Pontifícia Universidade Católica
do Rio de Janeiro

Vinícius Oliveira da Costa

Multistep Forecast Amazon Deforestation using Regression and Recurrent Neural Network Approaches

Dissertação de Mestrado

Dissertation presented to the Programa de Pós-graduação em Matemática da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Matemática.

Advisor : Prof. Sinésio Pesco
Co-advisor: Profa. Angélica Nardo Caseri

Rio de Janeiro
September 2024

Pontifícia Universidade Católica
do Rio de Janeiro

Vinícius Oliveira da Costa

Multistep Forecast Amazon Deforestation using Regression and Recurrent Neural Network Approaches

Dissertation presented to the Programa de Pós-graduação em Matemática da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Matemática. Approved by the Examination Committee.

Prof. Sinésio Pesco
Advisor
Departamento de Matemática – PUC-Rio

Profa. Angélica Nardo Caseri
Co-advisor
Matriz São Paulo – BASF

Prof. Abelardo Borges Barreto Junior
Departamento de Matemática – PUC-Rio

Prof. Hélio Côrtes Vieira Lopes
Departamento de Informática – PUC-Rio

Dr. Leonardo Bacelar Lima Santos
Centro Nacional de Monitoramento e Alertas de Desastres Naturais – Cemaden

Rio de Janeiro, September 27th, 2024

Vinícius Oliveira da Costa

Bachelor in Mathematics by Pontifical Catholic University of Rio de Janeiro in 2022.

Bibliographic data

Costa, Vinícius Oliveira da

Multistep Forecast Amazon Deforestation using Regression and Recurrent Neural Network Approaches / Vinícius Oliveira da Costa; advisor: Sinésio Pesco; co-advisor: Angélica Nardo Caseri. – 2024.

138 f. : il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Matemática, 2024.

Inclui bibliografia

1. Matemática – Teses. 2. Desmatamento;. 3. Aprendizado de Máquina;. 4. Previsão de Vários Passos.. I. Pesco, Sinésio. II. Caseri, Angélica Nardo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Matemática. IV. Título.

CDD: 510

Acknowledgments

I would like to thank my parents, sister, and grandparents for all their support and love throughout my life.

To my friends who made even the most difficult moments easier to live.

To my advisor Sinésio Pesco for all the support and help in decision-making since my graduation.

To my co-supervisor Angélica Caseri for countless meetings, ideas, discussions, and patience over the years.

To all the teachers, good or bad, I've had over the years, you helped shape the person I am today.

To PUC-Rio for the welcoming environment and all the opportunities offered.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Costa, Vinícius Oliveira da; Pesco, Sinésio (Advisor); Caseri, Angélica Nardo (Co-Advisor). **Multistep Forecast Amazon Deforestation using Regression and Recurrent Neural Network Approaches**. Rio de Janeiro, 2024. 138p. Dissertação de Mestrado – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

The Amazon rainforest, the largest tropical biome in the world, plays an essential role in both society and global environmental balance. Through its vast biodiversity and carbon storage capacity, it also supports local cultures and provides resources for sustainable development. Deforestation prediction occupies a significant role mainly in monitoring, control, and conservation planning. The ability to predict where and when deforestation will occur allows authorities and organizations to take more effective preventive measures, allocate resources more strategically and develop policies that can mitigate negative impacts.Therefore, the study of methods to predict deforestation has been increasingly developed in recent years. This work aims to apply supervised machine learning methods and statistical methods, such as autoregression, LightGBM, and Long Short Term Memory (LSTM) neural network to predict multi-step deforestation in the Brazilian Legal Amazon, using past observations of deforestation and climatic variables from the region. The research carried out showed that the most efficient results were presented in models that used autoregression. Furthermore, the study showed good results for classifying and pre-pointing anomalies in the series, characterized by their high deforestation values, as well as the general patterns of the series. The states of Pará and Mato Grosso and the municipality of Apiacás presented better results related to the classification of peak points, showing an average F1-Score for the predicted steps of 83%, 90%, and 85%, respectively. By enhancing strategies for monitoring and controlling deforestation, this study has the potential to positively impact public policies, promoting a balance between economic development and environmental preservation and climate regulation.

Keywords

Deforestation; Machine Learning; Multistep Prediction.

Resumo

Costa, Vinícius Oliveira da; Pesco, Sinésio; Caseri, Angélica Nardo. **Previsão Multi-Etapas do Desmatamento na Amazônia utilizando Abordagens de Regressão e Redes Neurais Recorrentes**. Rio de Janeiro, 2024. 138p. Dissertação de Mestrado – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

A floresta Amazônica, o maior bioma tropical do mundo, desempenha um papel essencial tanto na sociedade quanto no equilíbrio ambiental global. Através de sua vasta biodiversidade e capacidade de armazenamento de carbono, ela também apoia culturas locais e fornece recursos para o desenvolvimento sustentável. A previsão de desmatamento ocupa uma função significativa principalmente no monitoramento, controle e planejamento da conservação. A capacidade de prever onde e quando o desmatamento ocorrerá permite que autoridades e organizações tomem medidas preventivas mais eficazes, alocando recursos de maneira mais estratégica e desenvolvendo políticas que possam mitigar impactos negativos. Portanto, o estudo de métodos para prever o desmatamento tem sido cada vez mais desenvolvido nos últimos anos. Este trabalho visa aplicar métodos de aprendizado de máquina supervisionado e métodos estatísticos, como autorregressão, LightGBM e rede neural Long Short Term Memory (LSTM) para prever o desmatamento de múltiplos passos na Amazônia Legal brasileira, utilizando observações passadas de desmatamento e variáveis climáticas da região. A partir das pesquisas realizadas resultados mais eficientes foram apresentados nos modelos que utilizaram autorregressão. Além disso, o estudo mostrou bons resultados para classificar e prever pontos de anomalias da série, caracterizados por seus altos valores de desmatamento, assim como os padrões gerais da série. Os estados do Pará e Mato Grosso e o município de Apiacás apresentaram melhores resultados relacionados a classificação de pontos de pico, mostrando F1-Score médio para os passos previstos de 83%, 90% e 85%, respectivamente. Ao aprimorar as estratégias de monitoramento e controle do desmatamento, este estudo tem o potencial de impactar positivamente políticas públicas, promovendo um equilíbrio entre desenvolvimento econômico, preservação ambiental e regulação climática.

Palavras-chave

Desmatamento; Aprendizado de Máquina; Previsão de Vários Passos.

Table of contents

List of figures

List of tables

List of Abbreviations

AI – Artificial Intelligence

AM – Amazonas

AR – Autoregressive Model

ARX – Autoregressive Model with Exogenous Input

CNN – Convolutional Neural Network

EFB – Exclusive Feature Bundling

GOSS – Gradient One Side Sampling

GP – Gaussian Processes

GRU – Gated Recurrent Unit

LSTM – Long Short Term Memory

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

ML – Machine Learning

MLP – Multi Layer Perceptron

MT – Mato Grosso

NLP – Natural Language Processing

PA – Pará

PCA – Principal Component Analysis

RMSE – Root Mean Squared Error

RNN – Recurrent Neural Network

RO – Rondônia

SVD – Singular Value Decomposition

SVM – Support Vector Machine

X – Exogenous Input Regressive Model

Yes, it was my way.

Frank Sinatra, *My way.*

1
Introduction

The exploration of the Brazilian Legal Amazon was marked by cycles of economic development that, over the years, impacted the region and its biodiversity. Rubber, cocoa, soy, mineral extraction, and livestock farming were among the activities responsible for the widespread devastation of large areas in the region. Likewise, the government's incentive to populate the areas during the 1970s to integrate the region into Brazilian territory caused countless damages. Through the construction of roads, dams, and cities, millions of hectares of land were devastated, putting the region's fauna and flora at risk, as well as the culture and livelihood of hundreds of indigenous peoples.

Currently, the main causes of deforestation in the region are linked to the expansion of agribusiness (1), fires (2), and illegal exploitation of natural resources (3), such as wood and ores. To mitigate these impacts, the government launched projects focused on continuous forest monitoring, designed not only to detect deforestation early but also to enhance efforts to combat these destructive practices. In general, monitoring is done using remote sensing equipment, such as satellites, however, ground patrols and aerial surveillance are used to check deforested regions as well.

Over the years, numerous researchers have sought to model deforestation data to aid in its prevention and control. With the advancement of machine learning and deep learning techniques, several models related to these areas have emerged. Dominguez et. al (4) used a hybrid neural network (5) to predict deforestation in the Brazilian Legal Amazon. For deforestation data, an LSTM (Long Short Term Memory) (6) network was applied. In contrast, for static data, geographic and administrative variables, a dense neural network was applied, also called MLP (Multi-Layer Perceptron) (7). The use of neural networks is very common due to their high ability to discover patterns in series. LSTM networks are particularly effective at capturing short and long-term patterns, enabling in-depth and accurate data analysis. The study demonstrated that the proposed model achieved a coefficient of determination of 87%, that is, 87% of the variance in the dependent variables was explained by the model.

The deforestation problem is generally presented in two ways: with a regressive analysis of the data or to classify possible deforested areas in a given period. One of the ways to explore the second problem is through CNN (Convolutional Neural Networks) (8). Its use is more applied when the data are

images due to its ability to capture spatial and hierarchical characteristics of the data. Because of this, the network can extract important information from the series and can be used for segmentation, classification, and detection tasks, for example. Fodor et. al (9) used this type of network to predict deforestation in the Amazon and detect fires on a global scale. Thus, based on the use of different bands of satellite images, the model achieved positive results for fire detection, especially in South America, with 0.95 AUC, indicating that the model can distinguish the created classes very well.

Research on this topic does not stop in Brazil, Saha et. al (10) used machine learning algorithms to identify possible deforestation zones in Jaldapara National Park in India. Through the tests, it was concluded that Support Vector Machine (SVM) (11) achieved the best results, obtaining an accuracy of 90.7% when classifying deforested areas. The SVM algorithm works by finding the optimal hyperplane that best separates the available classes. It can be used in regression and classification problems, and, in this specific case, it was used to classify deforested areas.

A method that is also used to identify deforested regions is Bayesian neural networks (BN) (12). It combines traditional neural networks with Bayesian inference (13) allowing the modeling of uncertainty in predictions. Silva et. al (14) used BN to predict deforestation in northeastern Pará. Additionally, it used variables such as distance to hotspots and distance to degraded areas to assist in training the model. The results showed 80% accuracy in predicting events. Mayfield et. al (15) in addition to BN used Gaussian Processes (GP) (16) to predict deforestation values in Madagascar and Mexico. The research shows that while BN produced more stable results between different sampling methods, GP performed better considering fewer input variables. It is worth mentioning that GP, like BN, is a probabilistic model and is used to predict continuous distributions, providing an average estimate of uncertainty for each predicted point.

Regressive models (17) are also widely used to predict deforestation. These models have great relevance in the forecasting task, as, in addition to being less complex and having fewer parameters to train, compared to neural networks, they are effective in detecting temporal patterns in historical series. Başaran et. al (18) used linear regression applied to deforestation, climate, and geospatial variables to predict deforestation in the Mediterranean region of Turkey. The results presented indicated that the method applied could be used as a way to assist in decision-making to create regulatory policies in the region.

The problem of predicting deforestation remains open, mainly due to

the numerous challenges involved. One of the biggest difficulties is related to predicting outliers in the series. Bearing in mind that times of high deforestation rates are the most critical and must be combated with greater intensity, the responsible authorities need to carry out extensive control planning to mitigate its magnitude. Therefore, combined to completely map the behavior of the series the identification and prediction of peak deforestation values in the Brazilian Legal Amazon will be studied during this research.

Other difficulties permeate this problem, such as the quality of data that may be incomplete or inaccurate, in addition to the broad correlation of deforestation with other variables, such as climate, geopolitics, geospatial, and socioeconomic. Given the existence of these relationships, the complexity of predictions increases, as it can be difficult to predict the occurrence of geopolitical factors in a given region. During this study, only variables related to climate and deforestation will be considered. Another difficulty is presented by the size of the Legal Amazon, because of this, several regions behave differently, implying the creation of specific models to deal with their particularities.

Therefore, to meet the stated objectives, data from the DETER-B project will be used to predict deforestation values in selected states and municipalities in the Brazilian Legal Amazon. From this, three different methodologies will be proposed for applying the determined algorithms, autoregression (17), LightGBM (19), and LSTM (6). The models will be validated and compared based on the chosen regression and classification metrics. The classification metrics will be used to categorize moments of high and low deforestation rates and verify whether the models can capture unusual behaviors in the series. Furthermore, the research aims to anticipate the deforestation trend by creating new features and using the concept of sliding windows (13) to identify the ideal window size for predicting the values of the series.

It is important to highlight that classification metrics will be employed to enhance model reliability and improve anomaly detection in the series. Additionally, meteorological variables will be incorporated to explore the relationship between these factors and deforestation rates in each studied region. The data organization will also be scrutinized, considering both monthly models, where data is grouped by month, and weekly aggregated models, where data is organized weekly but predictions cover four-week intervals. These approaches aim to contribute to deforestation prediction studies and foster the development of more efficient models.

This study was divided into six chapters. Chapter 2 deals with the analysis of the theoretical part used during the research, such as algorithms,

metrics, hyperparameter optimizations, and related general concepts. Chapter 3 shows a discussion of the research study object, the Brazilian Legal Amazon, as well as the collection, treatment, and study of the data used. Chapter 4 introduces the proposed methodology, covering the forms of data organization applied in the models, the proposed methods for applying the algorithms, and the hyperparameter optimization made for each model. Chapter 5 analyzes the best models chosen for each selected region, as well as a general analysis of the results found. Finally, Chapter 6 proposes discussions of the results found, in addition to proposing topics to improve future research related to the topic.

2
Theoretical Concepts

This chapter will present the theoretical concepts covered during the research, including the metrics used, the models and algorithms adopted, and the areas of knowledge employed.

2.1
Artificial Intelligence and Machine Learning

The term "Artificial Intelligence" (AI) was coined in 1955 by John McCarthy during the event "Darthmouth Summer Research Project on Artificial Intelligence" (20). Even though there was great interest in leveraging the area, the evolution of AI was very limited in its early years due to the technological restrictions of the time. The reasons for this are quite straightforward, computers were still unable to process or store enough information, in addition to the operating costs being very high.

As the necessary technology began to become more powerful and cheaper, substantial advances in AI became more present. It can be mentioned the development of areas such as Machine Learning (ML) (21, 13), the creation of the first neural networks (22, 23) and the advancement in pattern recognition studies (13) and computer vision (24).

Although Machine Learning and Artificial Intelligence are closely related, they are not synonymous, as ML is a subset of AI. Other applications of AI include Natural Language Processing (NLP) (25, 26) and Visual Perception (27). The Machine Learning area aims to develop systems capable of learning and adapting without explicit instructions through algorithms and statistical models to infer patterns from data. It can be cited as examples of Machine Learning algorithms autoregressive models applied to time series (28), decision trees (13), and LightGBM (19).

Over the years, machine learning algorithms have become more powerful and accessible. Eventually, a new era of AI began, characterized by advances in deep neural networks and deep learning (29). The latter is a subfield of ML that uses artificial neural networks to model complex problems, as it is very effective with large numbers of data and where manual feature extraction is demanding. Examples of deep learning algorithms include recurrent neural networks (RNN) (29), such as LSTM (29, 30, 6) and GRU (29, 31), and convolutional neural networks (CNN) (29, 8). Such technologies have revolutionized big data analysis (32) and the automation of complex tasks.

Continued progress in computing power and increasing availability of large data sets continue to drive the rapid and efficient evolution of artificial intelligence. In this way, both ML and AI represent significant transformative potential capable of influencing the way we work, live, and interact with the digital and physical world.

2.2 Supervised Models

Supervised models are linked to the automatic learning of computational rules involving input and output relationships. They are generally used in regression and classification problems, in which the respective outputs for each model input is known, that is, the data set is labeled. Therefore, the main idea of supervised models is to find a relationship between inputs and outputs so that it is possible to minimize the error in this relationship.

When a supervised model is applied, the data set is divided into training, validation, and testing. Each set has a specific function so that the result obtained by the model is not biased. For this purpose, it is not interesting for models to share information, each one must contain unique parts of the total data set. Therefore, the training set will be used to train the model weights, the validation set to compare different models and hyperparameters, and the test set to prove that the constructed model works.

There are several important supervised learning algorithms such as linear regressions, decision trees, neural networks, and support vector machine (SVM). Its applications range from stock market asset prediction (33), to voice identification and music generation (34). Even though supervised models already have the advantage of knowing the intended outputs, that is, it is known what response it is needed to receive from the models, it also has disadvantages. The main ones revolve around the outputs, as they may be poorly labeled or defined, resulting in poor algorithm functioning.

2.3 Unsupervised Models

Unsupervised models deal with the automatic learning of computational rules that are described only with input data. In general, these rules are learned to simplify a database, making it easier to analyze and interpret the data or to apply a supervised model.

The main problems that are addressed with unsupervised learning are dimensionality reduction and clusterization. In general, the use of these algorithms allows the size of the input environment to be reduced or a smaller

number of representatives to be determined that can adequately represent the total set, simplifying the input set.

Principal component analysis (PCA) (35) and singular value decomposition (SVD) (36) are examples of algorithms that make it possible to reduce the dimension of the input set. K-means and hierarchical clustering are examples of algorithms that organize input data into groups. Applications of unsupervised models include anomaly detection, natural language processing, and recommendation systems.

However, it is important to mention that although supervised learning can find patterns in unlabeled data, it still has disadvantages. The main ones are that their results are subjective, as they are based only on human interpretation as there are no accuracy metrics, in addition to being subject to overfitting, as they are based on hypothetical patterns.

2.4 Time Series

Time series are defined as a sequence of points indexed in order of time that were collected in a certain period. For a fair analysis of time series, the indexed points must have consistent time intervals throughout the period. Time series analysis allows the study of a variable over time, the time variable provides an additional source of information to analyze dependencies between data.

Typically, for the analysis of time series data to be consistent and reliable, a large amount of data is required. Larger data sets allow for several representative samples and the analysis can ignore noisy data. Furthermore, it ensures that trends and patterns discovered are not atypical patterns (outliers) and could be responsible for seasonal variations.

A factor often examined in time series is stationarity. It is said that a time series is stationary when it presents constant mean, variance, and covariance in addition to not exhibiting seasonal patterns that are repeated at fixed time intervals. This last property can be relaxed in certain contexts if the seasonal patterns are well-defined and modelable. Stationary time series have more simplified statistical calculations. The use of models that use methods as moving averages, such as ARIMA (autoregressive integrated moving averages), are developed specifically for stationary series and assume that these properties are valid.

Another important fact to be examined in the study of time series is the autocorrelation of the series, that is, how relevant the past observations of the series are for predicting the step of interest. To discuss this, the concept of lag

needs to be addressed. The lag k of an observation in a time series is defined as the previous k-th step of that observation in the series. Therefore, to calculate the autocorrelation of the series, consider the correlation of the original series with the series of its lags. It is important to say that not all observations will have all of the considered lags, therefore only observations that have available lags should be considered in the calculation. Figure 2.1 presents the series of the first two lags of a time series. When calculating the autocorrelation with the first lag, the first observation must be excluded, since it does not have the first lag, the same goes for autocorrelation with the second lag, in which the first two observations must be excluded.

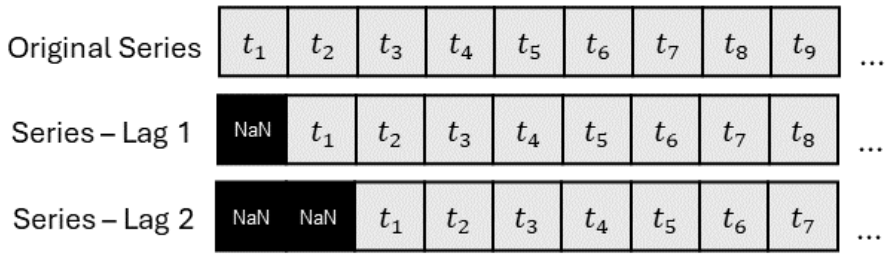


Figure 2.1: Series of the first and second lag of a time series.

In order to apply algorithms for time series forecasting, the sliding window concept (13) can be applied. It is based on defining the number of inputs and outputs of the models and then creating different subsets of continuous data within the series that satisfy the number of inputs and outputs. Figure 2.2 illustrates how the intervals for model creation are constructed.

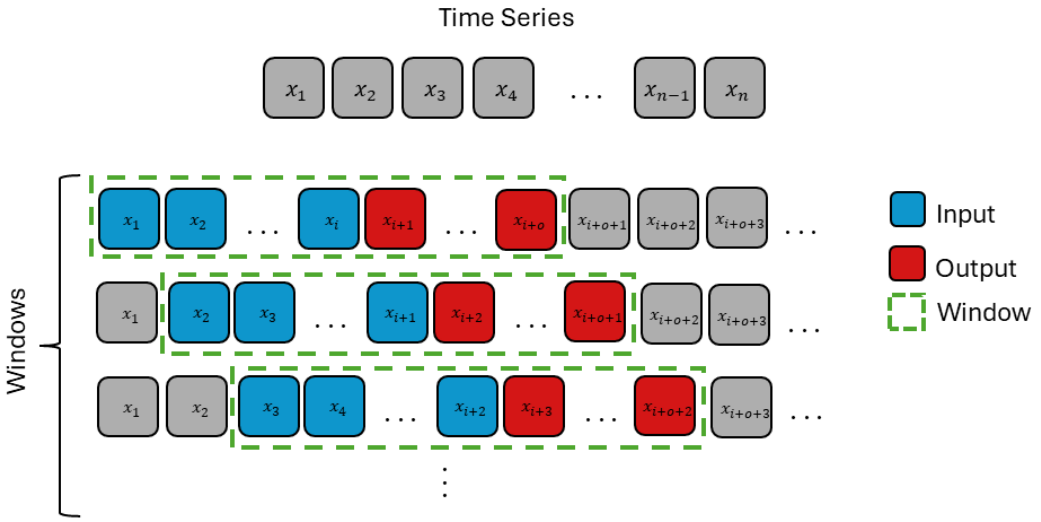


Figure 2.2: Sliding windows concept applied to time series considering inputs with i observations and outputs with o observations.

In addition to what has already been pointed out, although it is not necessary for some algorithms, data normalization is common for better model convergence, better comparison between different variables, and numerical stability since variables tend to be very large or small numbers. In this way, several types of data normalization can be used to obtain better results. Examples are Min-Max normalization, in which the data is normalized based on the minimum and maximum values of the series so that they are all between 0 and 1, and standardization in which the data is centralized so that they have an average 0 and standard deviation 1. In this research, Min-Max normalization will be applied to the series, as shown in Equation 2-1.

$$x_{Min-Max} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2-1}$$

2.5 Algorithms and Models

Now, the algorithms and models that were used to make predictions of deforestation values will be presented.

2.5.1 Autoregression

Autoregressive (AR) models (17) are statistical models widely used in econometrics, signal processing, among other areas. By using this model it is obtained a representation of time based on past observations of the time series considered. This relationship is depicted by the Equation 2-2, in which the time t of the time series is expressed through a linear combination of the last k time steps ($t > k$), where y_{t-i} represents the value of the time series $t - i$ and α_i the weight of this component ($1 \leq i \leq k$). In addition to past observations, there is a constant component, c, and a specific random error for each step that will be predicted e_t.

$$y_t = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_k y_{t-k} + e_t \tag{2-2}$$

In addition to autoregressive models, it can be introduced the concept of exogenous input models, which consist of representing values of a time series considering only values external to the series in question. Equation 2-3 shows the representation of this model, where x_{t-i} represents the value of the exogenous series at time $t - i$, β_i represents the weight that this component has in the regression ($1 \leq i \leq k$), y_t the expected value of the output at time t, c a constant component and e_t a specific random error for each step.

$$y_t = c + \beta_1 x_{t-1} + \beta_2 x_{t-2} + ... + \beta_k x_{t-k} + e_t \tag{2-3}$$

To this extent, it can be introduced the presence of a third type of model, called the autoregressive model with exogenous inputs (ARX). In addition to having entries from past observations of the series in question, we also consider external variables. The representation of this model can be seen in Equation 2-4, in which the sum of an autoregressive model and an exogenous input model is represented. Thus, y_{t-i} represents the value of the time series $t-i$ and α_i the weight of this component ($1 \leq i \leq k$), x_{t-j} represents the value of the exogenous series at time $t-j$, β_j represents the weight that this component has in the regression ($1 \leq j \leq l$), y_t the expected value of the output at time t, c a constant component and e_t a random error specific to each step.

$$\begin{aligned} y_t = c+&\alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_k y_{t-k} + \\ &\beta_1 x_{t-1} + \beta_2 x_{t-2} + ... + \beta_k x_{t-l} + e_t \end{aligned} \tag{2-4}$$

In order to determine the coefficients of the models created, Ordinary Least Squares (OLS) (37) will be used. This is a commonly used method for estimating coefficients within a linear regression. The objective is to find the set of coefficients that minimize the sum of the squared differences of the real (y_t) and predicted ($\hat{y_t}$) values as can be seen in Equation 2-5, where n is the number of observations and $\gamma_0, \gamma_1, \cdots, \gamma_m$ are the coefficients to optimize.

$$OLS = \min_{\gamma_0,\gamma_1,...,\gamma_m} \left(\sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{n} \right) \tag{2-5}$$

Performing the optimization shown in Equation 2-5 implies calculating $m+1$ partial derivatives, each regarding a coefficient γ_i with $i = 0, \cdots, m$, resulting in a linear system of $m+1$ variables and $m+1$ equations. The results will be estimators for each one of the coefficients, called $\hat{\gamma_i}$.

The estimators found within a regression require a certain quality and reliability. Therefore, some characteristics of the estimators are evaluated to infer their condition. The first is bias, which is determined by the difference between the expected value of the estimator and the true value of the parameter being estimated. An estimator, $\hat{\theta}$ is unbiased when the expected value of the estimator is equal to the true value of the parameter, θ, that is, $E\left[\hat{\theta}\right] = \theta$. Another property that can be mentioned is the consistency of the estimator, if the estimator is consistent then it converges in probability to the real value of the parameter as the number of samples increases. In addition to these characteristics, the efficiency and robustness of the estimators can be mentioned. An efficient estimator has the lowest variance among all unbiased

estimators, while a robust estimator refers to the ability of an estimator to withstand violations of model assumptions or the presence of outliers. Ideally, model estimators should be unbiased, consistent, robust, and efficient.

2.5.2 LightGBM

LightGBM (Light Gradient Boosting Machine) (19) is a recent ensemble framework created in 2017 that uses algorithms based on decision trees. It was designed to optimize training and improve model accuracy through techniques such as Gradient One Side Sampling (GOSS) (38) and Exclusive Feature Bundling (EFB) (19). Such techniques build a solid model by sequentially summing weaker models based on the gradient of the features considered. It is worth mentioning that decision trees are ML models used in classification and regression that aim to divide the input space into specific regions, using a tree structure to make decisions or predict values.

One of the main differences with LightGBM is that the tree being assembled will grow by leaves and not by levels, as is done with other algorithms. It is worth noting that this can generate overfitting on small bases or increase the complexity of the model if the same branch grows several times. A comparison of this characteristic can be seen in Figure 2.3. Therefore, one way to approach these problems is to limit the number of leaf divisions and the number of levels the tree can have.

