitation level (ppm_plot) and mean temperature (temp_plot), which are negatively related between them. Percentage of organic carbon (perCO) and organic matter (perMO) are completely positively correlated. Cation exchange capacity (CEC) is positively correlated to both perCO and perMO, but also to mean precipitation and to forest-related landscape metrics. The percentage of sand (persand), on the other hand, is very negatively correlated with both the percentage of clay (perclay) and the percentage of slime (perslime). These 3 variables put together result in 100% soil composition.

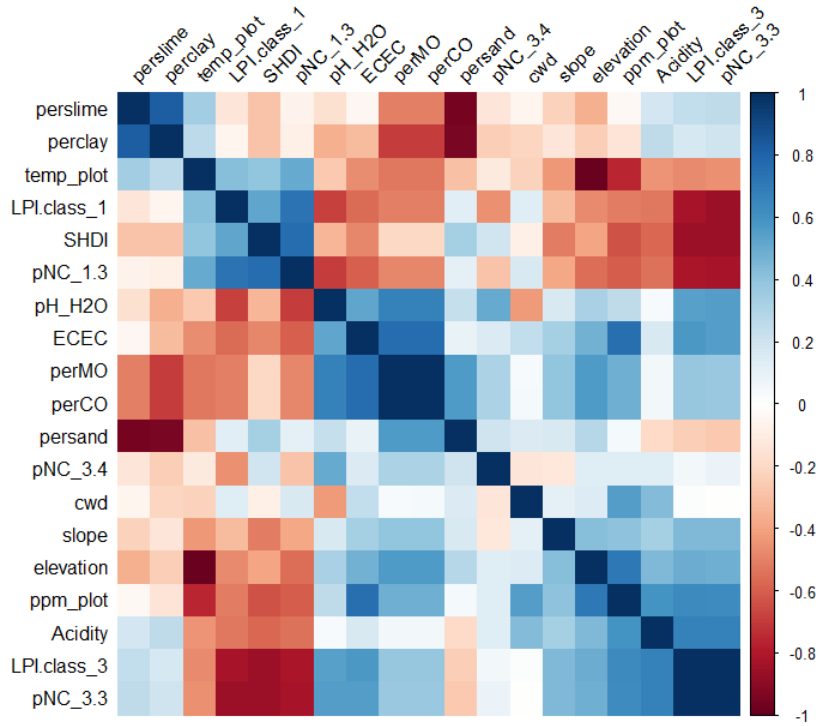

Figure 4: Correlation matrix between all environmental and landscape metrics.

- Covariables selected

Based on our hypothesis, the matrix correlation and after having tested each variable alone in the model, a list of nine covariables was established (Table 1). Landscape metrics (lines 1 to 3 in Table 1) were calculated using land use classification map with landscape ecology *Chloé* software (Boussard & Baudry, 2017[6]). The long-term Climatic Water Deficit (CWD) was

[6] Boussard, H. & Baudry, J. (2017) Chloe4.0: A software for landscape pattern analysis.

obtained from http://chave.ups-tlse.fr/pantropical_allometry.htm (Chave et al., 2014[7]). The slope was obtained thanks to raster map calculation using ArcGIS 10.3. Percentage of organic carbon and sand in soil and cation exchange capacity (CEC) data were obtained from CATIE database. Elevation values were obtained from SRTM raster.

Variable name	Description	Units/Values
pNC.1-3	Percentage of agriculture-forest interface pixels	[0;1]
pNC.3-3	Percentage of forest-forest interface pixels	[0,1]
pNC.3-4	Percentage of forest-pasture interface pixels	[0,1]
CWD	Long-term climatic water deficit	In mm/year
Slope	Slope	In degrees
perCO	% of organic carbon in soil first horizon (40 cm)	[0;100]
persand	% of sand in soil first horizon (40 cm)	[0;100]
ECEC	Effective Cation Exchange Capacity	Cmol/kg
elevation	Plot elevation	In meters (m)

Table 1: Names and description of the covariables included in the model.

[7] Chave J., et al. 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Global Change Biology. 20(10), p. 3177-3190.

Results

The maximum biomass (*AGBmax*) potential estimated by the model is 456.62 Mg/ha, with a standard deviation of 193.24 Mg/ha. The time to recover 50% of this potential value is on average 53.34 years (*mu_tm*) with a standard deviation of 12.33 years. The recovery time per plot is shown in Figure 5.