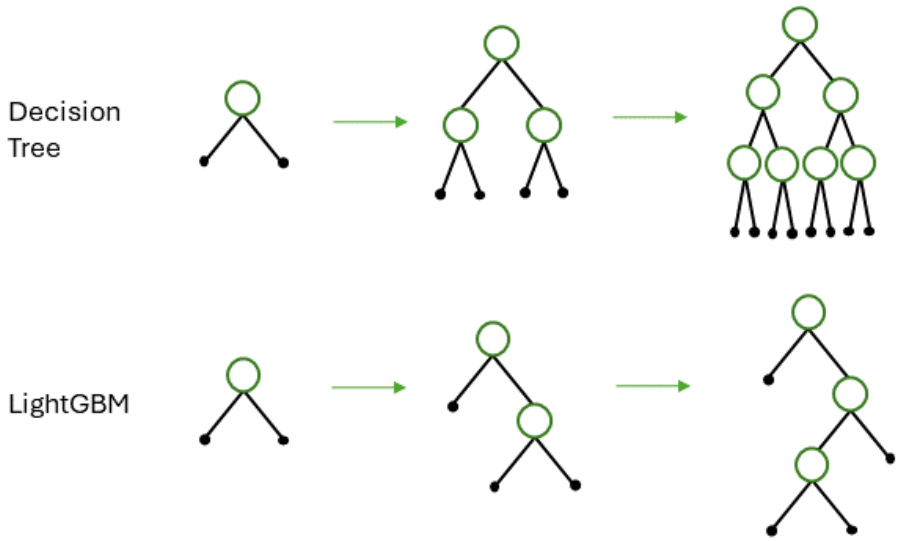


Figure 2.3: Comparison of LightGBM Architecture and Decision Tree in different iterations.

From this, it can be further dissected each part of this method, starting

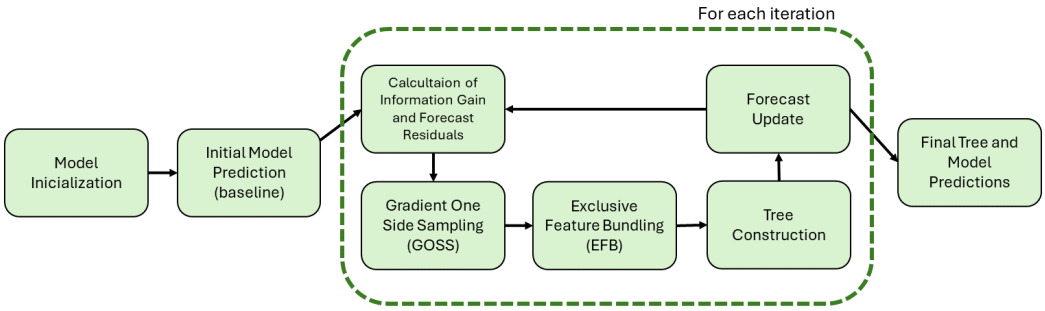


Figure 2.4: LightGBM operating scheme.

with GOSS. The main idea of this technique is to eliminate variables that do not have a high enough information gain (entropy). Instances with greater gradients have more information gains which helps more with model learning. Therefore, when determining an information gain threshold it is possible to separate variables that may be more useful for the model, that is, have entropy greater than or equal to the threshold, from those that are not useful for the model. From the group of variables with entropy lower than the threshold, random variables are removed from the model, while the remaining ones are retained. This practice allows the information gain estimator to remain accurate. This process is repeated for each iteration of the algorithm.

Next, EFB can be mentioned, which aims to combine features without causing much loss of information. Therefore, the approach will leverage the fact that if there are sparse features within the model, they can be combined, resulting in minimal loss of information. This way, the complexity decreases and the speed of the model increases.

In general, Figure 2.4 presents a summarized scheme of the algorithm. It is observed that initialization takes place in which a prediction of values occurs intending to compare them with the next tree that will be assembled. Then, a loop is entered, where the residuals of the series values will be calculated so that GOSS, EFB, and the assembly of the tree create a result that best approximates the predictions of the real data. This loop is repeated according to the number of repetitions determined at model initialization. Finally, the best result is chosen and final predictions are calculated.

2.5.3 LSTM

Long Short-Term Memory, or LSTM, neural networks (29, 30, 6) is a special type of recurrent neural network (RNN) designed to deal with the problem of sequence learning and solve some of the limitations of traditional RNNs, such as the challenge of learning long-term dependencies in temporal

sequences. Figure 2.5 shows how an LSTM network works and how its cells interact with each other. As can be seen, a cell always passes two values to the next cell, which are: the value of C_{t-1}, called the state of the cell, and h_{t-1}, called the hidden state of the cell. Furthermore, the cell at time t receives its input value, called x_t. Based on these three inputs, the outputs can be calculated: C_t and h_t. To perform cell computations, two functions will be used: sigmoid (σ) and hyperbolic tangent ($tanh$), represented respectively by Equations 2-6 and 2-7.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2-6}$$

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2-7}$$

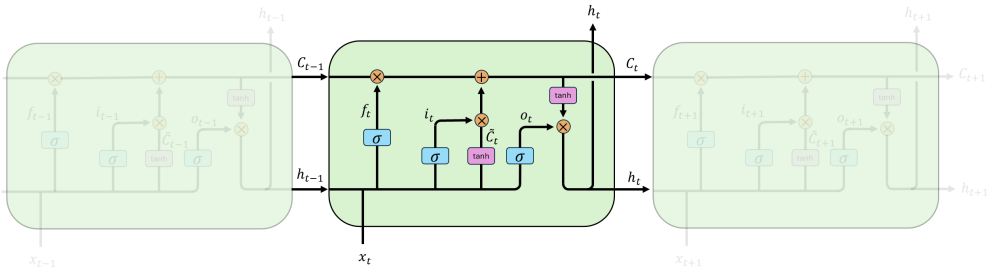


Figure 2.5: Sequential functioning of a LSTM network.

To understand the internal functioning of each cell, it is important to divide its processes into 3 parts, which will be treated individually. The first of them, shown in Figure 2.6, is called the forget gate. It will be decided which information will be left aside when calculating the cell state (C_t). This calculation is made by applying the sigmoid function (σ) to the vector obtained by concatenating h_{t-1} and x_t. Equation 2-8 shows the application of these calculations, considering the vector $[h_{t-1}, x_t]$ obtained by the concatenation of h_{t-1} and x_t, the bias vector b_f and the forget gate weight matrix W_f.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2-8}$$

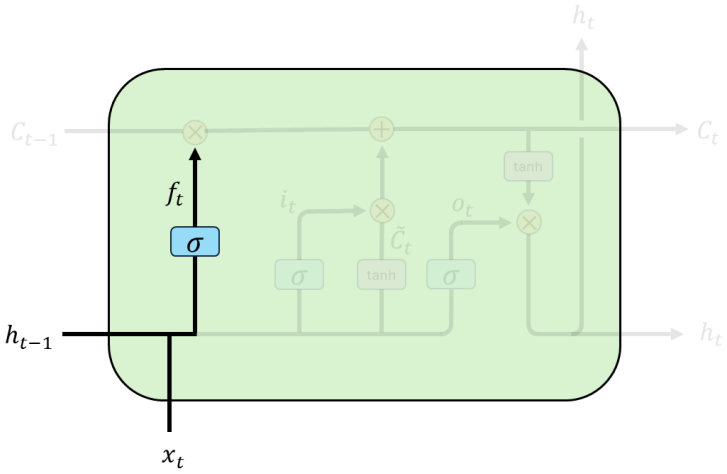


Figure 2.6: Forget gate in a LSTM cell.

Next, the values for the input gate or update gate are calculated, as depicted in Figure 2.7. The calculations performed at this stage determine which new information will be retained in the cell state and are divided into two parts. The first part, represented by Equation 2-9, determines the values to be updated by applying a sigmoid function. The second part, represented by Equation 2-10, generates a vector of new candidates that can be added to the cell state.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2-9}$$

$$\tilde{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{2-10}$$

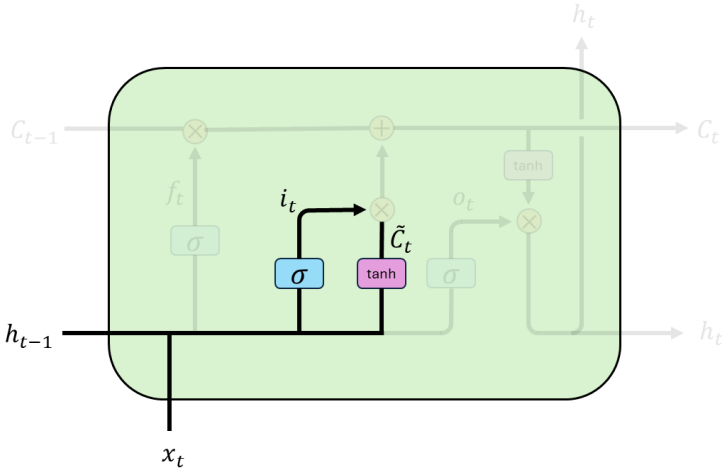


Figure 2.7: Input gate in a LSTM cell.

Based on the values calculated in the last two steps, it is possible to update the cell state value, C_t, simply perform the calculation shown

in Equation 2-11. Through this equation, the desired information from the previous cell, along with new and updated information from the current cell, will be obtained.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{2-11}$$

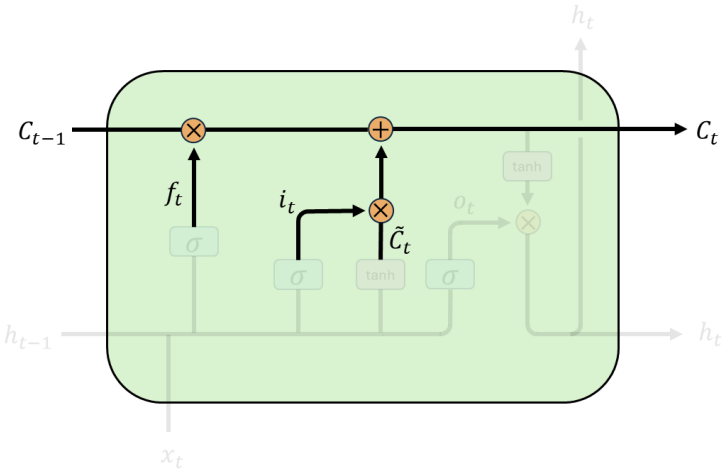


Figure 2.8: Cell state (C_t) in a LSTM cell.

Finally, the calculations of the cell's output values must be done, for this, the last compartment of the cell is entered, called the output gate, represented in Figure 2.9. First, the sigmoid will be applied to x_t and h_{t-1}, filtering which parts of the data will be passed to the output (h_t), represented by Equation 2.9. Then, the cell state (C_t) is applied to a *tanh* function and the values are multiplied point by point, as seen in Equation 2-13. This way, the value found can be passed to the next cell on a recurring basis.

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{2-12}$$

$$h_t = o_t \cdot tanh(C_t) \tag{2-13}$$

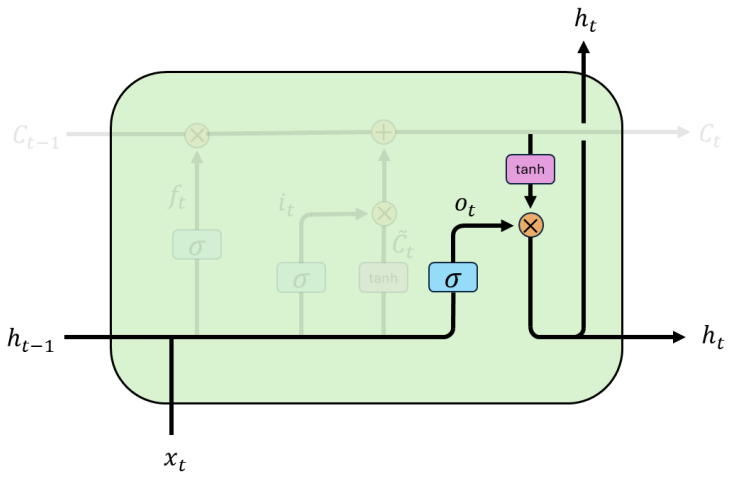


Figure 2.9: Output gate in a LSTM cell.

2.5.4 Principal Component Analysis

Principal Component Analysis, or PCA, (35) is a method used to reduce dimensionality in models with large amounts of data or features. The main objective of the method is to maintain the components that preserve the variability of the original model data. The theoretical application of PCA will be presented below considering the database with n observations and m variables, represented in 2-14.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{2-14}$$

For the algorithm to work, the data must be normalized, that is, for each variable in the model, the mean of the variable must be subtracted from each one of the observations and divided by the standard deviation, as presented in 2-15, it was considered $\bar{x_i}$ as the mean and σ_i as the standard deviation of the variable i. From this, the covariance matrix can be calculated, as can be seen in Equation 2-16.

$$R = \begin{bmatrix} \frac{x_{11}-\bar{x_1}}{\sigma_1} & \frac{x_{12}-\bar{x_2}}{\sigma_2} & \cdots & \frac{x_{1m}-\bar{x_m}}{\sigma_m} \\ \frac{x_{21}-\bar{x_1}}{\sigma_1} & \frac{x_{22}-\bar{x_2}}{\sigma_2} & \cdots & \frac{x_{2m}-\bar{x_m}}{\sigma_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x_1}}{\sigma_1} & \frac{x_{n2}-\bar{x_2}}{\sigma_2} & \cdots & \frac{x_{nm}-\bar{x_m}}{\sigma_m} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{2-15}$$

$$C = \frac{1}{n-1} X^T X \tag{2-16}$$

Then, the calculation of the eigenvectors and eigenvalues of the covariance matrix is done based on the singular value decomposition (SVD) (36) shown in 2-17. The equation shows three matrices, U, an orthogonal matrix of dimension $n \times n$, Σ, the diagonal matrix $n \times m$ with the diagonal entries being singular values, and V^T, where V is an orthogonal matrix $m \times m$ and V^T its transposed. The use of SVD helps in manipulating and understanding the matrix, facilitating tasks such as dimensionality reduction.

$$X = U\Sigma V^T \tag{2-17}$$

Using the described decomposition, it is possible to make the following calculations:

$$\begin{aligned} X^T X &= \left(U\Sigma V^T\right)^T \left(U\Sigma V^T\right) \\ \left(X^T X\right) V &= V\Sigma^T \Sigma V^T V \\ \left(X^T X\right) V &= V\Sigma^T \Sigma \rightarrow CV = \frac{1}{n-1} V\Sigma^2 \end{aligned}$$

Through these calculations, the problem of eigenvalues and eigenvectors is solved, as V is the matrix with the eigenvectors of $X^T X$ and $\Sigma^T \Sigma$ is the matrix whose diagonal entries are the eigenvalues. Therefore, when the eigenvectors are ordered in decreasing order of eigenvalues, the component space can be calculated. To decide the components to be used, a new matrix is calculated, presented in Equation 2-18, in which each column of the Z matrix refers to a main component, that is, the calculation of the i-th main component will be made from the multiplication of R with the i-th column of V.

$$Z = R\ V \tag{2-18}$$

2.6 Validation Metrics

Validation metrics are essential in evaluating machine learning and statistical models. They provide a way to quantify how models perform on data that was not used during training, helping to ensure that models generalize appropriately to new data. Choosing appropriate validation metrics depends on the type of problem being solved and the specific characteristics of the data.

2.6.1

RMSE

RMSE (Root Mean Square Error) (39) is the metric that indicates the standard deviation of residual values, or forecast errors ($y_i - \hat{y}_i$). RMSE assumes values greater than or equal to 0, values closer to 0 imply predictions with smaller errors. Errors are a measure of the distance between the regression line and the data points. Thus, RMSE represents a measure of dispersion of residual values, that is, indicating how concentrated the data are around the line of best fit. The Equation 2-19 indicates the RMSE formula, in which the real values (y_i), the predicted values ($\hat{y}_i$), and the number of predictions made (n) are considered.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2-19}$$

2.6.2 MAE

MAE (Mean Absolute Error) (39) is the validation metric used in regression problems that measures the average of the absolute errors between predictions and actual observed values. The Equation 2-20 consists of calculating the average of the n observations of the absolute values of the difference between the predictions ($\hat{y}_i$) and the actual values (y_i). MAE assumes non-negative values, and results closer to 0 indicate more accurate models. The MAE variation is less sensitive to outliers, unlike RMSE, so it penalizes all errors equally. The closer the RMSE and MAE values are, the more uniformly distributed the error values are. If the RMSE assumes much higher values, it indicates the presence of outliers in the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2-20}$$

2.6.3 MAPE

MAPE (Mean Absolute Percentage Error) (40) is a relative measurement that measures the MAE (Mean Absolute Error). The use of relative values allows the comparison of accuracy between time series methods, the lower the MAPE value, the more accurate the model. MAPE can assume values greater than or equal to 0. The result obtained indicates how much the values vary in percentage terms, for example, a MAPE equal to 0.5 indicates that the forecast values are within a range of 50% of the actual value. The Equation 2-21 indicates the calculation that is performed to obtain the MAPE, it is

considered the real values (y_i), the values obtained by the forecast ($\hat{y}_i$) and the total number of forecasts (n).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2-21}$$

2.6.4 Confusion Matrix

Confusion matrices (13) are used in classification problems so that it is possible to check correct and incorrect prediction values. In binary classification problems, confusion matrices have dimension 2×2, while in general, with n classes, confusion matrices have dimension $n \times n$. Generally, the actual labels are arranged on the vertical axis, while the prediction labels are arranged on the horizontal axis, as can be seen in Figure 2.10 for the binary classification example. The ideal when creating a confusion matrix is to only have numbers on the main diagonal of the matrix, as they represent values that are correctly classified. However, if there are many misclassification occurrences, this can identify problems within the model and help correct and improve it.

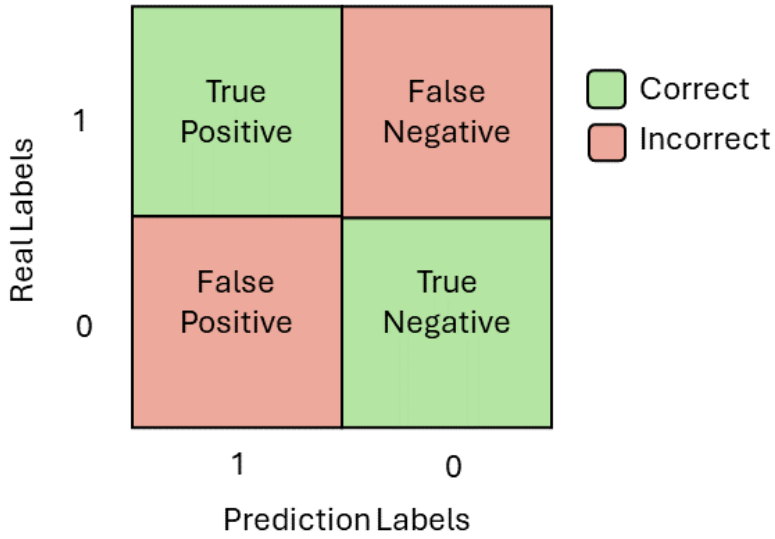


Figure 2.10: Confusion matrix for a binary classification.

2.6.5 Accuracy

To accurately represent the classified values within the confusion matrix, accuracy (13) can be defined. As shown in Equation 2-22, accuracy is calculated as the ratio of the sum of true positives (TP) and true negatives (TN) to the total number of predictions made by the model, which includes TP, TN, false

positives (FP), and false negatives (FN). Accuracy values range from 0 to 1, with 1 indicating that all predictions were correct and 0 indicating that all predictions were incorrect.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2-22}$$

2.6.6 Recall

The recall (13) of a model is measured based on the values calculated in the confusion matrix, it can also be called sensitivity or proportion of true positives and is used in classification problems. It intends to calculate the ratio between the number of TP and the sum of TP and FN, as can be seen in Equation 2-23. Their values can vary between 0 and 1, with results closer to 1 representing better classification. Examining Figure 2.10 it is presented that its calculation consists of dividing the number of TP by the sum of the values in the first line. In other words, recall answers the question: "Of the real number of values in this class, how many did I get right in the prediction?".

$$Recall = \frac{TP}{TP + FN} \tag{2-23}$$

2.6.7 Precision

Precision (13) represents another metric for classification problems. Its calculation is based on values from the confusion matrix, which consists of the ratio of the number of TP to the sum of TP and FP, as shown in Equation 2-24. Its values are in the range of 0 to 1, and the closer to 1, the more precise the model will be. From Figure 2.10 it is possible to see that this metric is calculated by the ratio of the number of VP to the sum of the values in the first column. Precision answers the following question: "Of the total number of elements predicted to be in this class, how many are actually part of this class?".

$$Precision = \frac{TP}{TP + FP} \tag{2-24}$$

2.6.8 F1-Score

The calculation of the F1-Score (13) can used for classification problems, especially when there is evidence of class imbalance. Its formula can be seen in Equation 2-25. This metric can take values between 0 and 1, with 1 being the perfect balance. Widely used when one class prevails much more than another,

implying that a model, even though it has high accuracy, may have a low F1-Score since all this model does is predict the same class that represents the majority of the data. Furthermore, the F1-Score can act when there is a need to balance precision and recall, implying that both the number of FN and FP have significant importance for the model.

$$F1\text{-}Score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{2-25}$$

2.7 Hyperparameter Tuning

In order to optimize model results, it is important to carry out a search so that the best hyperparameters are chosen. These are the attributes that control model training, that is, they are not learned by the model, but chosen before training. Therefore, two ways to search for the best set of hyperparameters will be presented: grid search and random search.

It is important to say that the cross-validation (13) technique is used in both processes. It is based on dividing the data set into k equal parts and training the model k times using one of the parts as a test set each time and the other $k - 1$ parts as a training set. For each time the model is trained, a performance metric is calculated and at the end of all training, the obtained metrics are combined to obtain a single estimate of the model's performance.

2.7.1 Grid Search

Grid search (41) is a way to perform hyperparameter optimization. It is based on a systematic exploration of a predefined set of possible hyperparameters, testing each combination of different hyperparameters. Cross-validation is generally applied so that the results found are reliable. For each combination used, a metric is calculated in order to compare the models, the best performance of the metric is attributed to the best model.

As can be seen, the grid search necessarily finds the best hyperparameter combination, as it tests all possible combinations, as seen in Figure 2.11. However, as the number of hyperparameters increases, the computational cost to apply this form of optimization becomes very high.

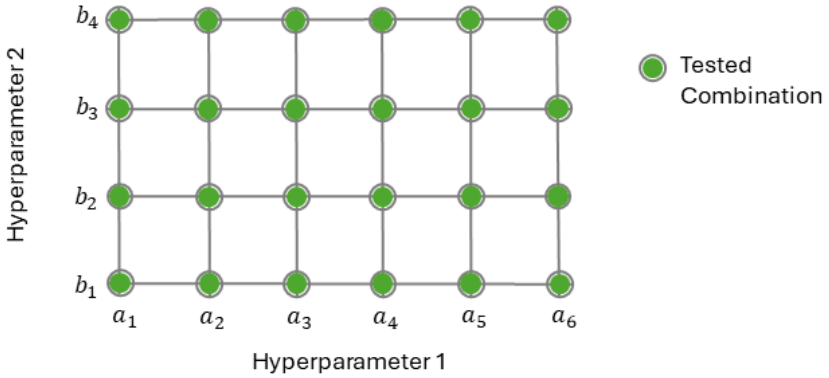


Figure 2.11: Performance of a grid search considering two hyperparameters.

2.7.2 Random Search

Unlike grid search, random search (41) does not explore all possible combinations of hyperparameters. A selection of combinations is made at random and only samples from this selection are tested. The use of random search has advantages when the hyperparameter space is very vast, however, it will not necessarily discover the best combination for that space, only the best combination concerning the selected cases.

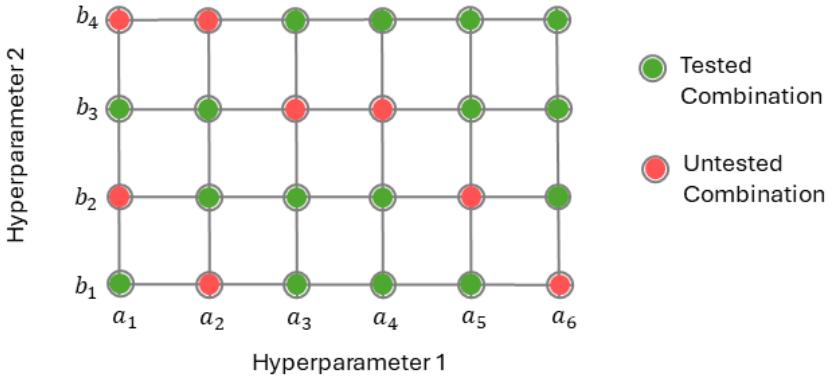


Figure 2.12: Performance of a random search testing 16 combinations considering two hyperparameters.

3
Data and Study Area

This chapter will expose the characteristics of the object of study covered in this research, the Legal Amazon, as well as the collection, processing, and analysis of data used during the research.

3.1
Study Area

The Brazilian Amazon forest occupies 3,300,000 km^2 of land, around 39% of the Brazilian territory. However, in 1953 the government established an area called the Legal Amazon with the objective and need to planning the inclusive and sustainable economic development of the region. Therefore, the Legal Amazon occupies 5,217,423 km^2, around 61% of Brazilian territory. Its extension entirely covers the states of Acre, Amapá, Amazonas, Mato Grosso, Pará, Rondônia, Roraima, and Tocantins and partially the state of Maranhão, encompassing 772 Brazilian municipalities.

Because it has such a large extension, the Legal Amazon occupies three Brazilian biomes: the Amazon, the Cerrado, and the Pantanal Matogrossense. Thus, it has a vast biodiversity with approximately 40 thousand species of plants, 400 species of mammals, 1300 species of birds, 3000 species of fish, and millions of insects.

In addition to the fauna and flora, the region has around 28 million inhabitants, around 13% of the Brazilian population, presenting a low demographic density of 5.6 inhabitants per km^2. It is worth mentioning that the Legal Amazon region covers 51% of the entire national Indigenous population, around 870 thousand people. According to data from 2021, 45% of its territory is made up of protected areas that are distributed in Conservation Units (19%), Indigenous Lands (23%), Environmental Protection Areas (3%), Lands Quilombolas (0.2%), among others.

The terrain is not predominantly flat; only 13.14% of the total area is flat, while 73.78% consists of irregular features, such as hills and ridges. 16% of the irregular terrain has rugged terrain, becoming vulnerable to erosion and consequently becoming more vulnerable to deforestation. Without vegetation cover, erosion is intensified, resulting in floods.

In the Legal Amazon region, two types of climate predominate: equatorial and tropical. The territory has high humidity throughout the year. However, in the part of the year with greater humidity, from November to March, a

substantial presence of rain and a lower average temperature is noticed, while in the period of lower humidity, from May to September, little or no precipitation in the region accompanied by higher temperatures is observed. It is common for an average of 2000 mm to rain in a year in the Legal Amazon, although there are regions where rainfall exceeds 3500 mm.

Despite attempts at more sustainable economic development, the engine of the Legal Amazon economy is based on the commerce of agricultural and mineral commodities. The rampant exploitation of these resources generates deforestation and carbon emissions into the environment. In addition to causing damage to the region, these practices do not generate significant economic growth and accentuate social inequality. Informal jobs increase, professional qualifications decrease and salaries are below the national average.

Just like legal economic practices within the region, illegal practices also take their place. Fires, logging, and illegal mining play a considerable role in the environmental degradation of the Legal Amazon along with the trafficking of animal and plant species at risk of extinction. In 2023 alone, deforestation reached an area of 18,000 km^2, although this value shows a decrease according to the previous year's values, it still represents a considerable area.

Monitoring of the region is carried out via satellite by the PRODES, DETER, and DETER-B projects. Encouraging and financing projects like these is essential for the responsible agents to be able to act with greater precision in combating such criminal activities. The development of techniques for predicting deforestation is already used, however, there is still great room for improvement.

Preserving forests is vital for several fundamental reasons. In addition to conserving their unique biodiversity, forests are essential for promoting a sustainable economy and protecting cultures and rights, especially of Indigenous communities, who, in part, directly depend on them. At the same time, the ability of trees to absorb carbon dioxide (CO_2) plays a crucial role in mitigating climate change and reducing the greenhouse effect. As rampant deforestation continues, the propensity for extreme temperature events will become more frequent (42), in addition to the increase in occurrences of infectious (43) and respiratory diseases (44), highlighting the importance of the topic addressed.

Through accurate data and reliable forecasts, it is possible to implement preventive actions to protect critical areas and, consequently, reduce carbon emissions. Therefore, advanced monitoring projects and forecasting models must be implemented so that these objectives are achieved and thus guarantee a better future for our forests.

3.2 Data

During the research exposed in this document, two data sets were used to assemble and train the models and algorithms used. The data described below were obtained from INPE (National Institute for Space Research) and INMET (National Institute of Meteorology). It will also be shown how the data was treated and studied.

In addition to the bases described above, two databases were used with the polygons of the states and municipalities that make up the Legal Amazon region. In Figure 3.1, it is possible to see the representation of the data found in these databases, there are nine states and 773 municipalities (45) (46).

(a) Legal Amazon - States (b) Legal Amazon - Municipalities

Figure 3.1: States and Municipalities that constitute the Legal Amazon

3.2.1 Deforestation Data

The collection of data related to deforestation in the Legal Amazon has been taking place since 1988 through PRODES (Satellite Monitoring Project in the Legal Amazon). The study of this data takes place using LANDSAT images and the data is made available annually to the public. From 2004 onwards, the DETER project was started and its objective was to release data on deforestation in the Amazon more quickly and accurately, such information was collected by the MODIS sensor on the Terra satellite.

In 2015, a new version of this project was started, called DETER-B (47). The information is collected by the WFI sensor on the CBERS-4 satellite (Sino-Brazilian Land Resources Satellite) and the AWiFS sensor on the IRS satellite (Indian Remote Sensing Satellite), making deforestation data more accurate. This adjustment was necessary because the deforested areas identified in the previous project only captured changes of at least 25 hectares. However, the

deforestation pattern has evolved, making it increasingly difficult to find areas of this size being continuously deforested. In this context, DETER-B can identify and map changes in vegetation cover with a minimum area of almost 3 hectares, making its data collection more optimized and accurate. However, in general, polygons made available to the public have a minimum area of 6.25 hectares.

Figure 3.2 indicates the projects that were carried out to monitor deforestation in the Legal Amazon, along with the periods in which they operated until the emergence of another project linked to monitoring. It is worth mentioning that the projects were not terminated, PRODES is still functioning, however, there are other better monitoring projects currently in Brazil. DETER-B remains the newest and most effective project regarding monitoring and data collection in this region.

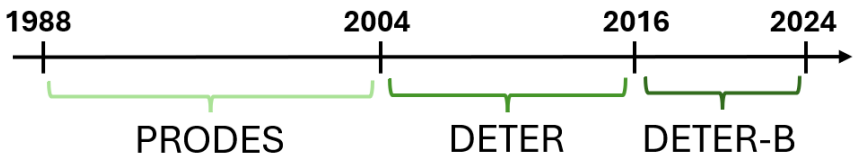


Figure 3.2: Projects on monitoring deforestation in the Legal Amazon

The data used in this research refers to the period from August 2016 to November 2023, collected from the DETER-B project (48), satellite images were not used directly, but the database containing the tabular values of the logging. It is worth noting that as the data are indirect measurements from satellites, the reliability of the data may be affected, for example, by adding bias. Each line of the data frame used indicates the value (km^2) and the polygon of the deforested area. In addition, information such as municipality, state, municipality's IBGE code, and date related to the occurrence of deforestation are also made available. In Table 3.1 it is possible to see a sample of the data lines present in the database used in this research.

Date	IBGE	State	Municipality	Area	Geometry
2016-08-02	1506807	PA	Santarém	0.201149	
2016-08-02	1505486	PA	Pacajá	0.125868	
	...	...	...	...	...
2023-11-24	1504950	PA	Nova Esperança do Piriá	0.420862	
2023-11-24	5104526	MT	Ipiranga do Norte	0.147052	

Table 3.1: Deforestation Database

After data collection, information must be processed to prepare it for exploratory analysis and assembly of models. Thus, to test whether changing the granularity of information would change the efficiency of future models, the data was grouped in four ways: by municipality and by week; by municipality and by month; by state and by week; by state and by month. Not all data will be predicted, that is, municipalities and states will be selected based on the information obtained in the exploratory data analysis.

Once the data was grouped, a search was conducted to identify missing values. For each date where deforestation information was absent, the value 0.01 was inserted. This choice was made because if an area in a region was not deforested, the deforestation value was either insufficient or nonexistent in the original database. The value 0.01 was selected instead of 0.0 to avoid potential issues when calculating future metrics. Additionally, this value is smaller than the smallest value present in the database used. For the deforestation databases created, there was no need to handle null values (NaN), as such values did not appear in the datasets.

3.2.2 Meteorological Data

In addition to the deforestation data, information from 148 meteorological stations located in the Legal Amazon was collected. These datasets, provided by INMET (49), include hourly measurements taken at each station. Among the available information, it was used the columns related to the date and time of information collection, total precipitation, in millimeters (mm), during the hour; maximum temperature in the previous hour, in degrees Celsius ($°C$); relative air humidity (%) during the hour; the speed of the maximum gust of wind, in meters per second (m/s), during the hour.

It is worth mentioning that INMET uses the anemometer to record the instantaneous wind speed, the rain gauge to measure the amount of precipitation in a given period, the psychrometer to measure the relative humidity of the air, and the maximum and minimum thermometer to indicate maximum and minimum temperatures. These devices indicate specific information from rain, wind, temperature, and humidity data. However, this data will be used for state and municipal regions. It will be studied whether changes in the granularity of the areas considered and periods reflect changes in the results.

After the data had been successfully imported, the databases of identical meteorological stations were concatenated, as the data was divided into folders by year. Therefore, an analysis must be made regarding which stations should be considered so that they can be included in future models. Based on this, it was first checked which stations had data in all years from 2016 to 2023, then which stations had more than 70% of non-missing data. The stations that covered these two rules were considered for use in the models, totaling 65 meteorological stations.

Since null values could still be found in the databases of each station, it was necessary to fill in such values. For rain data, null values were replaced by 0. Therefore, they were treated as if there was no precipitation at the given moment. Now, concerning humidity, temperature, and wind data, another method was used to treat missing data. As it was common to find sequential periods with null values, linear interpolation was used to calculate the missing values in the middle of the data frame. Meanwhile, for null values at the beginning of the desired periods, the "backfill" method was used, that is, the first non-zero value was propagated backward to the first base line and for null values at the end of the period, the "ffill", that is, the last non-null value was propagated forward to the last baseline. Figure 3.3 shows how the methods for filling in null values were treated.

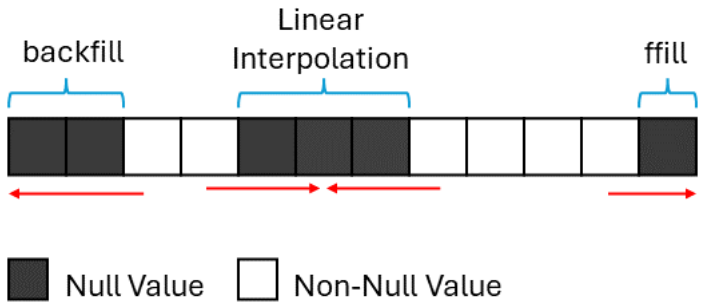
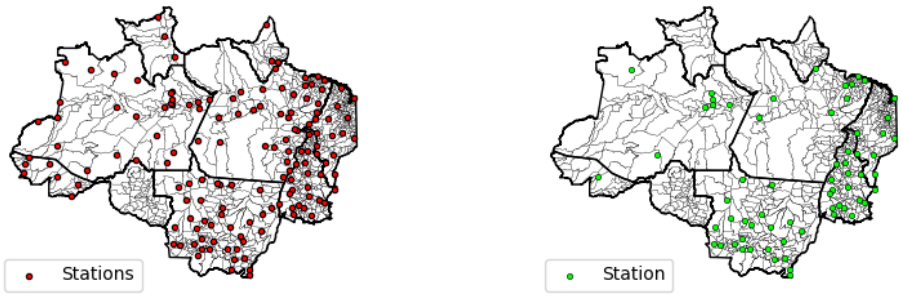


Figure 3.3: Data filling in climate databases

Taking into account Figure 3.4 it is possible to examine a comparison

of the total number of meteorological stations and the meteorological stations that are being considered. It is worth noting that there are states without any meteorological stations, such as Rondônia and Roraima. Furthermore, it is seen that most states do not have many stations in their territory, except for Mato Grosso (23) and Tocantins (16).



(a) All weather stations (b) Selected weather stations

Figure 3.4: Comparison between all weather stations and selected stations

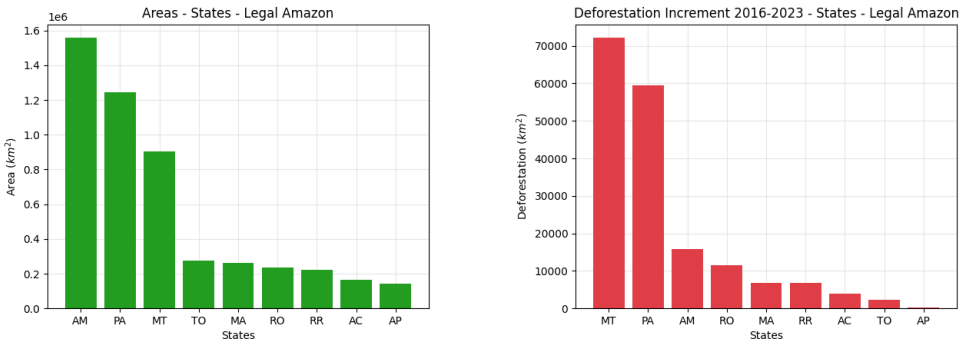
Then, the meteorological data were grouped so that they could be examined in the future in exploratory analysis and applied to the proposed models. The groupings occurred in a similar way as was done with the deforestation data, four groups were assembled: by state and month; by state and week; by station and month; and by station and week. When the groups were made, data related to rain were added together. While, data related to humidity, temperature, and wind were calculated through the average of the values per group. Regarding the data grouped by week, it was ensured that the days of the week were the same as those used for the groupings by week carried out with the deforestation data.

3.3 Exploratory Data Analysis

Given that the datasets are refined and complete, the selection of states and municipalities for the future construction of the models can proceed. First, the choice of states and the exploration of the data will be shown, then it will be seen how this choice was made for the municipalities.

The states of Amazonas (AM), Mato Grosso (MT), Pará (PA), and Rondônia (RO) were selected, as they are the states with the highest deforestation rates and largest areas within the Legal Amazon, as can be seen in Figure 3.5. Furthermore, the reason why it was not chosen states such as Maranhão (MA) and Tocantins (TO) was due to the high percentage of deforested forest

area presented in these states, as can be seen in Figure 3.6, even though they present state area larger than RO, for example.



(a) Areas of the states that constitute the Legal Amazon

(b) Aggregated area deforested by state of the Legal Amazon from Aug/2016 to Nov/2023

Figure 3.5: Total area and aggregate deforestation by state during the period of interest

Through Figure 3.6 it is possible to examine that the municipalities that show the greatest variation in the percentage deforested area are located in MT and PA. The main reason for this to happen is that they are located in the expansion of Brazilian agribusiness (50). Other factors that can act as catalysts for deforestation are: road construction, city expansion, and population growth (44). Therefore, the choice of MT and PA becomes more interesting since these are the largest current deforestation hotspots in the Legal Amazon.

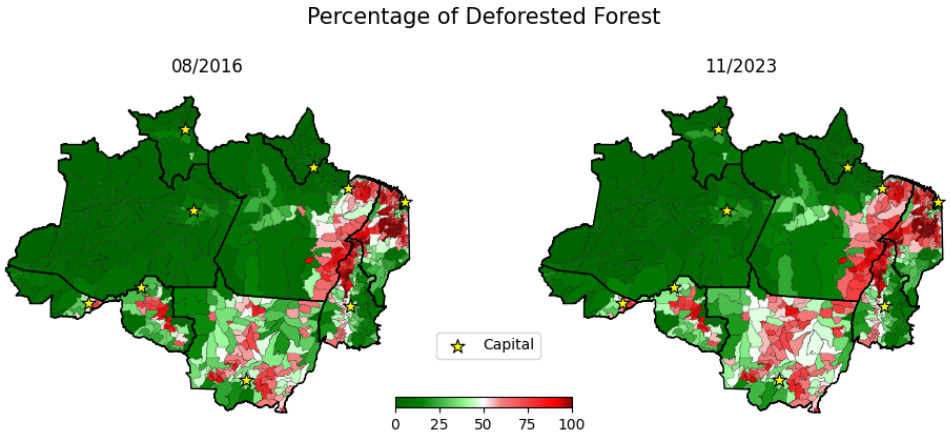


Figure 3.6: Comparison of the percentage of deforested forest in each municipality in the Legal Amazon

Furthermore, although it is not possible to inquire about many changes in the percentage of deforested areas in the states of Amazonas and Rondônia,

another reason for choosing these states is their proximity to PA and MT. This fact could imply possible outbreaks of deforestation in the future, as well as a growing presence of agribusiness. Thus, Figure 3.7 shows the states that were selected for the research. It is worth mentioning that although the states of Tocantins and Maranhão are major deforestation hotspots and have areas that have not yet been deforested, most of them are present in irregular and high-altitude terrains, factors that make deforestation in the region difficult.

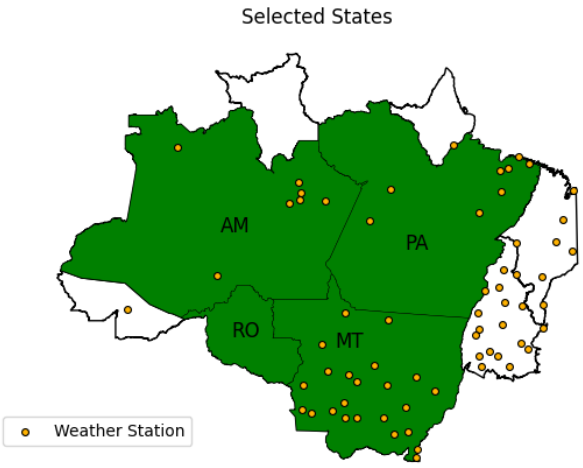


Figure 3.7: States selected for analysis and application of the proposed models

After selecting the states, it was possible to examine in more detail the increase in deforestation in each of them. Through Figure 3.8 deforestation presents a periodic behavior in all selected states, with peaks more present in July, August, September, and October. This information is reinforced in Figure 3.9 with aggregate deforestation data per month. Furthermore, it is observed that the periods with higher deforestation are more aggressive in the states of MT and PA, highlighting the data examined in Figure 3.6. Another point that highlights this information is the data found in Table 3.2 which indicates the values of the mean and standard deviation of the series and shows that the state of Mato Grosso is the one with the highest average deforestation in the given period.

Furthermore, using the standard deviation values it is possible to examine the variability of the data presented in each state, the values reinforce the idea of moments of low and high deforestation rates. The table also shows the mean and standard deviation of the weekly aggregated deforestation series, the concept of which will be explained in Chapter 4. Another point that can be addressed based on historical deforestation data is that in general substantial

variations and averages are seen as we analyze more current information. This is most notable in the states of Amazonas, Mato Grosso, and Pará. When we look at the weekly increment data the differences between the sets are not very marked except for two cases the Amazonas test set and the Pará validation set.

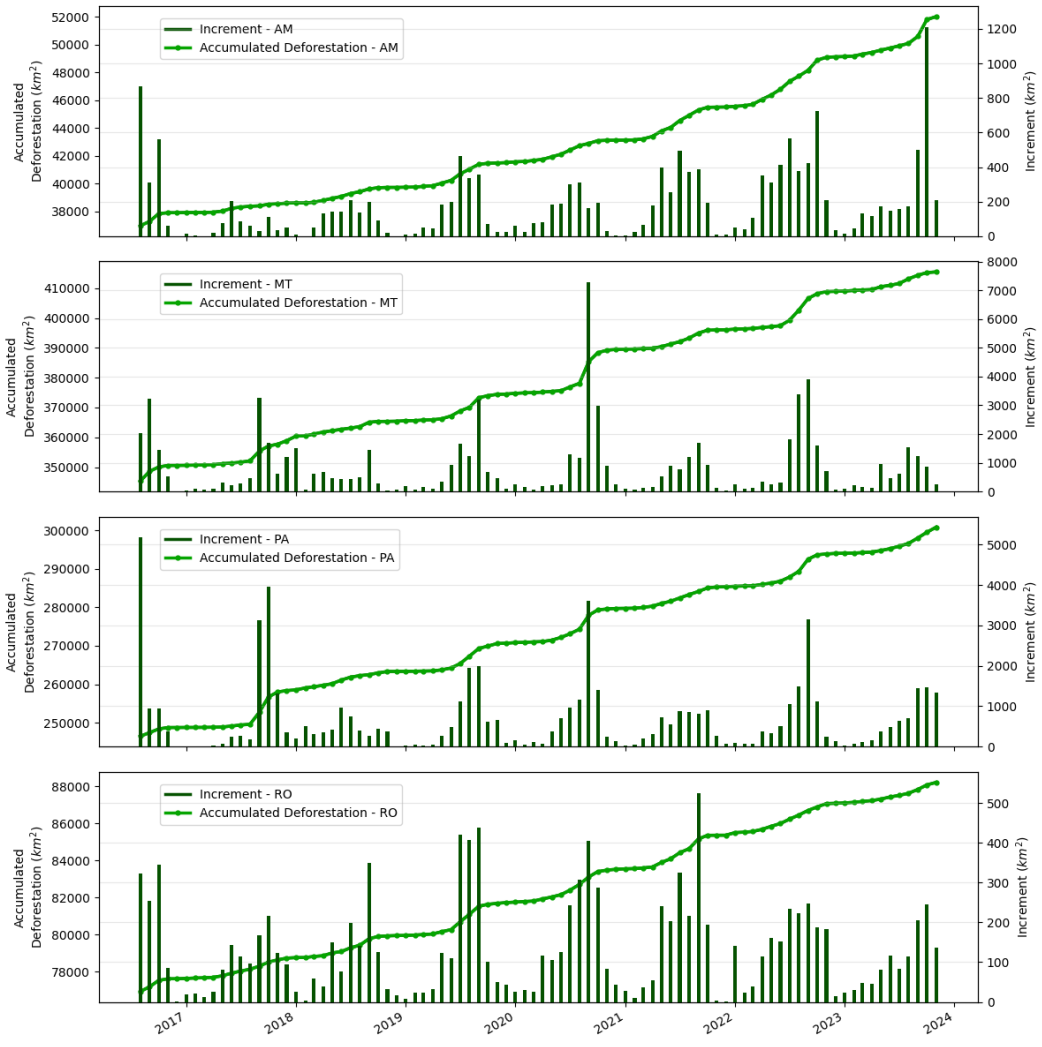


Figure 3.8: Accumulated deforestation and increase in deforested forest in selected states

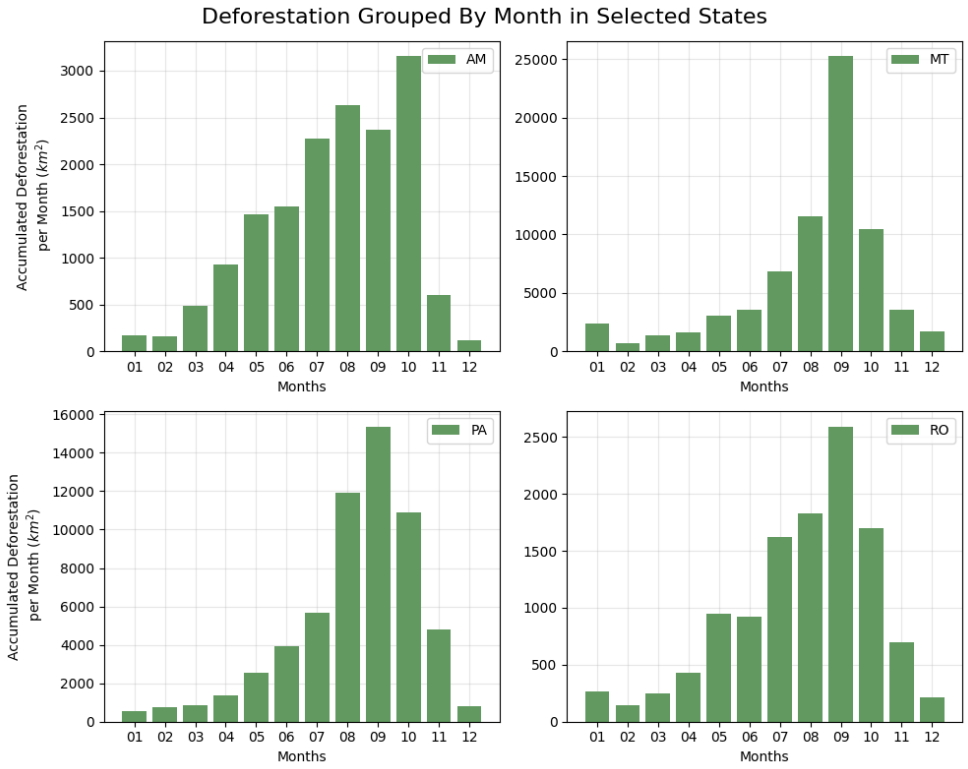


Figure 3.9: Total deforestation per month during the analyzed period

States	Monthly		Weekly Aggregated	
	Mean	Standard Deviation	Mean	Standard Deviation
AM	180.6	206.8	165.6	186.4
MT	819.9	112.48	762.0	1038.2
PA	676.4	913.8	612.3	827.8
RO	131.9	119.4	121.6	110.0

Table 3.2: Mean and standard deviation of deforestation series data from the selected states.

One of the possible reasons for periodic behavior in deforestation is due to the climate of the Amazon region, in which there is more rain during the summer and autumn and little or no rain during the winter and spring. Other climatic variables can also directly affect the rate of increase in deforestation, such as temperature, humidity, and wind. Figures 3.10, 3.11, and 3.12 presents graphs of climate data grouped by month for the states of AM, MT, and PA, respectively. It was not possible to carry out an analysis of RO's climatic variables, as all meteorological stations located within its borders were not selected. It is noticeable that the series of humidity, rain, and wind present an inversely proportional relationship with the increment variables. However, it must be analyzed the temporal lags of each of these series and verify their relevance for future data modeling.

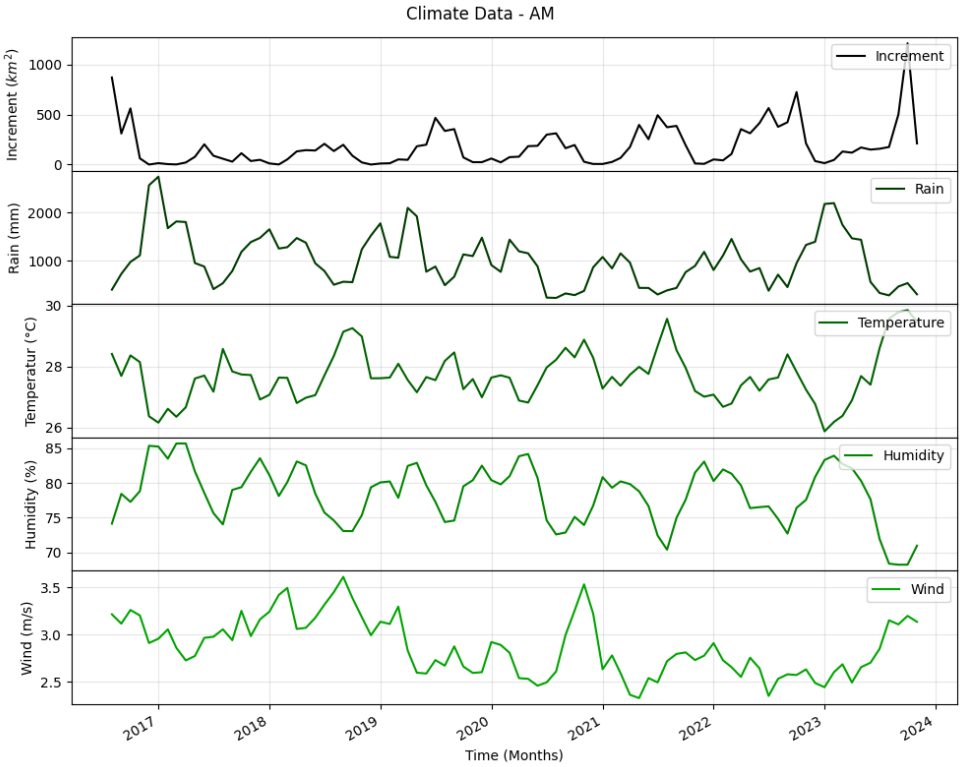


Figure 3.10: Climatic data by month from meteorological stations in the state of Amazonas

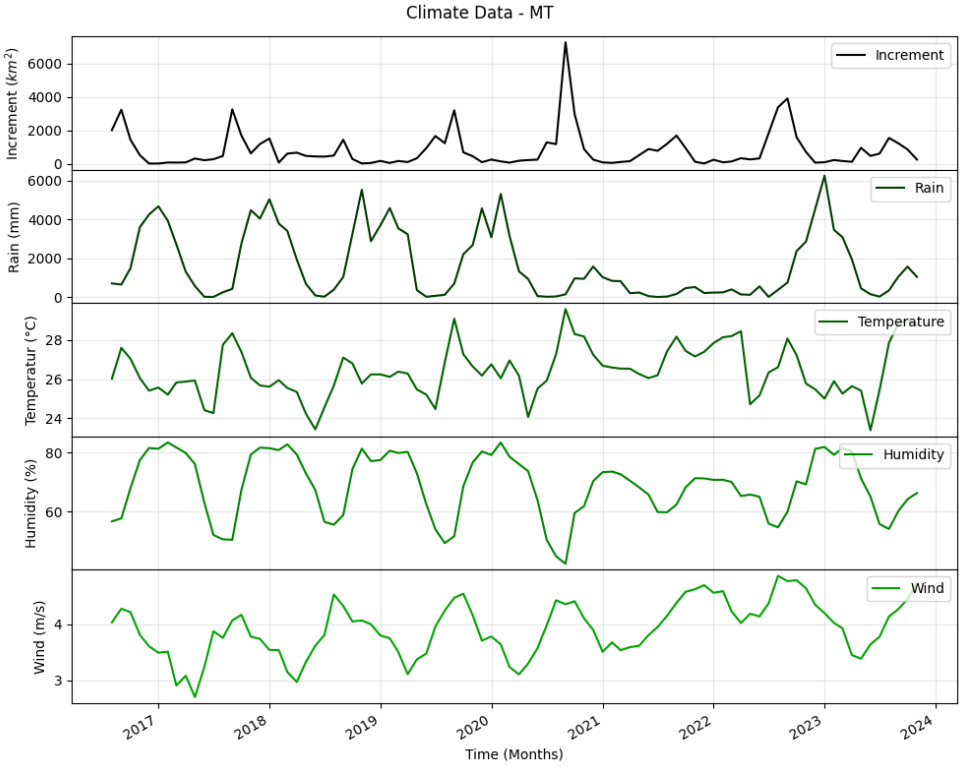


Figure 3.11: Climatic data by month from meteorological stations in the state of Mato Grosso

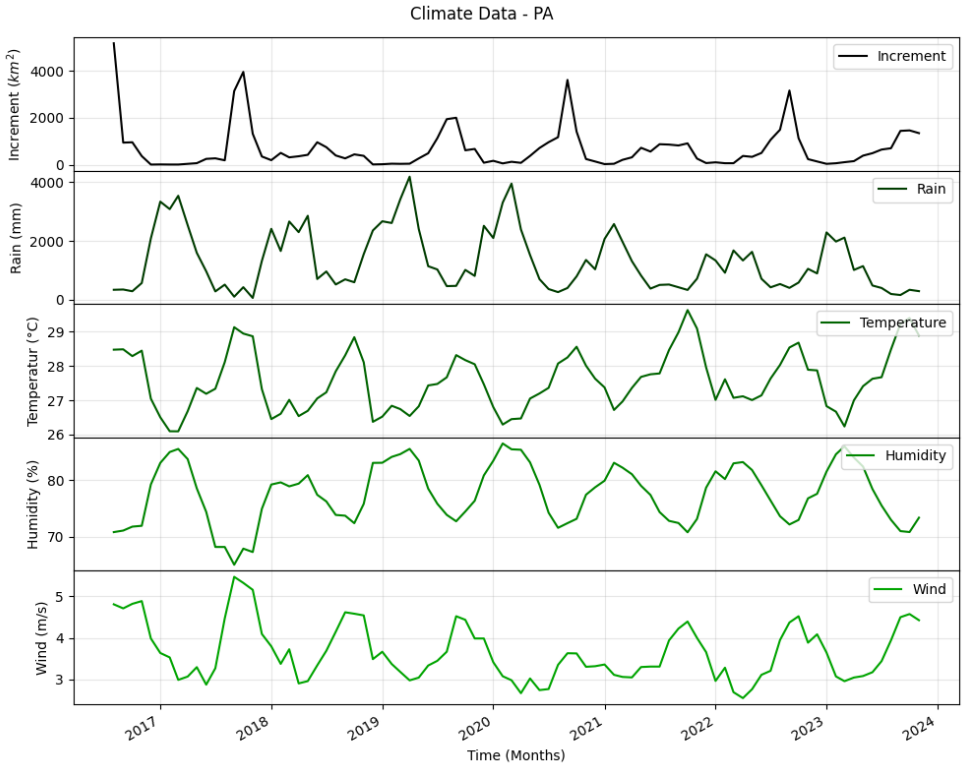


Figure 3.12: Climatic data by month from meteorological stations in the state of Pará

Now, in Figures 3.13, 3.14, and 3.15 it is possible to see the weekly data of humidity, rain, wind, and temperature series for the states of AM, MT and PA, respectively. A similar analysis can be made by noting that in periods with high humidity and precipitation, especially, deforestation is especially lower in the states, implying an inverse relationship between deforestation and the mentioned variables.