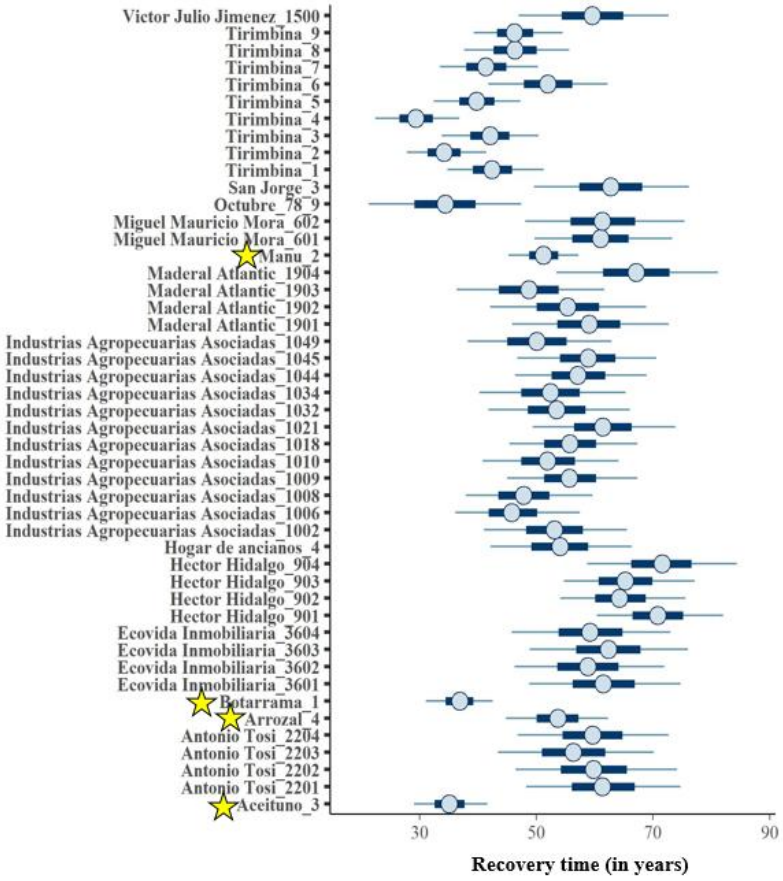

Figure 5: Recovery time (t_c) in year for each plot. Plots with a yellow star are the secondary forest plots.

A prediction curve was built for every forest plot, using the maximum likelihood of each parameter. Recovery rate trajectory is highly non-linear with accelerating recovery rates the first years, and decelerating rates or a stable rate after an inflexion point (Figure 6). Trajectories diverge between forest types (*e.g. Aceituno* and *Antonio Tosi* plots), but also between plots within the same forest type (e.g. *Hector Hidalgo* and *Tirimbina plots*). These differences can be due to the differences in sampling efforts between plots – some plots have between 10 and 17 inventories (*e.g. Aceituno, Arrozal,* or *Tirimbina*) whereas some plots have only 3 or 5 inventories (*e.g. Ecovida Inmobiliaria* or *Maderal Atlantic*). The date of disturbance may also explain these variations, as some forests have been logged or, conversely, abandoned for a longer or shorter period of time.