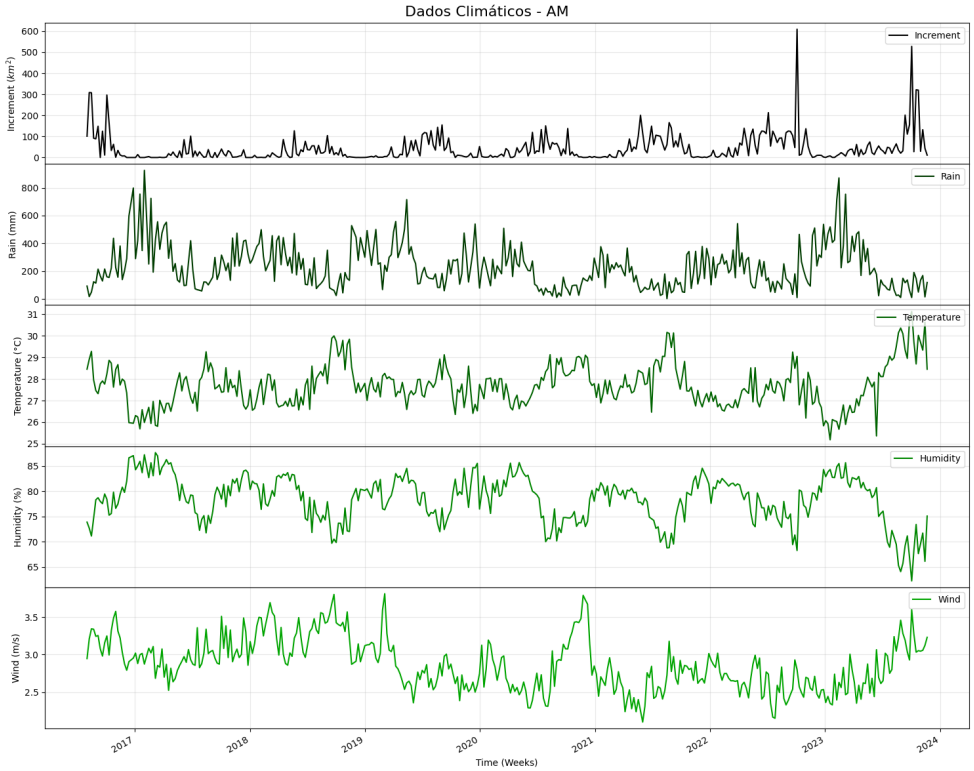


Figure 3.13: Climatic data per week from meteorological stations in the state of Amazonas

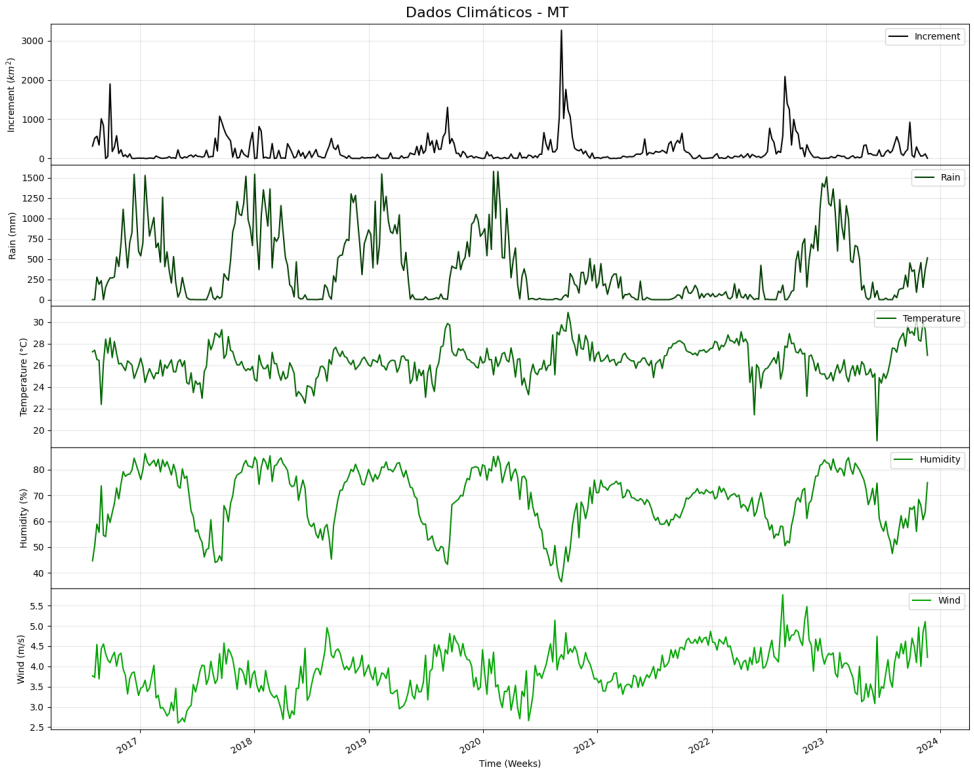


Figure 3.14: Climatic data per week from meteorological stations in the state of Mato Grosso

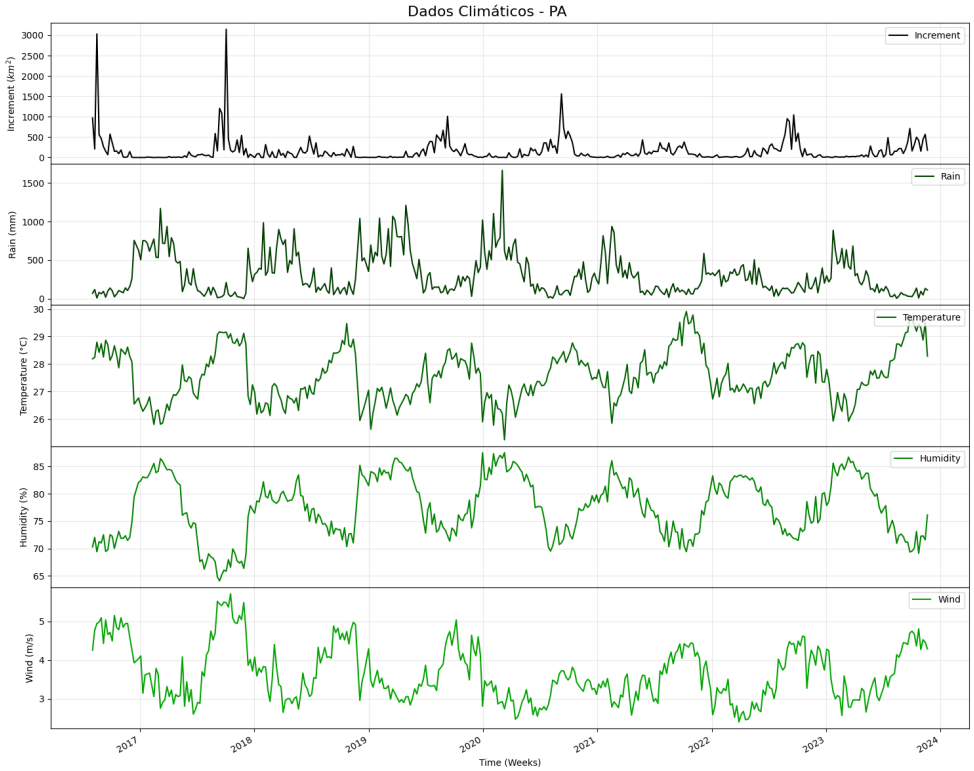


Figure 3.15: Climatic data per week from meteorological stations in the state of Pará

In addition to the variables extracted from the meteorological stations, other variables were created to assist in future data modeling. As they are being grouped by month and week, two variables were schematized so that more information can be obtained about each period. The first variable created was made by adding the number of hours in which the occurrence of rain was recorded. The other variable was created by the number of days it rained during the week or month, depending on the database considered. The Figures 3.16, 3.17, and 3.18 and the Figures 3.19, 3.20, and 3.21 show the values obtained with each one of the variables created from the data grouped by month and week respectively. Via the images, a periodic behavior is notable, mainly of the variable "Hours of Rain", this comes naturally since the climatic precipitation data also presents a seasonal behavior. Even though the "Rainy Days" variable also originated from precipitation data, the periodicity is not observed as much and it is not as discretized as the other feature. Furthermore, a lower magnitude of values can be seen in the year 2021 in the graphs, which may have been caused by droughts or lack of data in certain stations within the states.

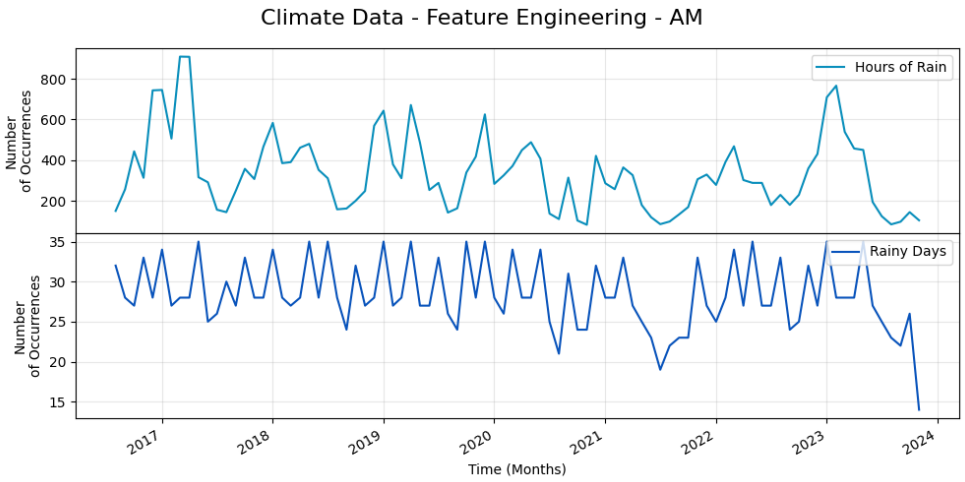


Figure 3.16: Climate data created by Feature Engineering by month from the meteorological stations in the state of Amazonas

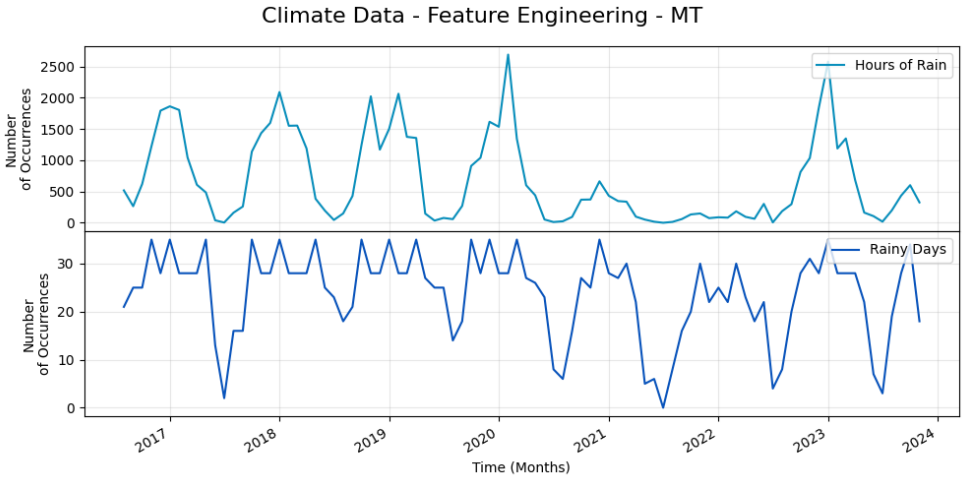


Figure 3.17: Climate data created by Feature Engineering by month from the meteorological stations in the state of Mato Grosso

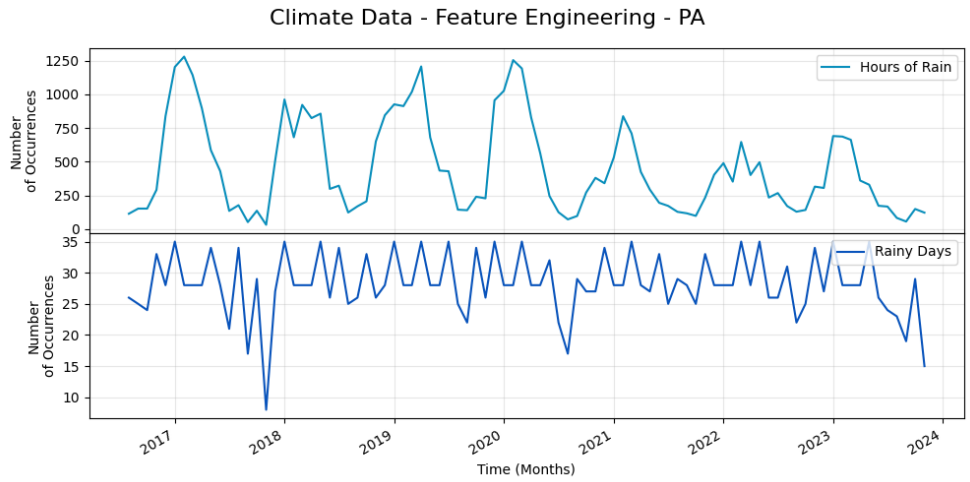


Figure 3.18: Climate data created by Feature Engineering by month from the meteorological stations in the state of Pará

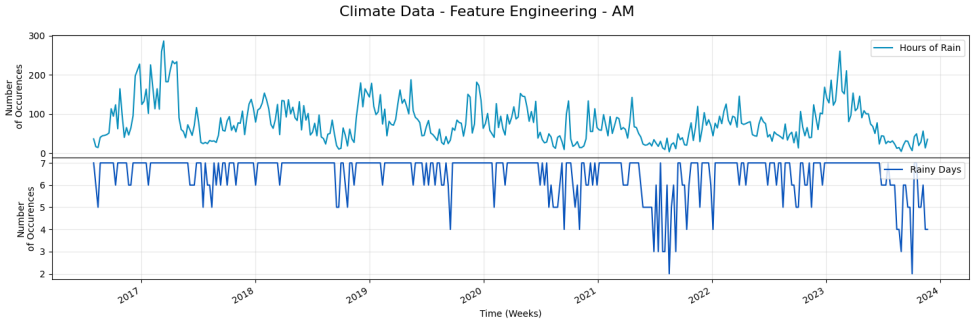


Figure 3.19: Climate data created by Feature Engineering per week from the meteorological stations in the state of Amazonas

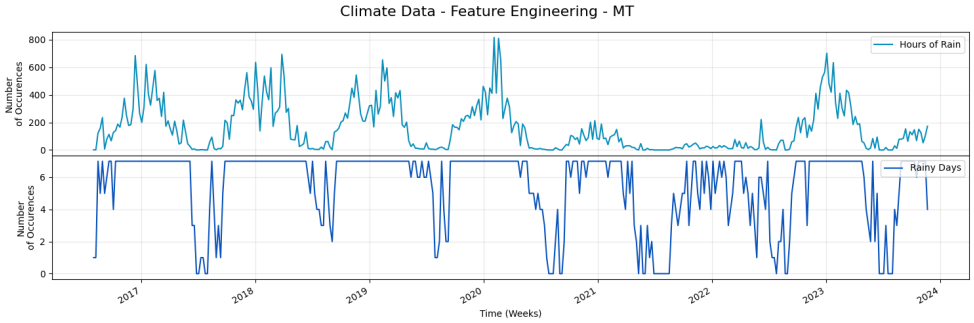


Figure 3.20: Climate data created by Feature Engineering per week from the meteorological stations in the state of Mato Grosso

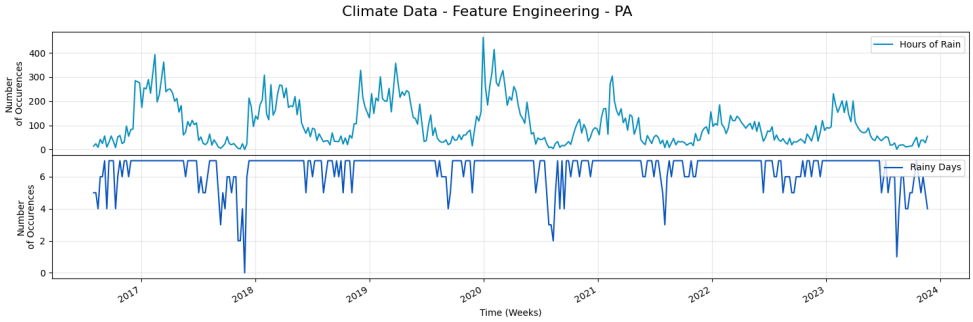


Figure 3.21: Climate data created by Feature Engineering per week from the meteorological stations in the state of Pará

To obtain the best model for each proposed algorithm, the correlation between the values of the increment series of each selected state and the past observations of the state series itself was calculated. Figure 3.22 shows how the series was considered to calculate the correlations.

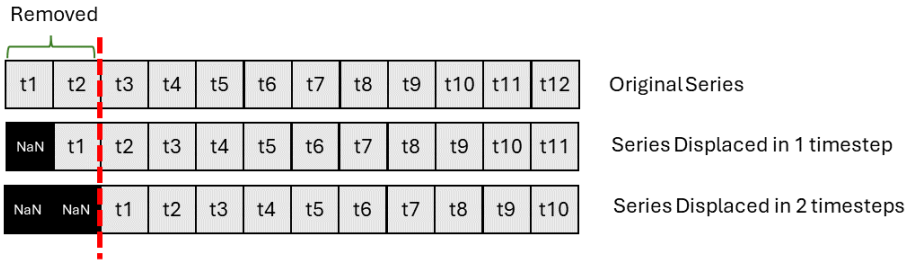


Figure 3.22: Assembling the Series for Calculating Correlations

The results obtained are represented in Figures 3.23 and 3.24, for monthly data correlations were calculated with up to 12 lags and for weekly data, up to 52 steps passed. It can be analyzed that, both in the weekly and monthly cases, the correlation starts positive and transitions to negative values, reinforcing the information collected previously that deforestation has periodic trends throughout the year, with moments of drops and increases in deforestation. Furthermore, more recent deforestation values tend to have more correlation with the values intended to be predicted, as well as values that are further away from the time step in question. Likewise, values close to 6 months or 26 weeks before the time of interest demonstrate a negative correlation, that is, they are inversely proportional to the values of interest.

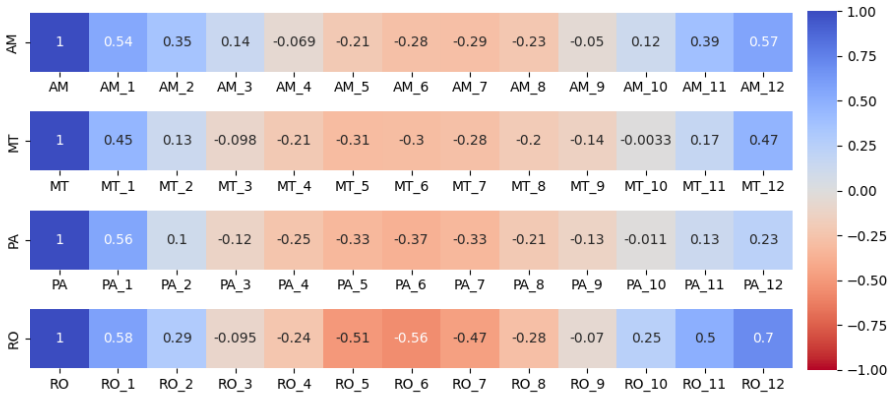
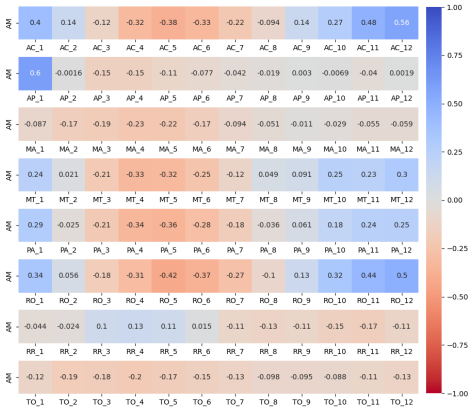


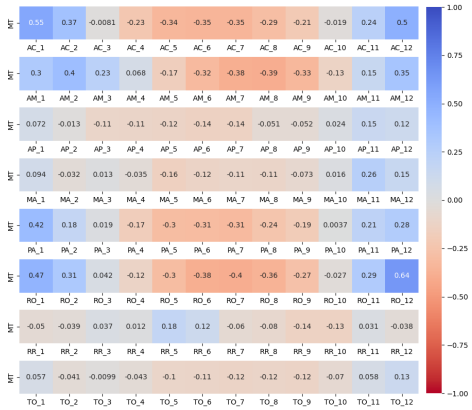
Figure 3.23: Correlations between the past times of the series, in months, and the time of interest

Figure 3.24: Correlations between the past times of the series, in weeks, and the time of interest

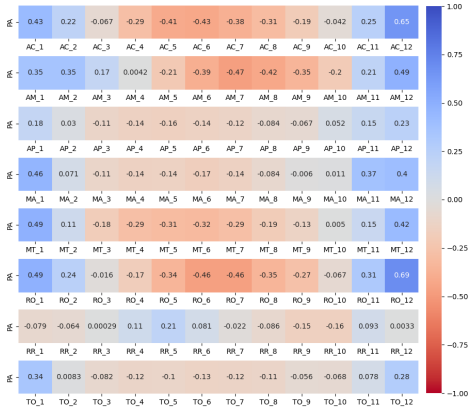
Then, the same correlation calculation was done with past values from other states. Figure 3.25 shows the correlations for monthly data considering the last 12 time steps, while Figure 3.26 shows the correlations for weekly data considering the last 52 time steps. Such correlations were calculated with the aim of identifying more variables that could help predict the data. Through the graphs shown, it is possible to observe that each state correlates with several time steps of other states, indicating that these may be variables that add information for the assembly of future models. However, not all variables will be useful for modeling, requiring future filtering.



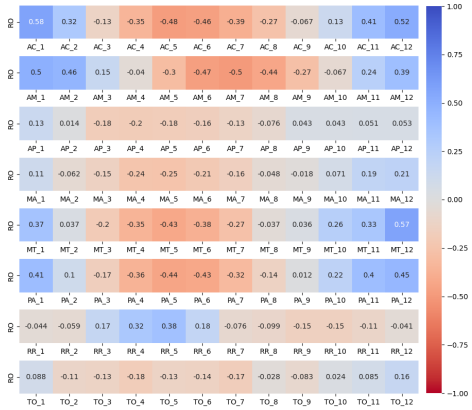
(a) Correlation with other states in the Legal Amazon - Amazonas (AM)



(b) Correlation with other states in the Legal Amazon - Mato Grosso (MT)

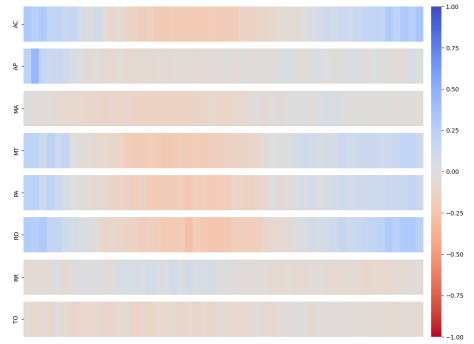


(c) Correlation with other states in the Legal Amazon - Pará (PA)

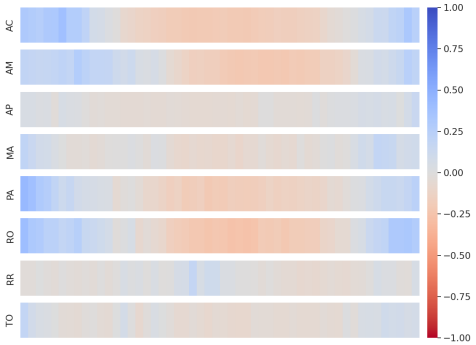


(d) Correlation with other states in the Legal Amazon - Rondônia (RO)

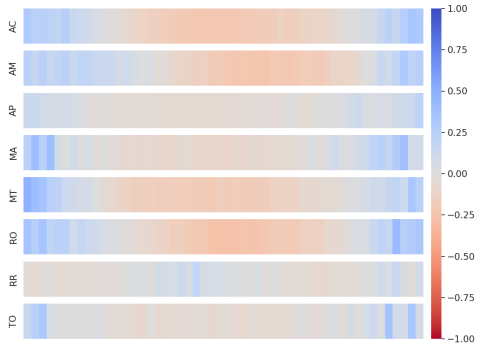
Figure 3.25: Correlations with other states - Months



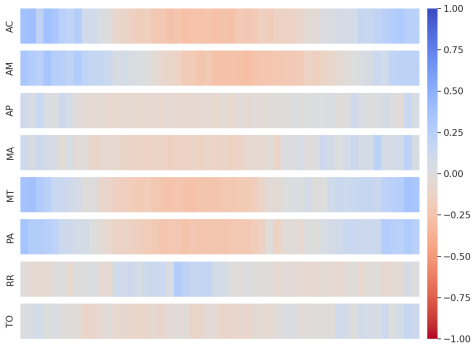
(a) Correlation with other states in the Legal Amazon - Amazonas (AM)



(b) Correlation with other states in the Legal Amazon - Mato Grosso (MT)



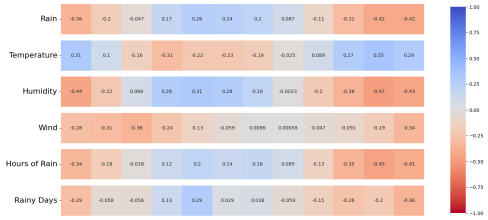
(c) Correlation with other states in the Legal Amazon - Pará (PA)



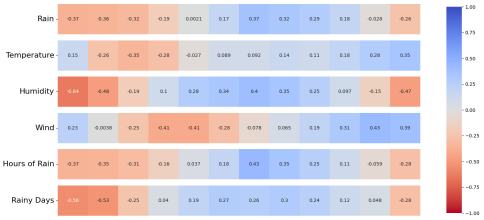
(d) Correlation with other states in the Legal Amazon - Rondônia (RO)

Figure 3.26: Correlations with other states - Weeks

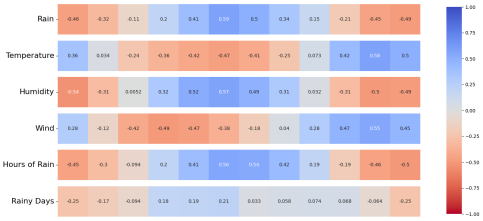
Finally, analyses were carried out on the correlations of climate variables concerning the series of deforestation increases in each state. In Figure 3.27 it is shown the correlations of monthly data with the last 12 time steps and in Figure 3.28 the correlations of weekly data with the last 52 time steps of climate data. Regarding monthly data, the states of MT and PA show a significant correlation with lags of humidity variables and days that had rain, AM did not show as significant correlation as the other two states. For monthly data, the correlations were not very high, however, the series of humidity and days of rain in MT and temperature, wind, humidity, and hours of rain in PA can be highlighted. In the state of Amazonas, no highlights were obtained in this regard.



(a) Correlation with climate variables - Amazonas (AM)

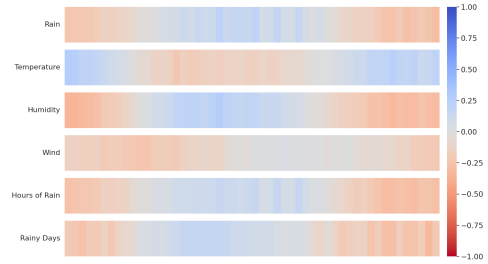


(b) Correlation with climate variables - Mato Grosso (MT)

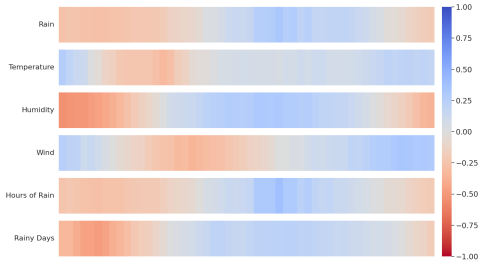


(c) Correlation with climate variables - Pará (PA)

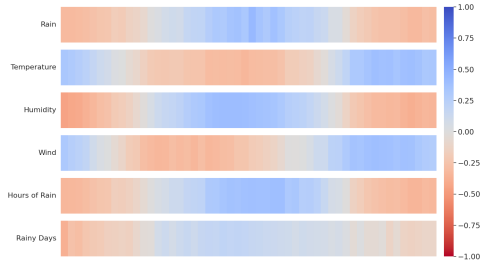
Figure 3.27: Correlations with climate variables - Months



(a) Correlation with climate variables - Amazonas (AM)



(b) Correlation with climate variables - Mato Grosso (MT)



(c) Correlation with climate variables - Pará (PA)

Figure 3.28: Correlations with climate variables - Weeks

Bearing in mind that all properties relevant to the work on the selected states have already been discussed, the same analysis will be carried out for the chosen municipalities. To do this, the analyses of Figure 3.29 need to be made, and from there the selection of municipalities that have the majority of non-missing data, both in the monthly and weekly analysis. Furthermore, municipalities were chosen that were close to at least one meteorological station and that have not been extremely deforested, as seen in Figure 3.6. Thus, the municipalities chosen were Lábrea (AM), Apiacás (MT) and Itaituba (PA). In Figure 3.30 the chosen municipalities are highlighted.

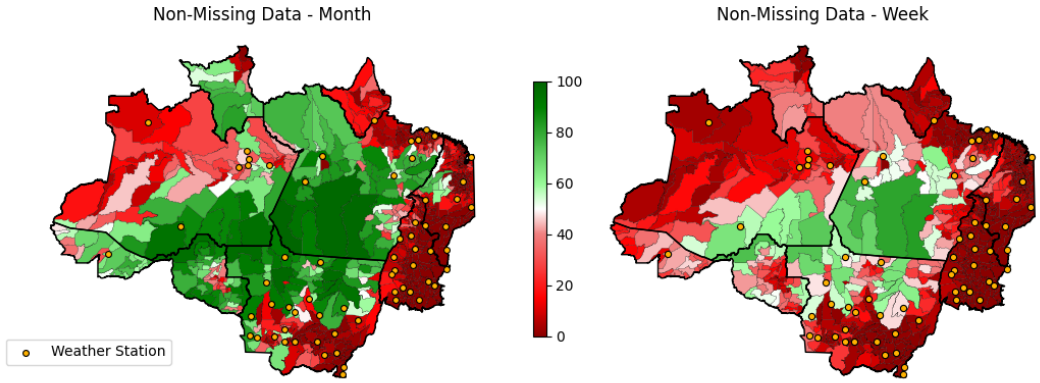


Figure 3.29: Percentage of non-missing data by municipality

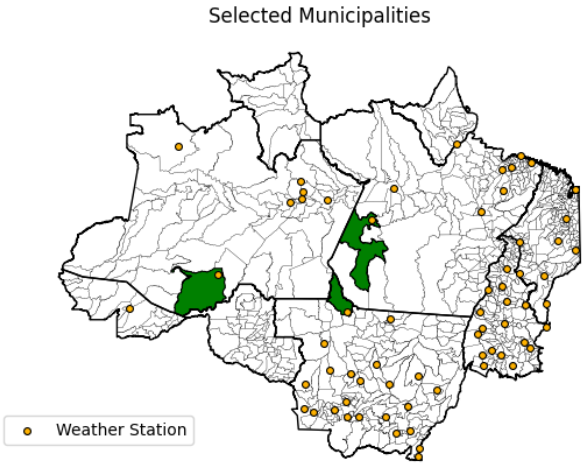


Figure 3.30: Municipalities selected for analysis and application of the proposed models

From the selection of municipalities, it is possible to analyze the deforestation series and increase in deforestation in each of them. Figure 3.31 presents these data in which it is possible to observe a periodic behavior, as demonstrated in the selected states, where higher rates of deforestation occur during the months of June to October. The indices show a decrease from 2022 to 2023, however, they remain, on average, at the same magnitude. Observing Figure 3.32 the months from June to October stand out for being the ones with the largest areas deforested during the period of interest, this is more evident in the municipalities of Apiacás and Itaituba, while in the municipality of Lábrea deforestation is more intense between the months of April to October. Table 3.3 indicates the mean and variance values for each of the monthly deforestation series in the selected municipalities.

An observation can be made about the training, validation, and test sets assembled for each location, unlike the case by state, sets of municipalities maintain a certain constancy in the mean and standard deviation values of such sets. Some changes can be noted concerning the behavior of the Itaituba deforestation series, in which the variations are less intense in the initial years, and about Apiacás, the emergence of more deforestation peaks from the second half of the data onwards. However, in general, the series tends to behave in a more standardized way when compared to the deforestation series of the states.

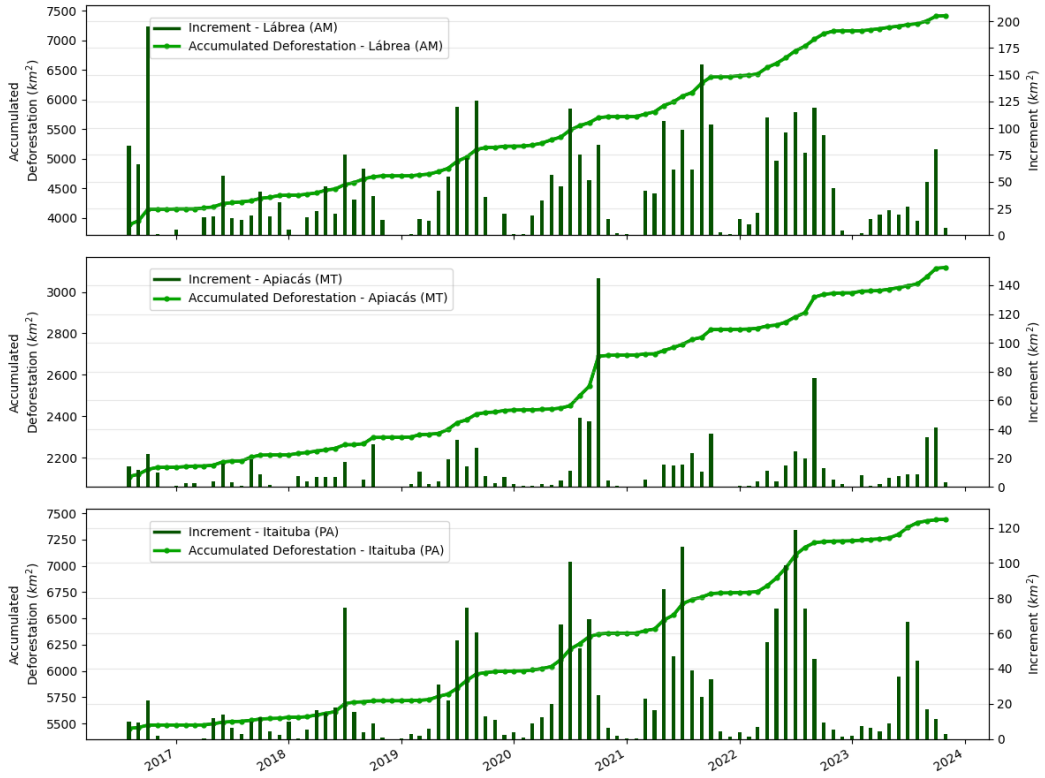


Figure 3.31: Accumulated deforestation and increase in deforested forest in the selected municipalities

Municipality	Mean	Standard Deviation
Lábrea	40.2	42.0
Apiacás	10.7	19.4
Itaituba	20.6	26.6

Table 3.3: Mean and standard deviation of the monthly data on the deforestation series of the selected municipalities.

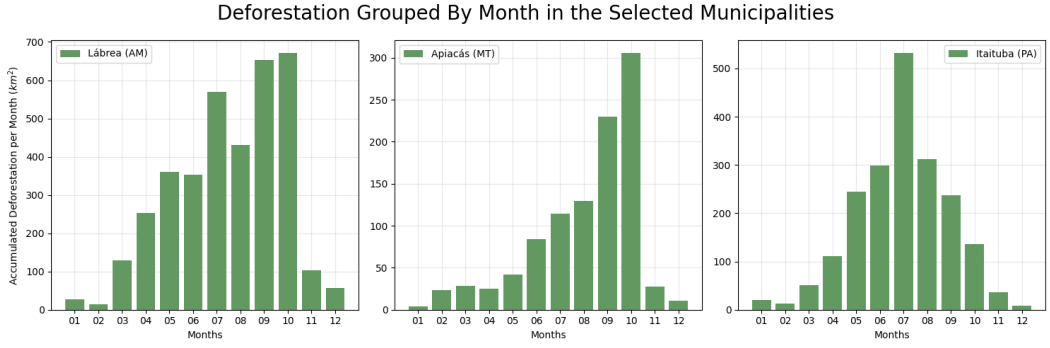


Figure 3.32: Total deforestation per month for each municipality chosen during the selected period

In such manner, it is possible to examine the graphs of the climate variables related to each one of the municipalities, represented by Figures

3.33, 3.34, and 3.35. For each municipality, the station closest to it was considered, that is, for Lábrea the station $INMET\_N\_AM\_A111$, for Apiacás the station $INMET\_CO\_MT\_A910$ and for Itaituba the station $INMET\ _N\_PA\_A231$. It is noticeable that for each station there are intervals in which the values function as linear functions, this is due to the fact that there was no data collection during these periods, implying a linear interpolation with the remaining data. This certainly prevents a complete analysis of the data, however, it is possible to analyze that generally periods with high precipitation and greater humidity present less deforestation and periods with lower precipitation and lower humidity, greater deforestation.

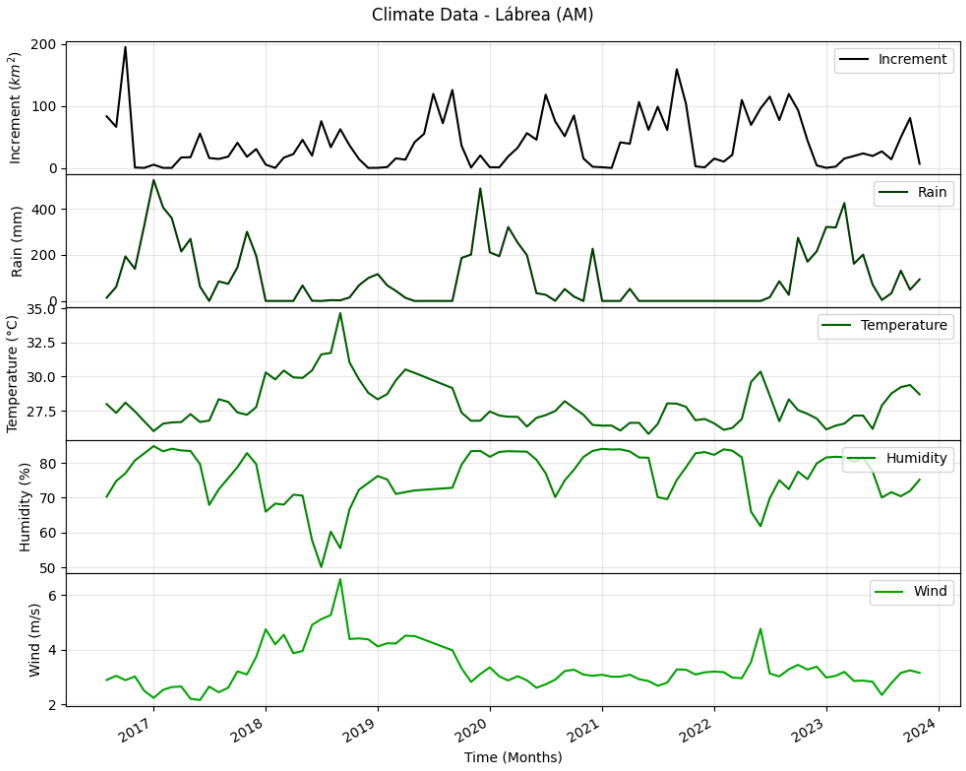


Figure 3.33: Climatic data by month from the meteorological station in Lábrea (AM)

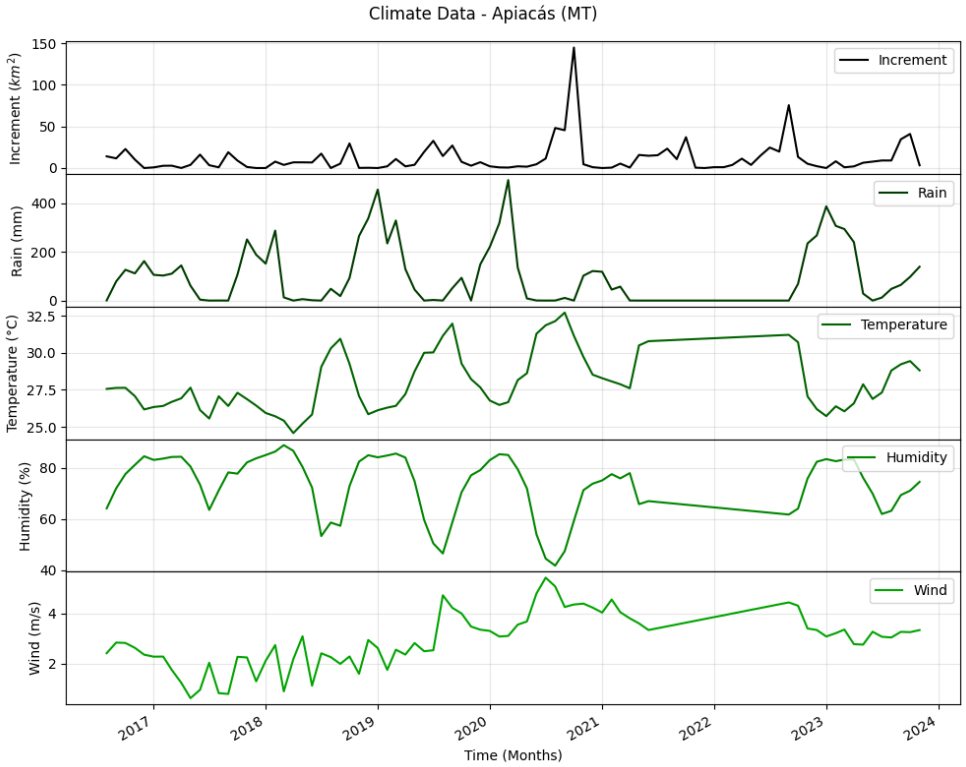


Figure 3.34: Climatic data by month from the meteorological station in Apiacás (MT)

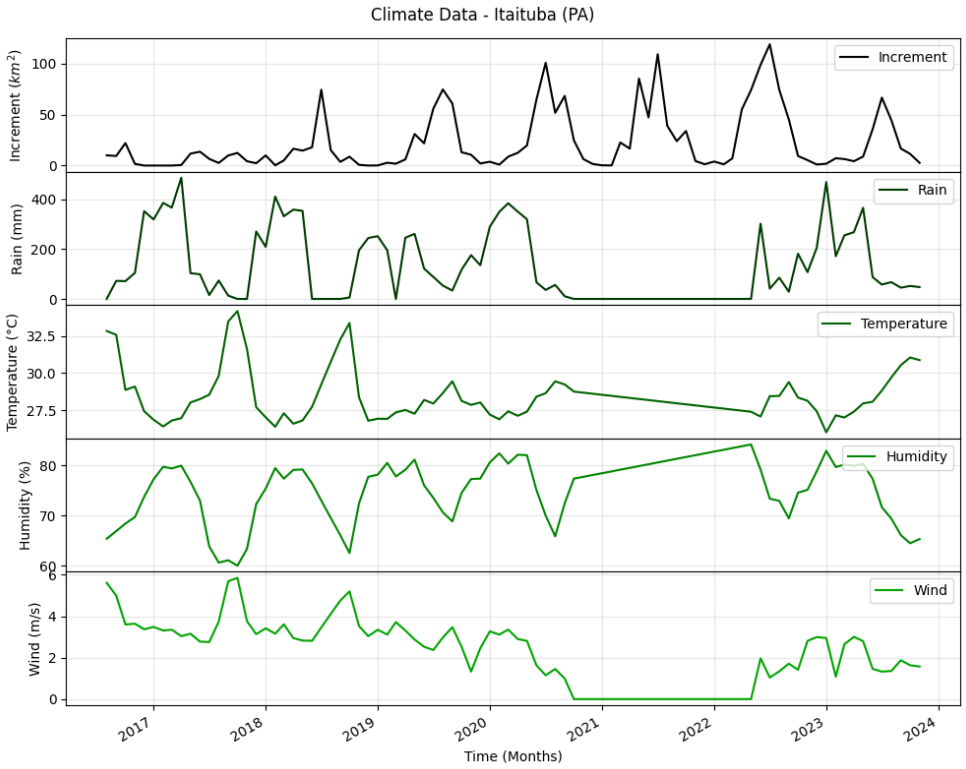


Figure 3.35: Climatic data by month from the meteorological station in Itaituba (PA)

Now, in Figures 3.36, 3.37, and 3.38 it is possible to see the weekly data of the humidity, rain, wind, and temperature series for the selected municipalities. Here it can also be pointed the same problems regarding missing data combined with interpolation of periods without data. Still, the same conclusions can be drawn regarding the humidity and precipitation series.

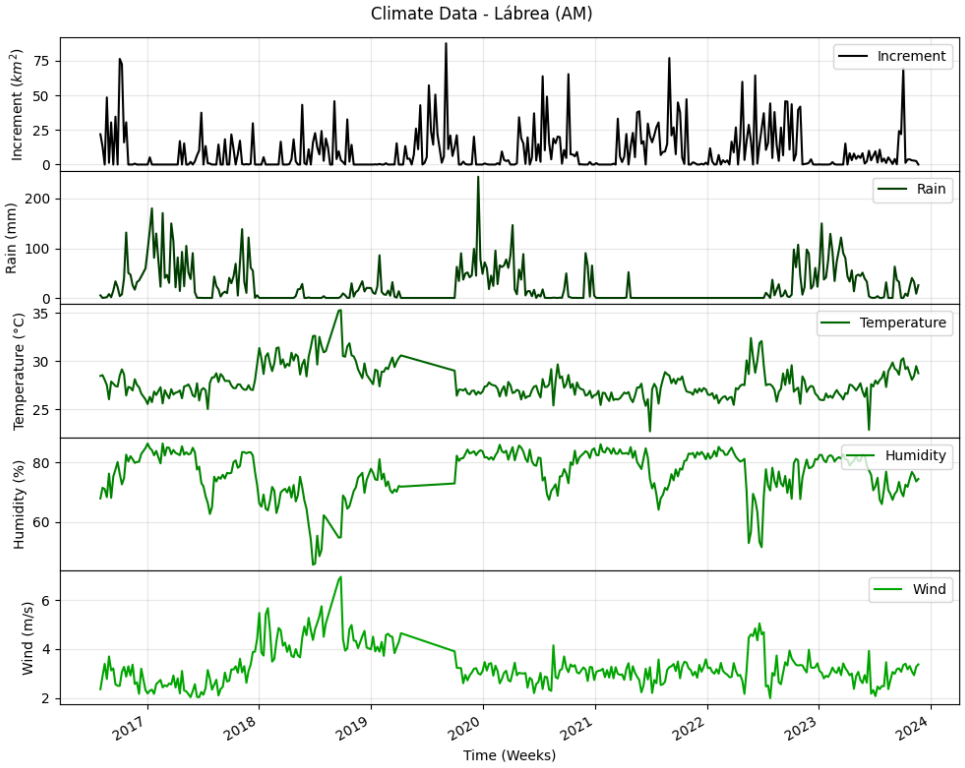


Figure 3.36: Climatic data per week from the meteorological station in Lábrea (AM)

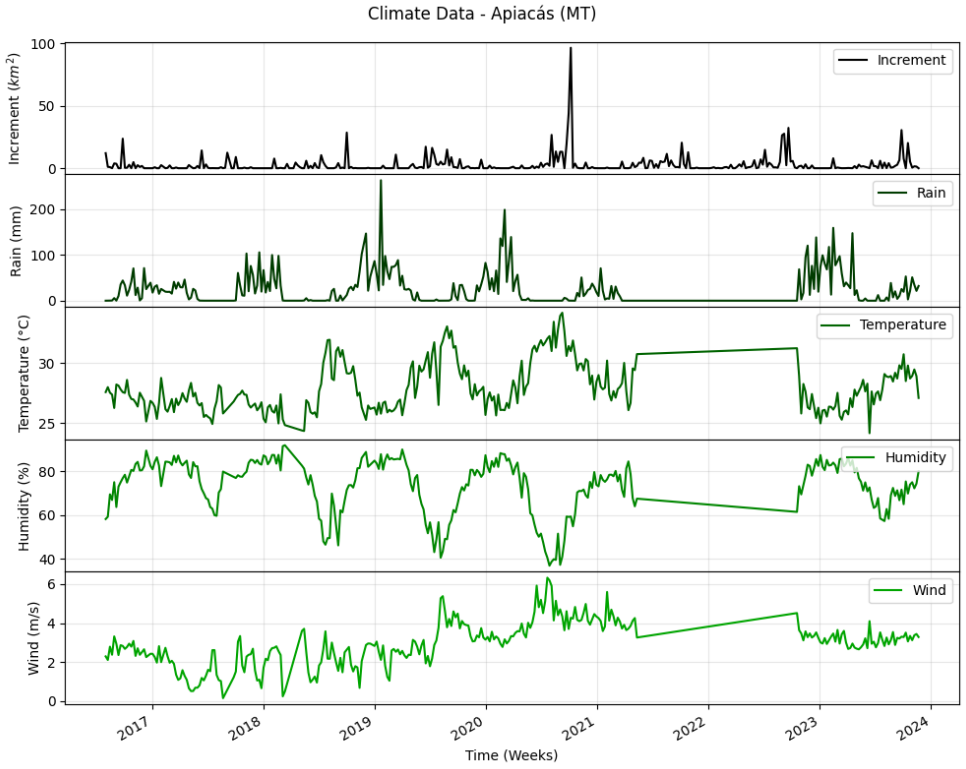


Figure 3.37: Climatic data per week from the meteorological station in Apiacás (MT)

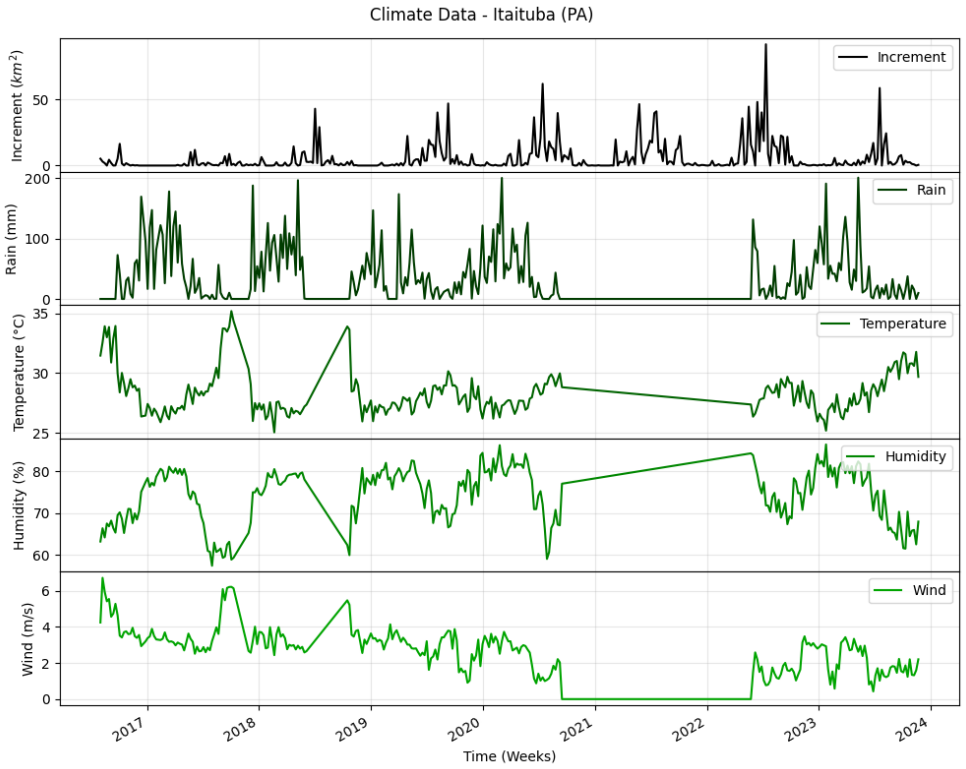


Figure 3.38: Climatic data per week from the meteorological station in Itaituba (PA)

The variable extraction process was also carried out with climatic data collected in the stations of interest to the municipalities. The two variables created were the number of hours in which the occurrence of rain was recorded and the number of days in which any occurrence of rain was recorded in the station. The data can be checked in Figures 3.39, 3.40, and 3.41 for the monthly analysis and in Figures 3.42, 3.43, and 3.44 for weekly analysis. Both variables present periodic behavior, considering that the precipitation series also presents periodic behavior, this is more noticeable for the weekly and monthly series in the municipalities of Apiacás and Itaituba.