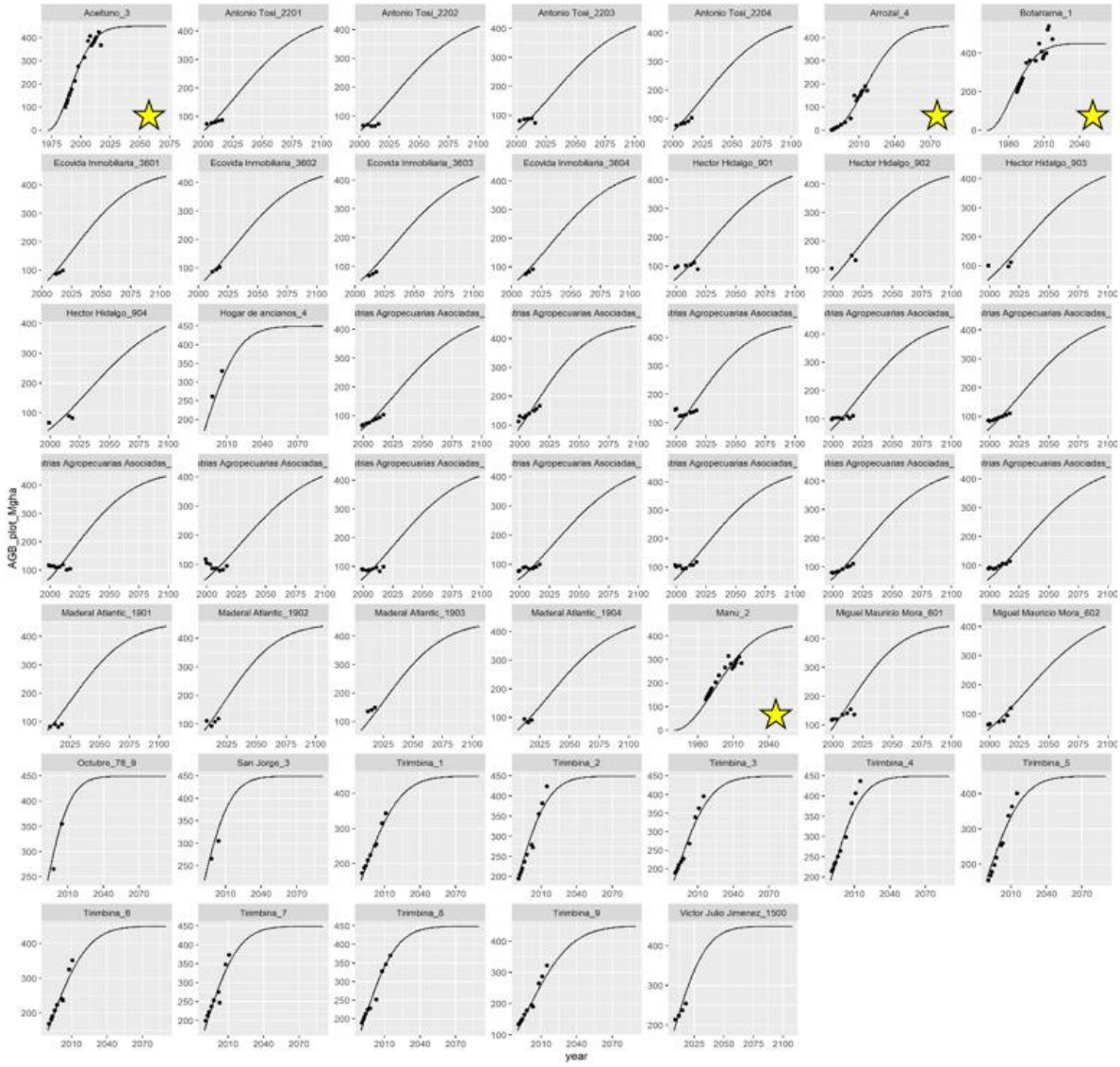

Figure 6: Prediction curves (black line) plotting biomass recovery according to time (year). Plots with a yellow star are for secondary forest plots. Black points represent observed data.

Estimating *lambda* parameter enables us to see if the covariates have an effect on the biomass recovery time, and if so, which ones. Figure 7 represents each of the nine variables integrated in the model according to its associated *lambda* value (Figure 7). All *lambda* values follow a normal low centred on 0 with a standard deviation of 10. In addition, all covariables were centred and scaled, so results could be compared. Thus, percentage of sand within soil first horizon (*persand*), the slope (*slope*) and climatic water deficit (*cwd*) do not seem to have a significant impact on the recovery time, as 95% of their lambda value are comprised between -5 and 5. Percentage of pixel couples forest-pasture (*pNC-3.4*) and forest-forest (*pNC-3.3*) do not impact recovery time, even if the second one show tendency to reduce this recovery time. Effective cation exchange capacity (*ECEC*) and elevation seem to have a relative effect on recovery time. Their associated *lambda* tend to high values, which suggest a negative effect on recovery time, *i.e.* a longer time period is necessary to recover 50% of AGBmax in plots with high ECEC and elevation values. The interface between forest and agriculture (*pNC.1-3*) also seems to increase recovery time. Values are very low, with none to 6% of crops-forest interface in some plots.