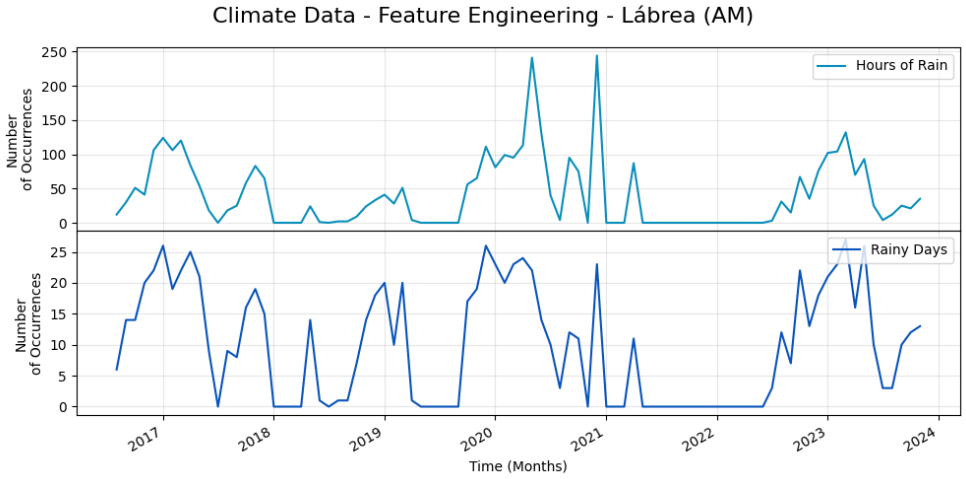


Figure 3.39: Climate data created by Feature Engineering per month from the meteorological station in the municipality of Lábrea (AM)

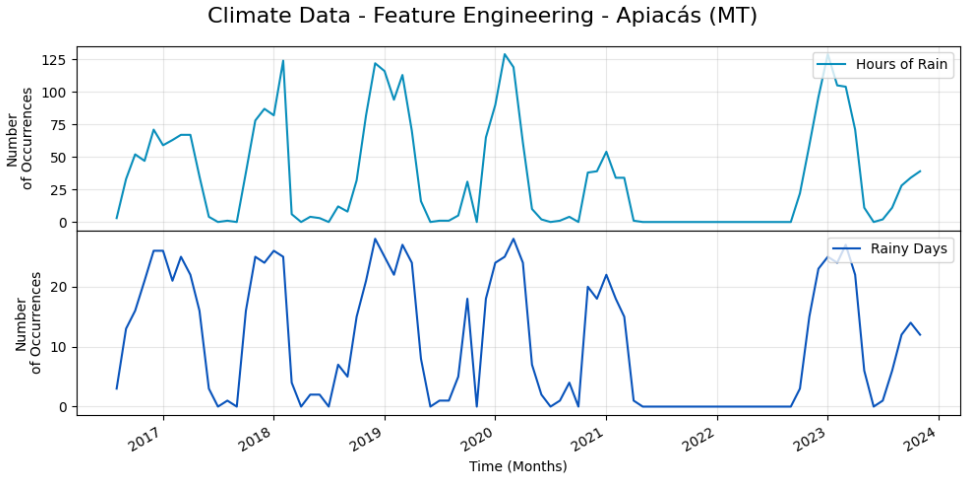


Figure 3.40: Climate data created by Feature Engineering per month from the meteorological station in the municipality of Apiacás (MT)

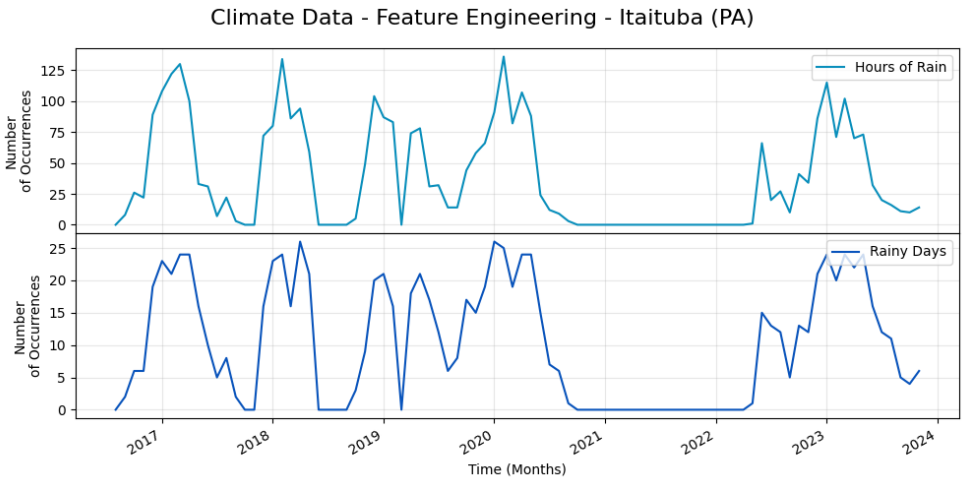


Figure 3.41: Climate data created by Feature Engineering per month from the meteorological station in the municipality of Itaituba (PA)

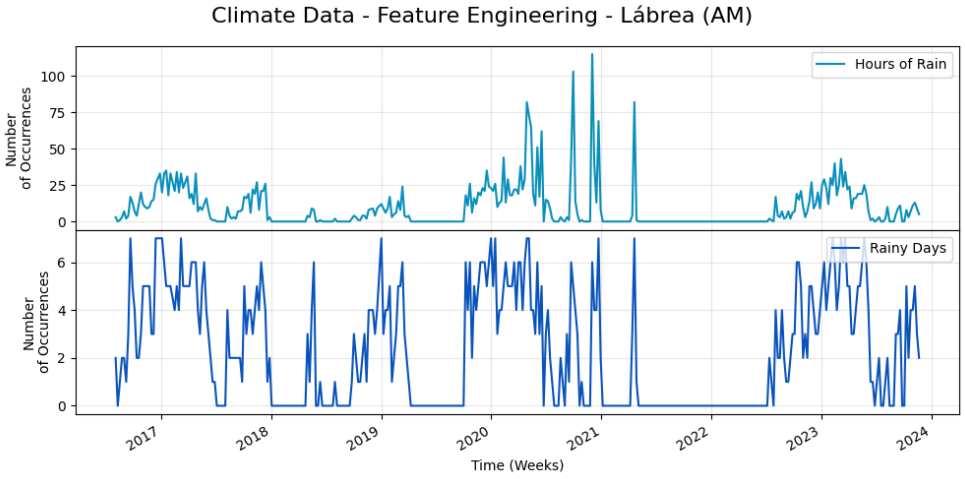


Figure 3.42: Climate data created by Feature Engineering per week from the meteorological station in the municipality of Lábrea (AM)

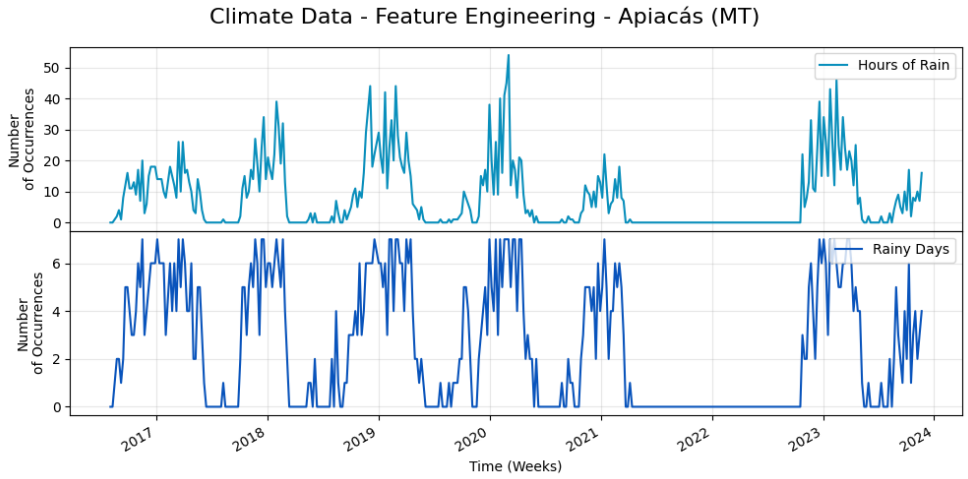


Figure 3.43: Climate data created by Feature Engineering per week from the meteorological station in the municipality of Apiacás (MT)

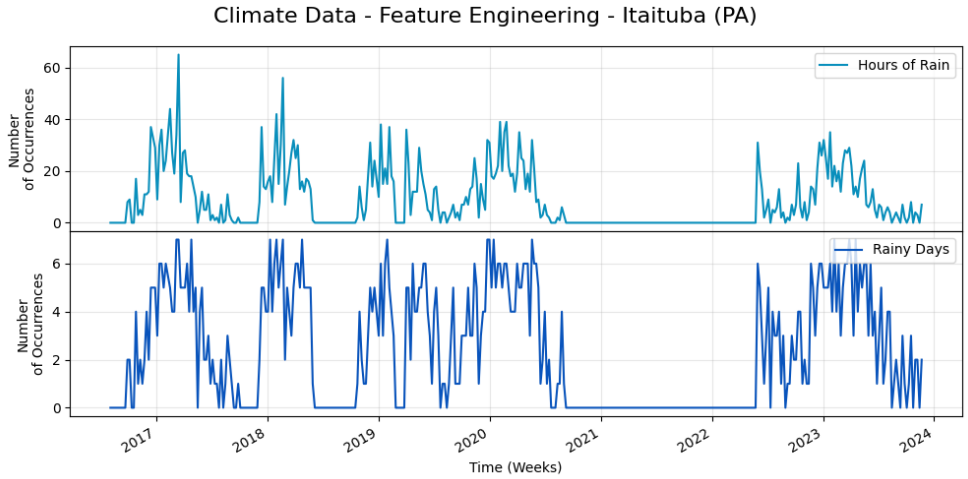


Figure 3.44: Climate data created by Feature Engineering per week from the meteorological station in the municipality of Itaituba (PA)

To study the relationships between the variables, the correlation was calculated concerning the previously shown series of climate and deforestation data according to lags from 1 to 12 for monthly data and with lags from 1 to 52 for weekly data. The study of correlations will be necessary to better assist in the choice of variables for the assembly of future models. Figure 3.45 represents the correlation of each municipality's deforestation series with the last 12 months lags, while Figure 3.46 represents the correlation of each municipality's deforestation series with its last 52 weekly lags. As expected, it is possible to see a positive correlation in the lags near the ends for both figures and a negative correlation for the lags closer to the middle of the lag range considered, reinforcing the data seen in the deforestation series in Figure 3.31.

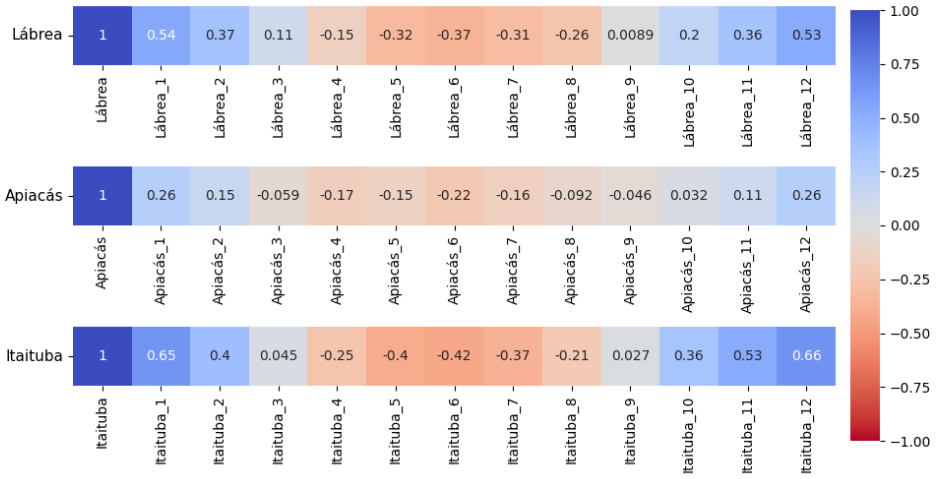


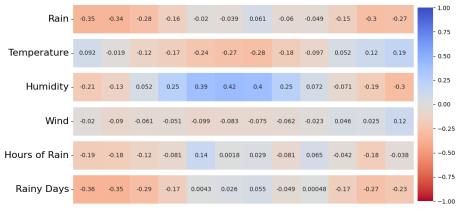
Figure 3.45: Correlations between the past times of the series, in months, and the time of interest

Figure 3.46: Correlations between the past times of the series, in weeks, and the time of interest

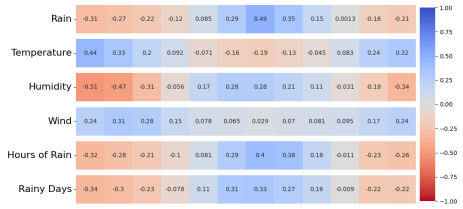
After studying the correlation of each municipality's deforestation series with its previous time steps, the calculation of the correlation between municipalities can also be calculated. As the Legal Amazon has more than 700 municipalities, graphs of correlations for all municipalities will not be shown. However, the highest correlations in the monthly data for all selected municipalities, that is, Lábrea, Apiacás, and Itaituba, are linked to small lags in other municipalities, usually the first or second lag. For weekly data, higher correlations are also more present in a series of smaller lags from other municipalities, between 1 and 6 lags. Remembering that the correlation calculation was made with a series of lags from 1 to 12, in the monthly data, and from 1 to 52, in the weekly data.

The same correlation analysis can be done with climate variables. In Figure 3.47 the correlations of the municipalities' monthly deforestation data with the lag series of 1 to 12 months of climate variables are shown. The same

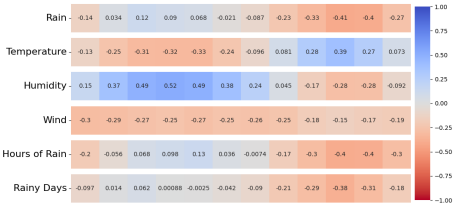
can be seen in Figure 3.48 with the correlations of the weekly deforestation series for each municipality with the lag series of 1 to 52 weeks of climate variables. For monthly data, the variables that show the highest correlation are humidity in the three municipalities, rain and rainy days in Lábrea, the temperature in Apiacás, and hours of rain in Itaituba. For weekly data, humidity also stands out with a high correlation for the three municipalities at different temporal lags, while other variables present a higher correlation, such as hours of rain, days of rain, and rain.



(a) Correlation with climate variables - Lábrea (AM)

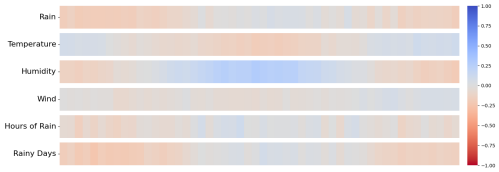


(b) Correlation with climate variables - Apiacás (MT)

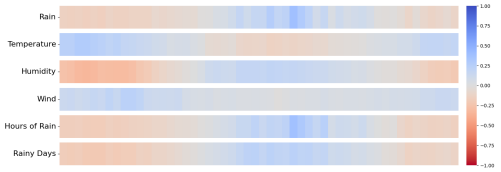


(c) Correlation with climate variables - Itaituba (PA)

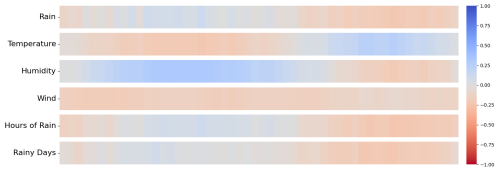
Figure 3.47: Correlations with climate variables - Months



(a) Correlation with climate variables - Lábrea (AM)



(b) Correlation with climate variables - Apiacás (MT)



(c) Correlation with climate variables - Itaituba (PA)

Figure 3.48: Correlations with climate variables - Weeks

4
Proposed Methodology

This chapter will examine how the chosen algorithms - autoregression, LightGBM, and LSTM - were applied to predict deforestation in the selected states and municipalities. Furthermore, it will be shown how the optimization of the hyperparameters chosen for these algorithms was carried out.

Through the analyses seen in Chapter 3 regarding periodicity, correlation with exogenous variables, autocorrelation, and possible deforestation trends in recent years, the creation of models that take these factors into account is necessary. Therefore, in addition to considering the algorithms and models already mentioned, using different methods has proven useful so that more comparisons and analyses can be carried out regarding the modeling of state and municipal data. Each model was applied to each algorithm so that nine models were assembled for every region considered, state, or municipality.

Initially, only models related to monthly data were applied. However, if necessary, weekly models were also constructed to improve the results obtained by the monthly models. Therefore, the processing of monthly and weekly data will be explained below. Then, the methods are applied to the data and algorithms, and finally, the hyperparameter tuning and the results of these optimizations are applied to the adopted algorithms.

It is worth noting that for each model training, validation, and test sets were considered, representing 60%, 20%, and 20% of the data respectively. These sets were defined sequentially, that is, the data referring to the training set occurred first, then the data from the validation set, and lastly the data from the test set. Furthermore, Min-Max normalization was applied in cases where LSTM was used; it was not necessary to use normalization in models that used LightGBM since we were dealing with trees in this case and regression typically doesn't request normalization. The normalization was done according to the values of the training set, that is, the values of the test and validation sets were normalized based on the minimum and maximum of the training set. To evaluate the results, the data considered from the test set was denormalized.

The processes set out below were applied to each chosen region. The need to create specific models for each state and municipality proved necessary due to the difference in the behavior of the deforestation series in each location. As interesting as creating a general model is, its results end up removing important properties from each series, such as deforestation peaks and possible trends. Therefore, the creation of individual models is important given the objective

of capturing and modeling the particularities of each data sequence.

4.1 Data Organization

The first part of the proposed methodology is based on the organization of model inputs and outputs. For each data granulation, weekly and monthly, a type of treatment will be proposed. The main purpose is to separate multiple outputs for each model so that it is possible to make predictions for various times in the future. Predicting several steps is a way of helping to prevent deforestation, as it enables advanced planning and more informed decision-making.

4.1.1 Monthly Data

For models that used monthly data, predictions will be made for the next two time steps, that is, two months. As much as predicting more temporal steps is interesting, the amount of data does not allow for a large number of steps to be predicted. Therefore, a two-time step analysis provides a reasonable initial perspective. Furthermore, it is important to mention that naive models were calculated to make predictions. The assembling consisted of repeating the last observation available for the next two steps.

The inputs are based on past values of the variables chosen according to the analysis of their relationships. For each method, variables will be chosen and assigned as model inputs. Figure 4.1 indicates an example of how the data can be considered, taking into account the deforestation series in the region and other variables that have temporal lags related to deforestation in the region in question. It is worth mentioning that the figure shows a general case, so it does not completely reflect what was adopted in all the methods exposed.

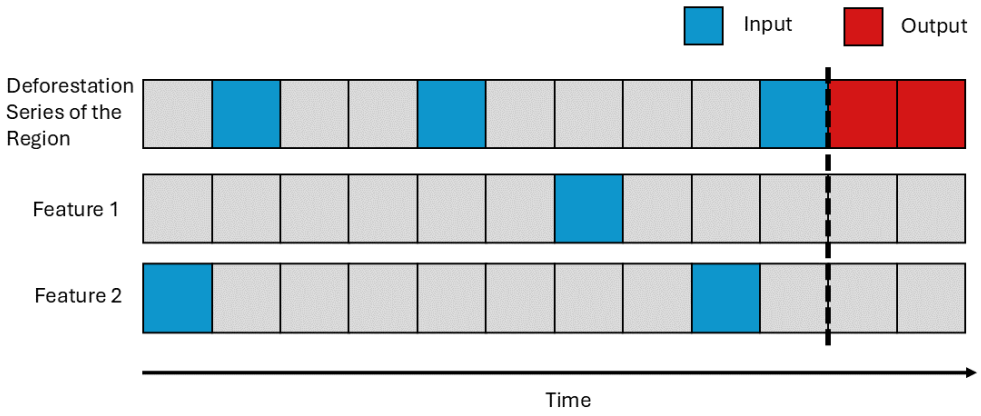


Figure 4.1: Example of model organization using monthly data.

4.1.2 Weekly Data

For models that consider weekly data, the objective will also be to predict several temporal steps. However, the output values will be established by the sum of four time steps in a row, this will be repeated twice. In this way, the weekly models will predict two groups of four consecutive time steps added together, resulting in approximately two months of forecast, just like the monthly models. The purpose of carrying out this process and not predicting and calculating the metrics of interest with each step individually is to make the predictions more robust. The weekly deforestation series in each region is very complex to make predictions, mainly due to the large fluctuations from one step to another combined with the low amount of data. Naive models were also considered concerning the weekly data contained in the repetition of the last observation available for the predicted steps, followed by the sum of the values in groups of four.

The inputs will be based on past values of the variables chosen for each method and algorithm. Figure 4.2 shows an example of how data can be considered in organizing inputs and outputs of models that group variables by week. The example considers the deforestation series of a region and other series that showed certain relationships with the variable of interest.

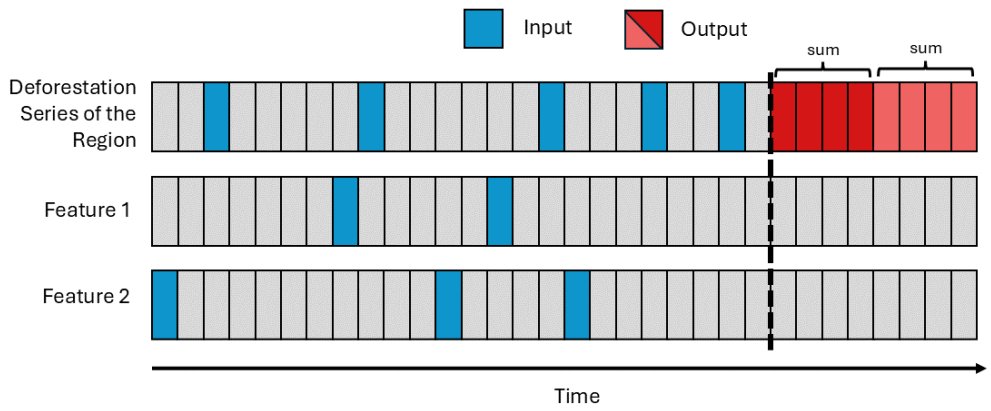


Figure 4.2: Example of model organization using weekly data.

4.2 Methodology

From the treated and processed data, three methods for applying the algorithms were devised. Therefore, they will be presented below according to what was performed for each of them.

4.2.1 Method 1

The first method is based on the sensitivity test about the lags of the deforestation series in each region. No other variables were considered, only past values of the series of interest. Therefore, tests were carried out to examine how many consecutive time steps the model would result in better metrics. Models were tested that considered 1 to 12 consecutive time steps, for monthly cases, and 1 to 53 consecutive time steps, for weekly cases. The choice of numbers 12 and 53 for the models proved interesting due to the analysis of the deforestation series in each region since it is possible to see an annual periodicity through them. Therefore, the number of lags taken into account in this method specifically also enters into hyperparameter tuning.

Even though the 12-month or 53-week windows are large for the amount of data available, the analysis of the metrics and graphs will indicate whether the choices made by the models will have a major impact or not. Furthermore, considering large windows does not imply the final choice of a model that considers large windows, as there are other models to be tested and graphics to be analyzed. Figure 4.3 illustrates the organization of the steps involved in implementing the method.

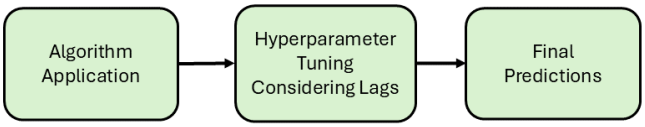


Figure 4.3: General structure of the first proposed method.

4.2.2 Method 2

To add variables that can better explain the sequence of interest, correlation was considered. For monthly data, variables with a correlation above 0.4 in absolute value were considered, while variables with a correlation above 0.3 were considered for weekly data. If the number of variables exceeded 10, for monthly data, and 20, for weekly data, only the 10 or 20 most correlated variables would be considered.

Only the correlations of the last 6 and 26 lags of each variable were considered for the monthly and weekly models respectively. Furthermore, the correlations considered were related to climate variables and deforestation variables from other regions, in the case of states, all states were considered and in the case of municipalities, all municipalities were considered. Thus, Figure 4.4 represents the organization of the exposed method.

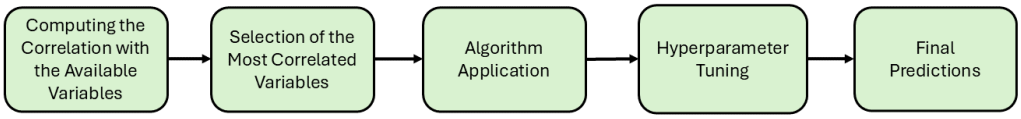


Figure 4.4: General structure of the second proposed method.

Essentially, this method was applied in the same way to all algorithms, however, the inclusion of two variables was made for models involving LightGBM, in which two temporal variables were added. The first indicates the month or week of the year and the other identifies the season of the year in which the period is located, being 0 summer, 1 autumn, 2 winter, and 3 spring. This proves useful because LightGBM does not necessarily identify the temporal moment in which the step is located, unlike the other models applied, LSTM and autoregressive.

4.2.3 Method 3

The last method considered consists of a modification of the previous method. The correlations of the last 6 and 26 lags will also be considered for the monthly and weekly models, respectively, concerning climate and deforestation variables. However, in addition to calculating the correlation, principal component analysis (PCA) will be applied to select the variables that best explain the space of observations considered.

Therefore, after selecting variables with a correlation above 0.3 for weekly models and 0.4 for monthly models, in absolute value, PCA will be applied to select the variables of interest. Firstly, the graph of explained variance by each component of the PCA must be examined, through which the minimum number of components considered for the models will be chosen. In the case of states, the minimum number of explained variance will be 95%, while in the case of municipalities, it will be 80%. Then, the variables that contribute most to each component are separated and these will be used to make the predictions.

The organization of the method can be examined in Figure 4.5. It can be seen that after applying the PCA, three parts must be executed. The first would be to calculate the variance explained by each component, the goal is to sum these percentages to be close to the previously defined values. Then, after deciding on the number of components, it is possible to select the variable that had the most impact on each component and separate it for future application in the model. This way, hyperparameter tuning can be performed and final predictions made.

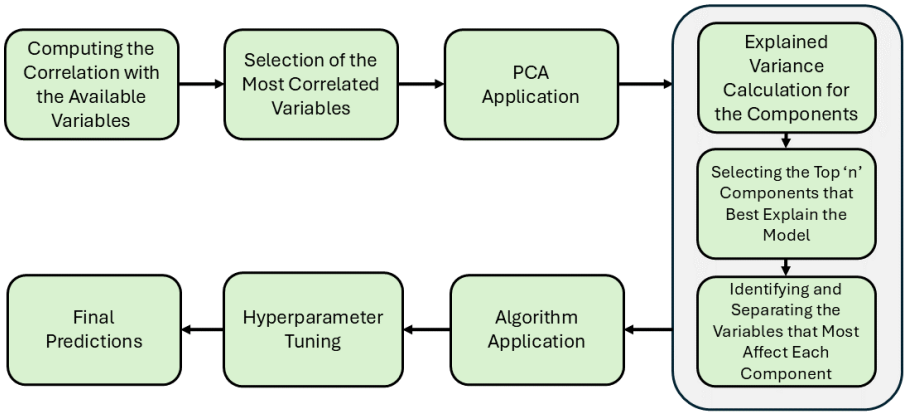


Figure 4.5: General structure of the third proposed method.

As in the second method, the inclusion of two variables was made for the

models that used LightGBM. The variables aim to provide a temporal location for the model, as one transmits information about the month or week in which the period is situated, while the other indicates the season of the year for the period. These variables were not added to the models that used autoregression or LSTM, since the algorithms themselves have their ways of dealing with the temporal position of the data.

4.3 Hyperparameter Tuning

The way the models and algorithms were applied remained the same for all the methods shown. However, in each one of the algorithms, a tuning of its hyperparameters was applied. This implies more optimized models according to the chosen metrics. This way, the hyperparameters considered for each model will be shown, in addition to the respective results of the tuning performed.

4.3.1 Autoregression

For models that used regression, two hyperparameters were used: *period* and *trend*. The first indicates the number of binary variables to be inserted in the model, each variable refers to a time step of the period, indicating 1 if it is in the selected period and 0 otherwise. The second hyperparameter concerns the data trend, which can be without trend ('n'), constant trend ('c'), temporal trend ('t'), and constant and temporal trend ('ct'). The hyperparameters considered for the monthly and weekly models can be examined in Table 4.1. Furthermore, the selection of hyperparameters was done through a grid search.

Hyperparameter	Monthly	Weekly
period	2 to 25	2 to 60
trend	'n'; 'c'; 't'; 'ct'	'n'; 'c'; 't'; 'ct'

Table 4.1: Hyperparameters considered for optimization of regressive models.

The hyperparameter optimization results for each method can be viewed in the Tables 4.2, 4.3, and 4.4 considering the data from the selected states and in Tables 4.5, 4.6, and 4.7 considering the data from the selected municipalities. In the tables it is possible to see two hyperparameter values for each monthly model, the first indicates the results for the prediction of the first step, while the second indicates the result related to the prediction of the second step. Meanwhile, there are eight values for the weekly models, symbolizing the eight weeks ahead that the model is predicting, each value relates to one week.

From this, forecasts are calculated for eight consecutive weeks. They will be organized into groups of 4 weeks so that metrics related to the aggregate values of the groups can later be computed. It can also be observed that there is no weekly model data in the selected municipalities or in the states of MT and PA. This is due to the efficiency of the monthly models considered for these regions, meaning that weekly model training is not necessary.

Method 1 - Autoregressive - State						
States	Monthly			Weekly		
	lags	*period*	*trend*	*lags*	*period*	*trend*
AM	2; 1	18; 13	't'; 'ct'	4;4;4;4; 46;46;46;46;	52;52;52;52; 52;52;50;50;	't';'t';'t';'ct'; 'ct';'ct';'t';'t'
MT	8; 1	8; 24	'ct'; 'c'	x	x	x
PA	10; 11	24; 24	'n'; 'n'	x	x	x
RO	8; 9	2; 5	'n'; 'c'	34;34;34;34; 29;29;29;29	27;27;27;27; 27;27;27;27	'c';'n';'n';'n'; 'n';'n';'n';'n'

Table 4.2: Hyperparameter Tuning - Method 1 - Autoregression - Weekly and Monthly - States. Each set of values (x; y; ...) in the table represents the tuning for each predicted time step.

Method 2 - Autoregressive - State				
States	Monthly		Weekly	
	period	*trend*	*period*	*trend*
AM	12; 13	'ct'; 't'	52; 55; 52; 52; 55; 55; 55; 55	'ct'; 't'; 'c'; 'c' 'c'; 'ct'; 't'; 't'
MT	8; 24	'c'; 'c'	x	x
PA	12; 18	't'; 'c'	x	x
RO	6; 12	'c'; 'n'	55; 5; 33; 55; 27; 27; 54; 27	'n'; 'n'; 'c'; 'n' 'n'; 'n'; 'n'; 'n'

Table 4.3: Hyperparameter Tuning - Method 2 - Autoregression - Weekly and Monthly - States. Each set of values (x; y; ...) in the table represents the tuning for each predicted time step.

Method 3 - Autoregressive - State				
States	Monthly		Weekly	
	period	*trend*	*period*	*trend*
AM	24; 13	't'; 't'	55; 52; 52; 55; 55; 55; 55; 55	'c'; 'c'; 't'; 't' 'c'; 'ct'; 't'; 'ct'
MT	8; 24	'c'; 'n'	x	x
PA	12; 18	'c'; 'c'	x	x
RO	6; 12	'c'; 'c'	55; 55; 55; 5; 27; 5; 55; 27	'n'; 'c'; 'c'; 'n'; 'c'; 'c'; 'n'; 'n'

Table 4.4: Hyperparameter Tuning - Method 3 - Autoregression - Weekly and Monthly - States. Each set of values (x; y; ...) in the table represents the tuning for each predicted time step.

Method 1 - Autoregressive - Municipal						
Municipalities	Monthly			Weekly		
	lags	*period*	*trend*	*lags*	*period*	*trend*
Lábrea	2; 2	12; 12	'n'; 'n'	x	x	x
Apiacás	12; 4	23; 23	'c'; 'c'	x	x	x
Itaituba	6; 2	12; 24	'n'; 'c'	x	x	x

Table 4.5: Hyperparameter Tuning - Method 1 - Autoregression - Weekly and Monthly - Municipalities. Each set of values (x; y) in the table represents the tuning for each predicted time step.

Method 2 - Autoregressive - Municipal				
Municipalities	Monthly		Weekly	
	period	*trend*	*period*	*trend*
Lábrea	12; 12	'n'; 'n'	x	x
Apiacás	13; 15	'ct'; 'ct	x	x
Itaituba	6; 6	'n'; 'c'	x	x

Table 4.6: Hyperparameter Tuning - Method 2 - Autoregression - Weekly and Monthly - Municipalities. Each set of values (x; y) in the table represents the tuning for each predicted time step.

Method 3 - Autoregressive - Municipal				
Municipalities	Monthly		Weekly	
	period	*trend*	*period*	*trend*
Lábrea	13; 18	'n'; 'c'	x	x
Apiacás	12; 5	'c'; 'c'	x	x
Itaituba	18; 12	'c'; 'c'	x	x

Table 4.7: Hyperparameter Tuning - Method 3 - Autoregression - Weekly and Monthly - Municipalities. Each set of values (x; y) in the table represents the tuning for each predicted time step.

4.3.2 LightGBM

The hyperparameter optimization of applications that used LightGBM accounted for 6 hyperparameters that can be seen through Table 4.8. This way, a brief discussion about the function of each hyperparameter can be made. The intention of defining a maximum height of the tree can be determined through *max_depth*, the value -1 indicates that the tree has no height limit. Defining this characteristic is important so the model does not suffer from overfitting. Other hyperparameters that have this same function are *min_child_samples* and *num_leaves*. While the first regulates the minimum number of samples that each leaf needs to have to be able to divide, the second indicates the maximum number of leaves that the tree will possess. Now, about *learning_rate* and *n_iterations*, the model's learning rate and the number of iterations that will be performed to create the tree are determined, respectively. Finally, *feature_fraction* is the variable that determines the number of features that will be disregarded when performing GOSS. It is worth mentioning that the hyperparameter tuning method that was used for this algorithm was grid search.

Hyperparameter	Monthly	Weekly
max_depth	-1; 5; 10	-1; 10; 20
learning_rate	0.05; 0.1; 0.15	0.05; 0.1; 0.15
min_child_samples	3; 5	3; 5; 10
num_leaves	7; 15; 31	7; 15; 31
n_iterations	100; 500	100; 500
feature_fraction	0.8; 0.9; 1	0.9; 0.95; 1

Table 4.8: Hyperparameters considered for LightGBM optimization.

The results that were obtained through the optimization of the hyperparameters can be inspected in Tables 4.9, 4.10, 4.11, 4.12, 4.13, and 4.14 for state models and in Tables 4.15, 4.16, and 4.17 for municipal models. The weekly models from municipalities and some states were not represented, as they were not necessary to make the forecasts, meaning the monthly models were sufficient.

Method 1 - LightGBM - Monthly - State							
States	*lags*	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
AM	11	-1	0.1	3	7	100	0.8
MT	8	5	0.1	5	15	100	0.8
PA	12	5	0.05	3	31	100	0.8
RO	3	5	0.05	5	15	100	0.8

Table 4.9: Hyperparameter Tuning - Method 1 - LightGBM - Monthly - States.

Method 2 - LightGBM - Monthly - State						
States	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
AM	-1	0.05	5	15	100	0.8
MT	5	0.05	5	15	100	0.9
PA	-1	0.05	5	7	100	0.9
RO	-1	0.05	5	7	100	0.8

Table 4.10: Hyperparameter Tuning - Method 2 - LightGBM - Monthly - States.

Method 3 - LightGBM - Monthly - State						
States	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
AM	5	0.05	5	15	100	0.8
MT	-1	0.05	3	7	100	0.8
PA	-1	0.05	3	7	100	0.8
RO	-1	0.05	5	7	100	0.8

Table 4.11: Hyperparameter Tuning - Method 3 - LightGBM - Monthly - States.

Method 1 - LightGBM - Weekly - State						
States	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
AM	-1	0.05	10	7	100	0.9
MT	x	x	x	x	x	x
PA	x	x	x	x	x	x
RO	-1	0.05	10	7	100	1

Table 4.12: Hyperparameter Tuning - Method 1 - LightGBM - Weekly - States.

Method 2 - LightGBM - Weekly - State						
States	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
AM	-1	0.05	10	7	100	0.9
MT	x	x	x	x	x	x
PA	x	x	x	x	x	x
RO	-1	0.05	5	7	100	0.9

Table 4.13: Hyperparameter Tuning - Method 2 - LightGBM - Weekly - States.

Method 3 - LightGBM - Weekly - State						
States	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
AM	-1	0.05	10	7	100	0.95
MT	x	x	x	x	x	x
PA	x	x	x	x	x	x
RO	-1	0.05	5	7	100	0.9

Table 4.14: Hyperparameter Tuning - Method 3 - LightGBM - Weekly - States.

Method 1 - LightGBM - Monthly - Municipal							
Municipalities	*lags*	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
Lábrea	12	-1	0.05	5	7	100	0.8
Apiacás	4	-1	0.05	5	7	100	0.8
Itaituba	10	-1	0.05	5	7	100	0.8

Table 4.15: Hyperparameter Tuning - Method 1 - LightGBM - Monthly - Municipalities.

Method 2 - LightGBM - Monthly - Municipal						
Municipalities	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
Lábrea	5	0.05	3	7	100	0.8
Apiacás	-1	0.05	3	7	100	1
Itaituba	-1	0.05	5	15	100	0.8

Table 4.16: Hyperparameter Tuning - Method 2 - LightGBM - Monthly - Municipalities.

Method 3 - LightGBM - Monthly - Municipal						
Municipalities	*max_ depth*	*learning_ rate*	*min_child_ samples*	*num_ leaves*	*n_ iterations*	*feature_ fraction*
Lábrea	10	0.1	3	31	100	0.9
Apiacás	5	0.05	3	7	100	0.9
Itaituba	-1	0.15	3	7	100	0.8

Table 4.17: Hyperparameter Tuning - Method 3 - LightGBM - Monthly - Municipalities.

4.3.3 LSTM

LSTM networks were optimized for the number of layers (*layers*), the number of neurons for each layer (*units*), the percentage of connections that would be disconnected from one layer to another (*dropout_rate*) and the learning rate used by the chosen optimizer (*learning_rate*). All values considered for each hyperparameter can be examined in Table 4.18. It is worth mentioning that when the number of layers considered is less than the maximum number, 5, the hyperparameters related to layers that weren't classified will not be taken into the tuning process and will be represented by '-' in the following tables. Furthermore, it is worth pointing out that the optimizer chosen for the LSTM models was Adam (Adaptive Moment Estimator) (51), which adjusts the learning rates individually for each parameter based on estimates of the first and second moments of the gradients, allowing for smoother parameter updates. Finally, as the number of hyperparameters was potentially vast, a random search was applied, exploring 50 combinations for each set of lags in method 1 and 200 combinations for methods 2 and 3.

Hyperparameter	Monthly	Weekly
layers	2 a 5	2 a 5
units_1	5; 20; 35; 50	5; 20; 35; 50
units_2	5; 20; 35; 50	5; 20; 35; 50
units_3	5; 20; 35; 50	5; 20; 35; 50
units_4	5; 20; 35; 50	5; 20; 35; 50
units_5	5; 20; 35; 50	5; 20; 35; 50
dropout_rate_1	0; 0.5; 0.1	0; 0.5; 0.1
dropout_rate_2	0; 0.5; 0.1	0; 0.5; 0.1
dropout_rate_3	0; 0.5; 0.1	0; 0.5; 0.1
dropout_rate_4	0; 0.5; 0.1	0; 0.5; 0.1
dropout_rate_5	0; 0.5; 0.1	0; 0.5; 0.1
learning_rate	0.01; 0.001; 0.0001	0.01; 0.001; 0.0001

Table 4.18: Hyperparameters considered for optimization of LSTM networks.

Now, it is possible to observe the results of the hyperparameter tuning for each model that used LSTM through the Tables 4.19, 4.20, 4.21, 4.22, 4.23, and 4.24 for the state models and Tables 4.25, 4.26, and 4.27 for the municipal models. Each table has the names of the simplified hyperparameters, *'u'* for *'units_'*, *'dr'* for *'dropout_rate_'*, and *'lr'* for *'learning_rate'*. Furthermore, Figure 4.6 indicates a general representation of the LSTM network architecture for better understanding.

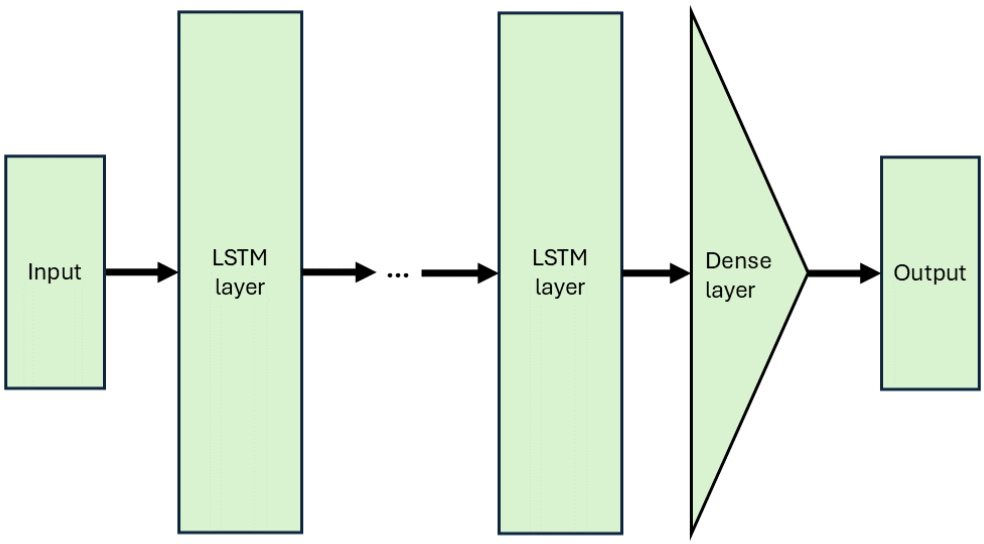


Figure 4.6: General architecture of the LSTM networks used.

Method 1 - LSTM - Monthly - State													
States	lags	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
AM	1	4	20	50	20	35	-	0.05	0.1	0.1	0.1	-	0.01
MT	12	4	5	5	5	35	-	0.05	0	0	0	-	0.0001
PA	7	4	35	50	20	35	-	0.1	0	0	0.05	-	0.001
RO	8	5	35	20	30	35	35	0	0.05	0.1	0	0.05	0.01

Table 4.19: Hyperparameter Tuning - Method 1 - LSTM - Monthly - States.

Method 2 - LSTM - Monthly - State												
States	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
AM	3	35	20	20	-	-	0	0	0.05	-	-	0.01
MT	3	35	50	5	-	-	0	0	0.1	-	-	0.01
PA	2	50	35	-	-	-	0.1	0.1	-	-	-	0.01
RO	4	20	50	50	20	-	0.1	0.05	0	0.05	-	0.01

Table 4.20: Hyperparameter Tuning - Method 2 - LSTM - Monthly - States.

Method 3 - LSTM - Monthly - State												
States	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
AM	2	50	35	-	-	-	0.1	0.1	-	-	-	0.01
MT	2	50	5	-	-	-	0.05	0	-	-	-	0.01
PA	2	20	50	-	-	-	0.1	0	-	-	-	0.01
RO	3	35	35	35	-	-	0.1	0	0	-	-	0.01

Table 4.21: Hyperparameter Tuning - Method 3 - LSTM - Monthly - States.

Method 1 - LSTM - Weekly - State													
States	lags	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
AM	3	4	20	35	20	35	-	0	0.05	0.1	0	-	0.01
MT	x	x	x	x	x	x	x	x	x	x	x	x	x
PA	x	x	x	x	x	x	x	x	x	x	x	x	x
RO	19	3	35	20	20	-	-	0	0	0.05	-	-	0.01

Table 4.22: Hyperparameter Tuning - Method 1 - LSTM - Weekly - States.

Method 2 - LSTM - Weekly - State												
States	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
AM	4	20	35	20	35	-	0	0.05	0.1	0	-	0.01
MT	x	x	x	x	x	x	x	x	x	x	x	x
PA	x	x	x	x	x	x	x	x	x	x	x	x
RO	4	20	20	50	5	-	0.1	0.05	0.05	0.05	-	0.01

Table 4.23: Hyperparameter Tuning - Method 2 - LSTM - Weekly - States.

Method 3 - LSTM - Weekly - State												
States	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
AM	2	50	5	-	-	-	0.05	0	-	-	-	0.01
MT	x	x	x	x	x	x	x	x	x	x	x	x
PA	x	x	x	x	x	x	x	x	x	x	x	x
RO	2	50	35	-	-	-	0	0	-	-	-	0.01

Table 4.24: Hyperparameter Tuning - Method 3 - LSTM - Weekly - States.

Method 1 - LSTM - Monthly - Municipal													
Municipalities	lags	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
Lábrea	7	5	35	5	50	20	50	0.1	0	0.1	0	0.1	0.01
Apiacás	6	3	20	50	5	-	-	0.05	0.1	0.05	-	-	0.01
Itaituba	3	5	5	50	50	35	35	0	0.05	0	0.1	0	0.01

Table 4.25: Hyperparameter Tuning - Method 1 - LSTM - Monthly - Municipalities.

Method 2 - LSTM - Monthly - Municipal												
Municipalities	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
Lábrea	3	20	50	5	-	-	0	0.05	0.05	-	-	0.01
Apiacás	4	5	50	35	5	-	0.05	0	0.05	0.1	-	0.01
Itaituba	2	35	5	-	-	-	0	0.05	-	-	-	0.01

Table 4.26: Hyperparameter Tuning - Method 2 - LSTM - Monthly - Municipalities.

Method 3 - LSTM - Monthly - Municipal												
Municipalities	layers	u1	u2	u3	u4	u5	dr1	dr2	dr3	dr4	dr5	lr
Lábrea	4	35	20	35	50	-	0.05	0.1	0	0	-	0.001
Apiacás	3	5	5	5	-	-	0	0.1	0.05	-	-	0.01
Itaituba	5	5	50	20	5	50	0.1	0.05	0	0	0.05	0.01

Table 4.27: Hyperparameter Tuning - Method 3 - LSTM - Monthly - Municipalities.

5
Results

Using the models presented in Chapter 4 it is possible to analyze the results found. In this way, for each selected state and municipality, a study of all applied models will be carried out. If it is necessary to apply weekly aggregated models, their evaluation will be carried out afterward. The comparison with values found in Tables 3.2 and 3.3 were made with the purpose of explaining the use of the models. Finally, a brief discussion of the results obtained will be conducted.

Due to the number of models tested for each location, graphs of all models will not be presented; only those that obtained the best RMSE concerning each algorithm that will be presented. Through this selection, the choice of the best generated model will be given by comparing the graphs and opinions of some of the metrics used. The final model will be detailed so that more observations and criticisms can be made. The intention of choosing RMSE as the initial metric for evaluating the models was due to the fact that it gave more weight to the model's largest errors. This is important since models that cannot correctly predict peak points should be penalized more, as the main objective is to be able to identify and study periods of higher deforestation rates.

It is important to say that although the best models were selected based on their RMSE values, other metrics were also considered for better conclusions. The mean absolute error (MAE) is one of them, functioning as a metric less sensitive to large deviations and having statistical stability, being more useful in cases where the series of interest does not have many outliers. Furthermore, classification metrics were applied to identify periods in which deforestation was above normal. For this purpose, the third quartile of each deforestation series was calculated, and predicted values above the third quartile were classified as 1 and below as 0. This way, the calculation of precision, recall, and F1-Score can be done to identify whether the models corrected the behavior of deforestation in the period and did not necessarily provide the exact value.

It is worth reinforcing the three methods that were applied in each algorithm. The first method uses a sensitivity test between the possible lags selected from the deforestation series in each location to predict their values. The second considers only the variables most correlated with the series of interest, including lags from the series itself, climate variables, and deforestation values from other states or municipalities. Finally, the third

method uses PCA on the most correlated variables to identify the features that add the most information to the space they define. Therefore, the results can be analyzed.

5.1 States Results

Initially, the analysis of the results of state models will be carried out considering the four previously selected states: Amazonas (AM), Mato Grosso (MT), Pará (PA), and Rondônia (RO). It is worth mentioning that as the objective will be to predict 2 steps forward in all models, monthly or weekly aggregated, the metrics will be calculated individually for each step. Then, the metrics can be averaged in order to evaluate the model as a unit. It is essential to analyze both the average and the individual values of each step since the objective is to obtain models that can effectively predict both the first and the second steps. Even though the metric gives a perspective of analyzing the model in general, the analysis of each step must also be done.

5.1.1 Amazonas

The first state to be examined will be AM, it is possible to analyze the comparison of the RMSE of the models through Figure 5.1. It is worth noting that the value obtained by the naive model is recorded in the figure legend. From these data, it is possible to see that the models in which LSTM and LightGBM were applied present higher RMSE, some of them even present worse metrics than the naive model adopted. The models in which autoregression was applied were the best, presenting values in the range of 242 to 271. The model with the lowest RMSE was the one that applied autoregression in method 2.

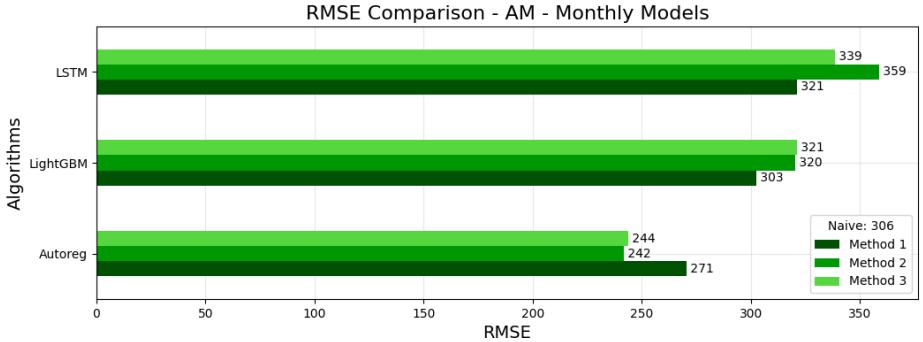


Figure 5.1: Comparison of the RMSE of each monthly model proposed for the state of Amazonas.

By analyzing the RMSE values, the graphs with the best metrics were separated so that a comparison could occur between the predictions made by each of them in addition to the other calculated metrics. Figure 5.2 shows the results of the selected models, establishing the visualization of the prediction graphs for each temporal step as well as the joint prediction of the predictions, that is, all the values that each window predicted. Meanwhile, Table 5.1 exposes the MAE, RMSE, and F1-Score values of each predicted step and the average of the steps. Through this, it is possible to observe that the autoregressive model presents better values of average RMSE and average F1-Score, while the model that uses LightGBM presents better MAE.

By observing the particular values of each predicted step, most of the best results are present in the autoregressive model. Furthermore, it can be seen that for the model that used LSTM, the second predicted step presented the value 0, symbolizing that there were no predictions greater than the determined threshold. Looking at the prediction graphs, it is noticeable that this model maintained lower predictions, taking into account that given that the series for the state of Amazonas presents several periods with low deforestation values, as seen in the image 3.8. From the analyses, the best model considered for the monthly AM forecast was the one that used the second method in the autoregressive model.

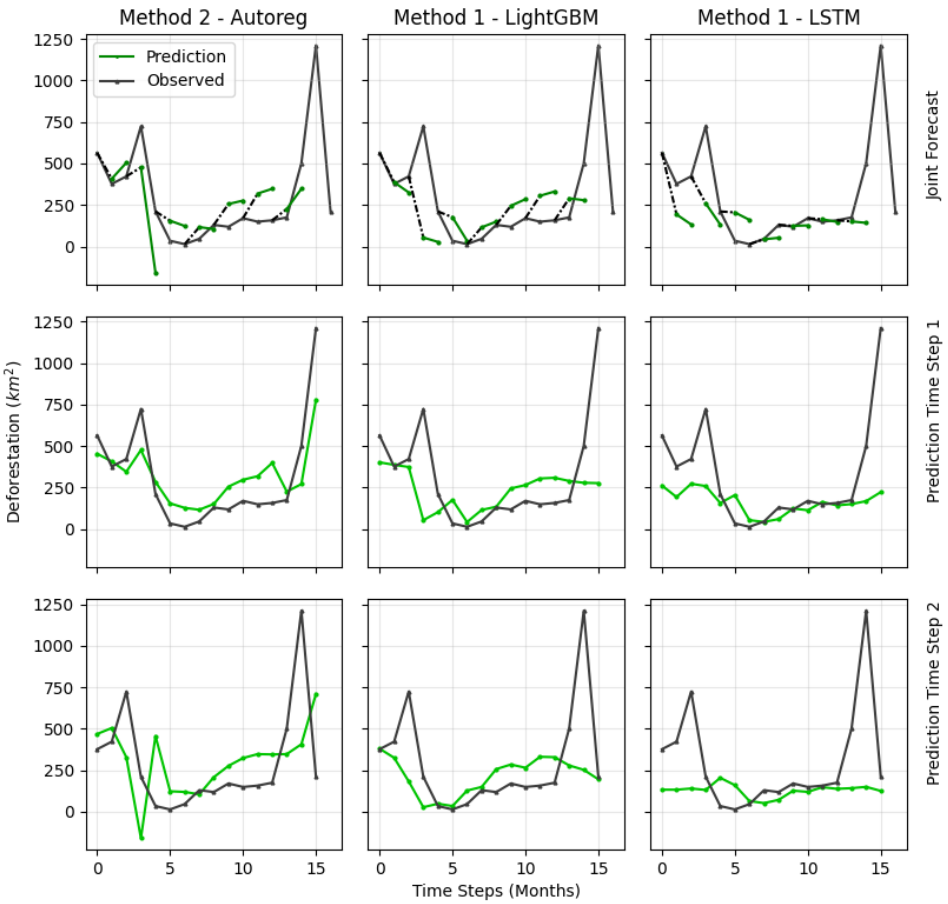


Figure 5.2: Best monthly models predictions for each algorithm used - Amazonas.

AM - Best Monthly Models

Step	2 - Autoregression			1 - LightGBM			1 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.667	141	174	0.625	190	308	0.800	179	305
2	0.625	235	310	0.571	178	297	0	205	337
Mean	0.646	188	242	0.598	184	303	0.400	192	321

Table 5.1: Comparative table of the best monthly models metrics by algorithm - AM

Figure 5.3 and Table 5.2 show the graph and the rest of the metrics related to the best monthly model chosen. This model cannot properly identify the peak moments of the series. However, even though the series maintains a periodic behavior, the data for the period in question were higher compared to recent years. However, it is important to say that the model managed

to perfectly map all the points that remained above the adopted threshold, obtaining a perfect recall. However, not all points predicted to be greater than the threshold are in the correct place, implying lower accuracy. In this way, models will be applied with aggregated weekly data in order to improve forecasts.