Covariables pNC-3.3 and ECEC are partially correlated (around 50%), which could explain the non-effect of forest proximity on biomass recovery. Indeed, in our plots ECEC values are quite low (between 4 and 7 cmol/kg) and a soil with low CEC is a good indication that a soil is sandy with little or no organic matter that cannot hold many cations.

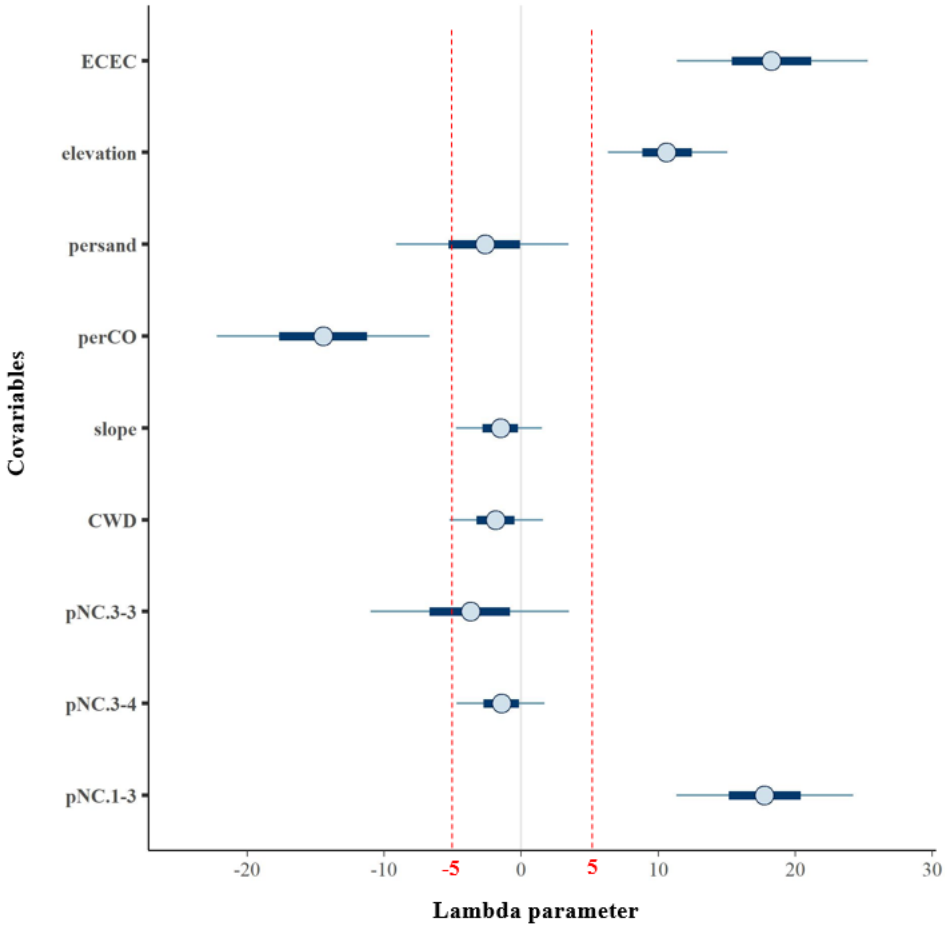

Figure 7: *Lambda* values for each covariable. A *lambda* value overlaid to zero means no significant effect of the associated variable on biomass recovery time. The light blue bar includes 95% of potential values that lambda can take. The large blue bar includes 80% of these values. The blue point represents the median value. See Table 1 for covariables description.

Perspectives

In the light of these first results, landscape structure and configuration seems to have a relative effect on biomass recovery time, but more investigations are needed to better understand and estimate the extent to which these variables impact biomass recovery, by testing the significance of these effects for example.

More information on plots history and successive treatments would be required to better encompass forest dynamics in these plots.

This Bayesian approach makes it possible to predict the potential for biomass recovery from relatively little data, and to see which parameters influence this dynamic.

For the continuation and end of this work, the selection of variables will be refined in order to integrate into the model the most relevant parameters to best explain the biomass recovery capacity. The interpretation of the results will also be continued in order to clearly answer the initial questions.