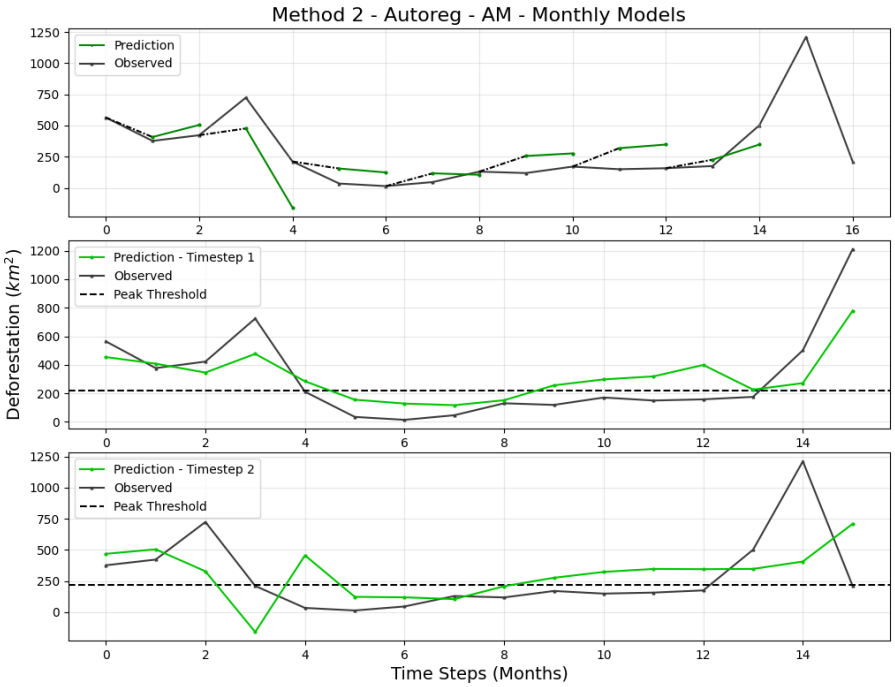


Figure 5.3: Best monthly model predictions - Amazonas.

AM - Metrics - Best Monthly Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	0.500	1	0.667	141	174
2	0.455	1	0.625	235	310
Mean	0.477	1	0.646	188	242

Table 5.2: Metrics table of the best monthly model by predicted time step - AM

The same process will be seen for the aggregated weekly models, the RMSE comparison can be found in Table 5.4. It is possible to observe behaviors similar to those found in the monthly models, that is, method 2 in the autoregressive models performed better, while method 1 obtained lower metrics in LSTM and LightGBM. Furthermore, all models had lower RMSE than the naive model. From these results, the separation and analysis of the best models of each algorithm can be made.

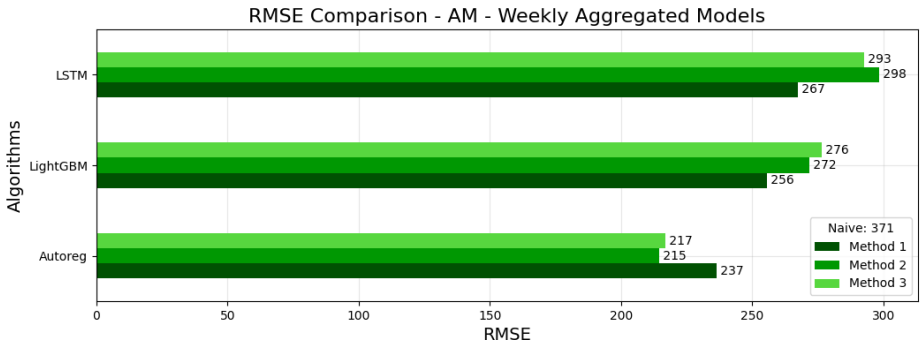


Figure 5.4: Comparison of the RMSE of each weekly aggregated model proposed for the state of Amazonas.

The biggest obstacle encountered in the monthly model was the fact that it was unable to predict peak points. However, from the application of weekly aggregated models, this problem is not completely resolved, as can be seen in Figure 5.5 with the comparison of the best models of each algorithm. Even so, it is noticeable that the moments of the highest deforestation values are better modeled by the autoregressive model. Furthermore, it presents a better classification metric when compared to the other two models through Table 5.3. The autoregressive model presents better results in all metrics for all steps in the table and can be considered the best model without dispute.

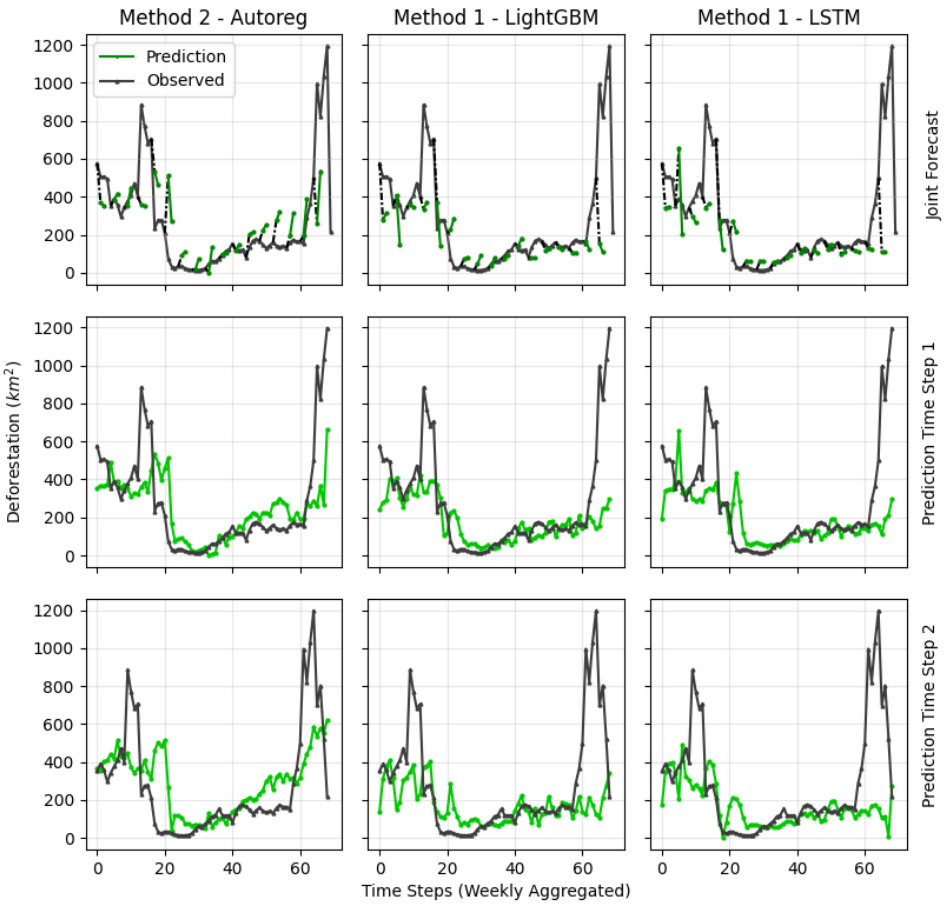


Figure 5.5: Best weekly aggregated models predictions for each algorithm used - Amazonas.

AM - Best Weekly Aggregated Models

Step	2 - Autoregression			1 - LightGBM			1 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.885	132	210	0.880	132	234	0.800	138	247
2	0.754	150	219	0.634	162	277	0.634	165	287
Mean	0.819	141	215	0.757	147	256	0.717	152	267

Table 5.3: Comparative table of the best weekly aggregated models metrics by algorithm - AM

The ultimate purpose of choosing the best weekly aggregated model is to check if it behaves better than the monthly model shown previously. Therefore, observing Figure 5.6 and Table 5.4 the most evident fact is that the classification of points above the peak threshold was higher. The recall result

remains perfect for all steps, just like in the monthly model, but the precision metrics show an increase, resulting in a better F1-Score.

It should be noted that the standard deviation of the aggregated weekly deforestation series is equal to 186, that is, smaller than the RMSE and larger than the MAE of the model. This is not ideal, but as RMSE penalizes larger errors more, this shows that the series presents problems when predicting outliers. Likewise, as the MAE is less sensitive to these errors, it ends up becoming smaller, indicating that the model makes relatively accurate predictions, while the RMSE suggests that there are predictions with errors much larger than the average.

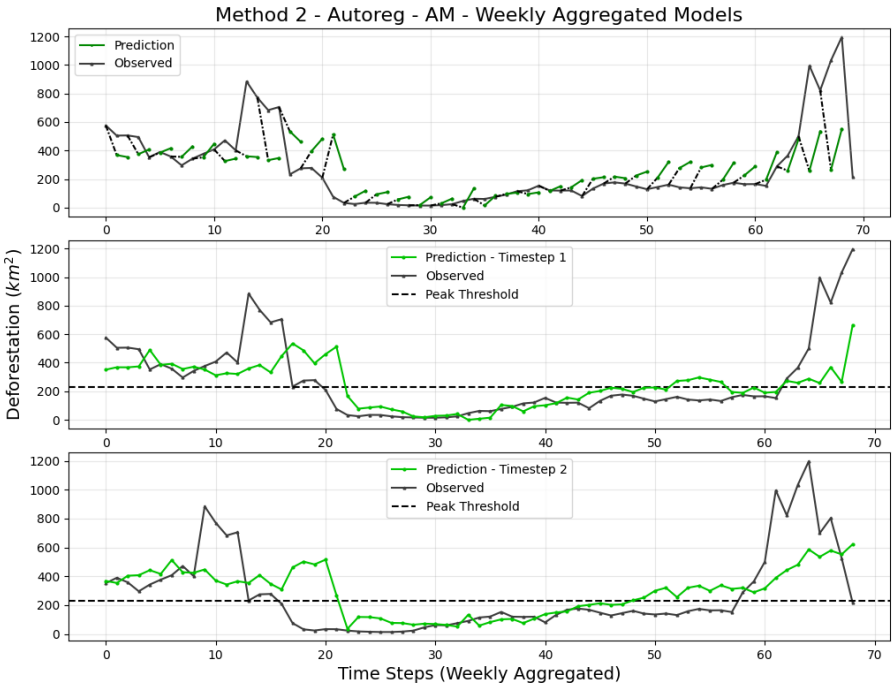


Figure 5.6: Best weekly aggregated model prediction - Amazonas.

AM - Metrics - Best Weekly Aggregated Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	0.794	1	0.885	132	210
2	0.605	1	0.754	150	219
Mean	0.699	1	0.819	141	215

Table 5.4: Metrics table of the best weekly aggregated model by predicted time step - AM

5.1.2 Mato Grosso

Moving on to the results for the state of Mato Grosso, the comparison of the RMSE of each model can be seen in Figure 5.7. The situation is similar to the state of Amazonas, since LSTM and LightGBM have worse results on average, while the autoregressive models show better results. Nevertheless, all models performed better than the base model concerning this metric, that is, they obtained RMSE lower than 1190. Method 3 performed best for models that used LSTM and autoregression, while method 2 performed best for models that used LightGBM.

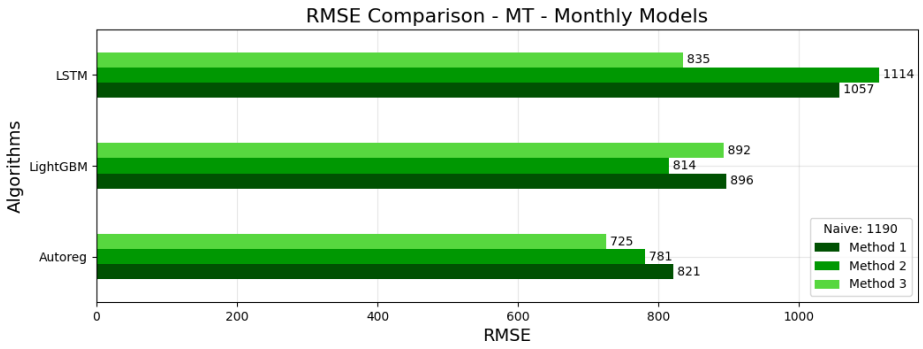


Figure 5.7: Comparison of the RMSE of each monthly model proposed for the state of Mato Grosso.

The best results can be examined in more detail using Figure 5.8 and Table 5.5 which present a comparison of the graphs and metrics of the best models of each algorithm. From the graphs, it is clear that the models were able to predict the magnitude of the peaks much more accurately than the monthly AM models. A model that draws the most attention is the autoregressive model, as it can predict moments of high deforestation rates more consistently and with smaller errors. The metrics displayed in the table reflect this behavior, in which the best results on average are present in the autoregressive model. Therefore, it was selected as the best model to predict monthly data for the state of Mato Grosso.

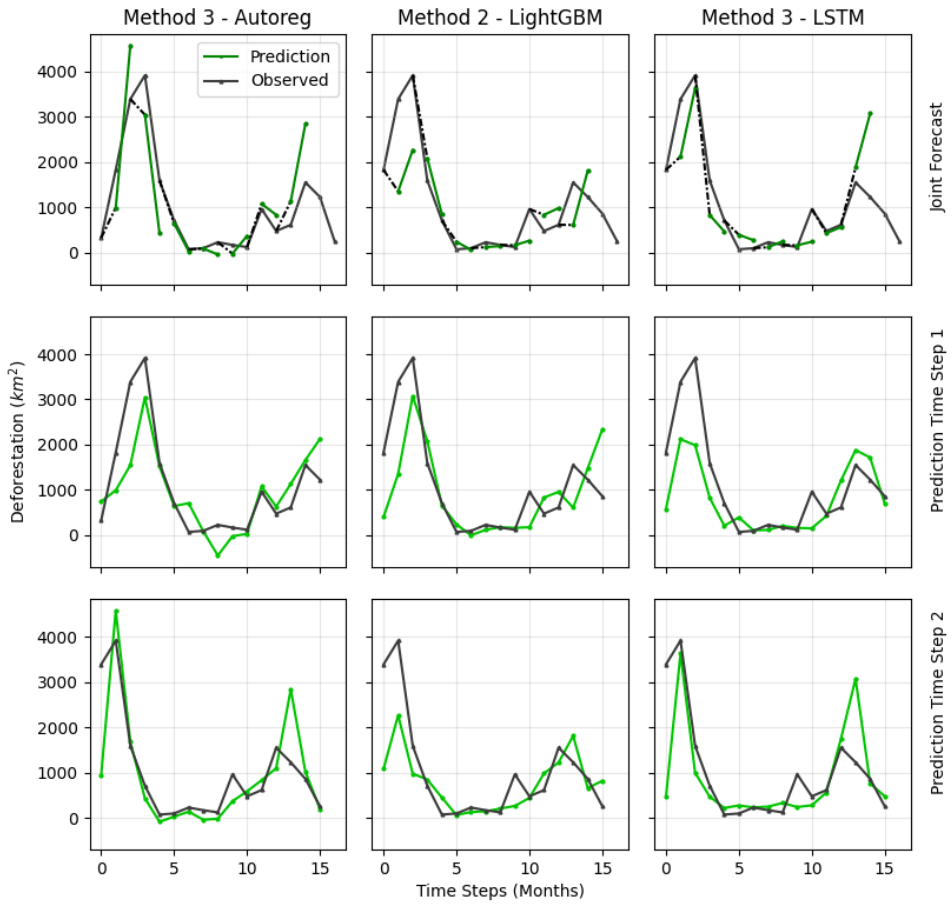


Figure 5.8: Best monthly models predictions for each algorithm used - Mato Grosso.

MT - Best Monthly Models

Step	3 - Autoregression			2 - LightGBM			3 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.909	470	665	0.727	589	836	0.727	542	760
2	0.750	460	786	0.750	507	793	0.750	501	911
Mean	0.830	465	725	0.739	548	814	0.739	521	835

Table 5.5: Comparative table of the best monthly models metrics by algorithm - MT

Therefore, the analysis in more detail of the best model chosen can be seen in Figure 5.9 and Table 5.6. They reinforce what has already been said in which forecasts consistently approach peak points. Furthermore, they present perfect precision for all predicted steps, that is, all elements predicted to be above the peak threshold are part of this class. This is a difficult

task considering that there is greater difficulty in predicting the model's outliers. Therefore, the higher the threshold adopted, the more challenging the classification of outliers becomes.

The prediction of the first time step is better for classification metrics and RMSE. However, the second step also presents good metrics and is not far from the first. Therefore, the application of weekly aggregated models was not necessary, indicating that the use of the third method applied to autoregression presented good results for predicting deforestation data for the state of MT.

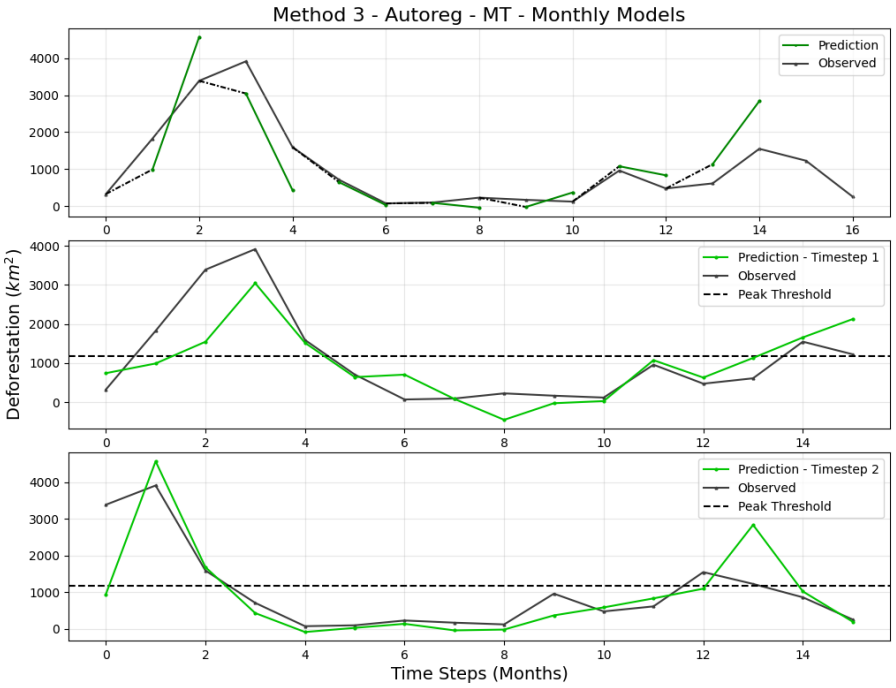


Figure 5.9: Best monthly model predictions - Mato Grosso.

MT - Metrics - Best Monthly Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	1	0.833	0.909	470	665
2	1	0.600	0.750	460	786
Mean	1	0.717	0.830	465	725

Table 5.6: Metrics table of the best monthly model by predicted time step - MT

5.1.3 Pará

The RMSE results of the models found for the state of Pará are presented in Table 5.10. Through this, it is notable that method 1 obtained the best

results in all algorithms, implying that each of them will be put aside for later comparison. Furthermore, all results, except for method 3 applied to LSTM, presented better metrics than the naive model.

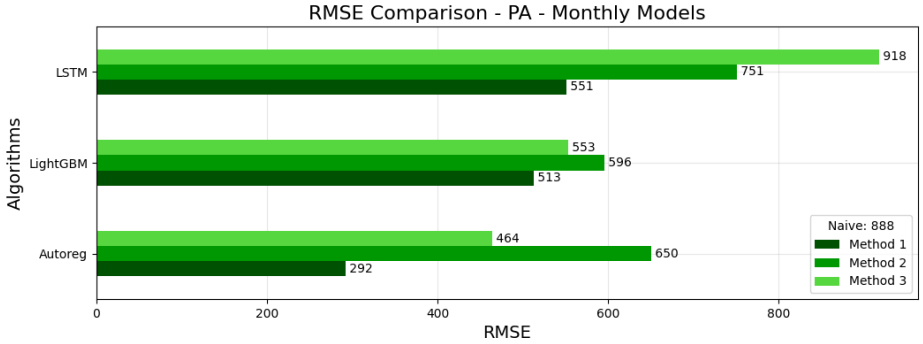


Figure 5.10: Comparison of the RMSE of each monthly model proposed for the state of Pará.

Based on Figure 5.11 the comparison of the best models of each algorithm can be carried out. At first, the predictions of the three models appear to identify well the deforestation patterns presented in PA. All of them can identify moments of high and low trends in the state, however, one model becomes more evident when the values of the metrics in Table 5.7 are observed, the autoregressive model. It is observed that on average the model presents the best metrics, in addition to presenting RMSE and MAE well below the standard deviation of the deforestation series, which is equal to 913. This indicates that predictions are more accurate than simple data variability, resulting in better predictions.

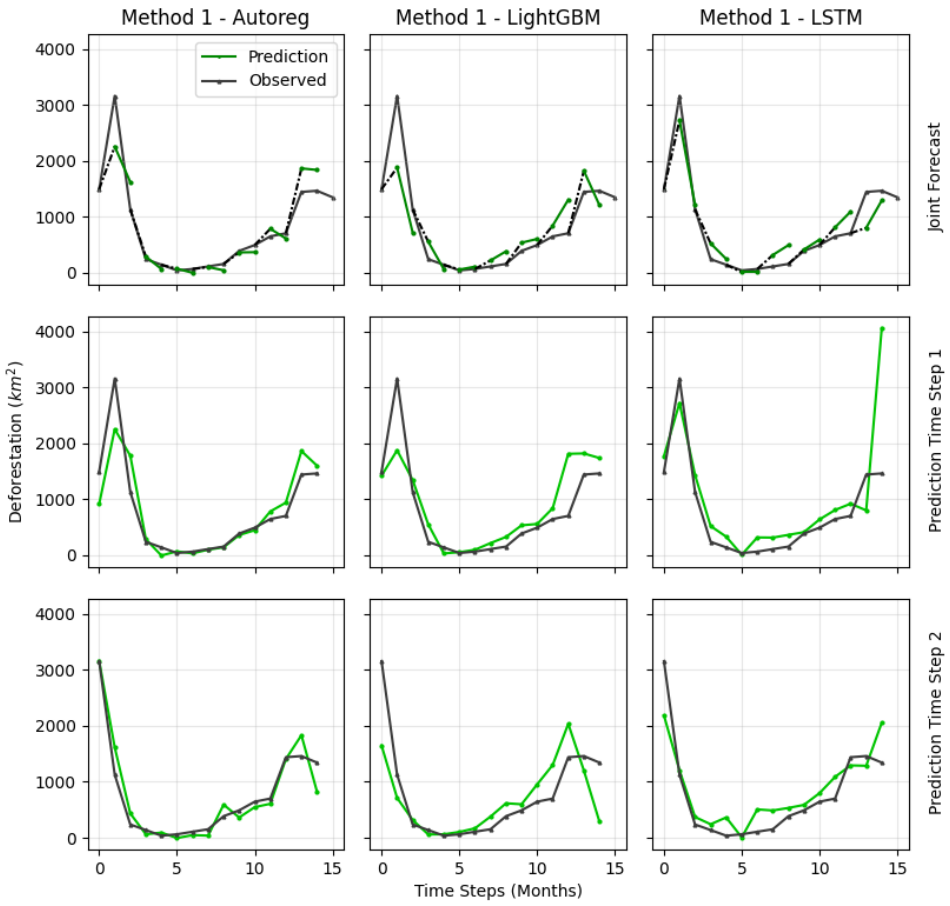


Figure 5.11: Best monthly models predictions for each algorithm used - Pará.

PA - Best Monthly Models

Step	1 - Autoregression			1 - LightGBM			1 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.909	228	292	0.909	299	474	0.800	399	726
2	0.889	167	230	0.600	373	552	0.909	281	376
Mean	0.899	198	292	0.755	336	513	0.855	340	551

Table 5.7: Comparative table of the best monthly models metrics by algorithm - PA

According to the comparison analysis, the graph of the best model can be seen in Figure 5.12, as well as its metrics in Table 5.8. An interesting fact that can be drawn from this data is that the predicted second step presents better accuracy, RMSE, and MAE than the first step. However, in general, the model can map the behavior of the series very well, as well as classify its

points according to the adopted threshold. Based on what has already been seen in other states, the forecast in Pará was the one that obtained the best F1-Score, in addition to being better able to predict peak values for the period in question. This fact is important since it helps to identify outliers more effectively.

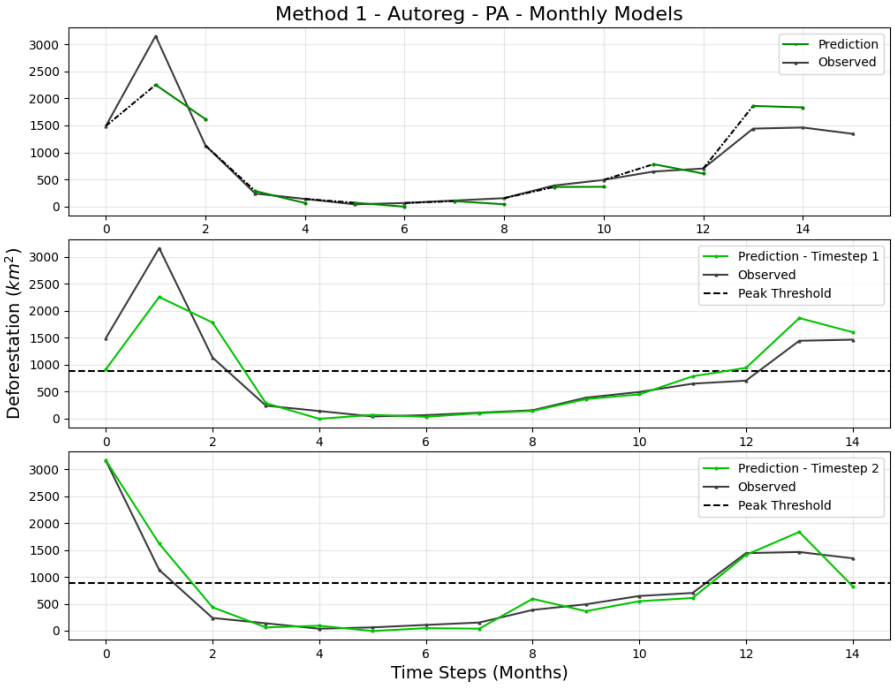


Figure 5.12: Best monthly model predictions - Pará.

PA - Metrics - Best Monthly Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	0.833	1	0.909	228	354
2	1	0.800	0.889	167	230
Mean	0.917	0.900	0.899	198	292

Table 5.8: Metrics table of the best monthly model by predicted time step - PA

5.1.4 Rondônia

The last state to be studied will be Rondônia, the values found for the RMSE of the models can be seen in Figure 5.13. It is clear that for all the algorithms tested, the best method was the first, in which the variation in the temporal lags of the deforestation series was considered. Another evident

point is that, except for the best models of each algorithm, all others have higher RMSE than the naive model. Based on this, it is possible to make a comparative analysis of each of the models chosen for each algorithm.

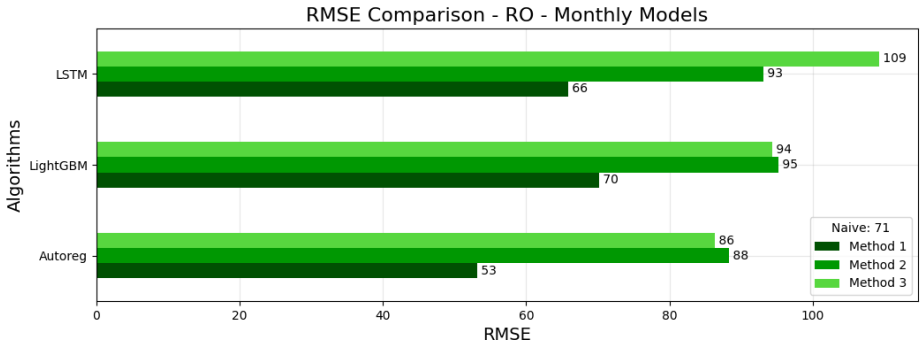


Figure 5.13: Comparison of the RMSE of each monthly model proposed for the state of Rondônia.

Examining the comparative results shown in Figure 5.14 and Table 5.9 it can be seen that in general, the models can predict the behavior of the function of interest. However, it is observed that mainly in the prediction of models that used autoregression and LightGBM, false peaks are created, this does not occur with the LSTM model. This fact is reinforced by the F1-Score of the models, even though the MAE and RMSE present lower values on average in the autoregressive model. Hence, the LSTM model will be defined as the best monthly model.

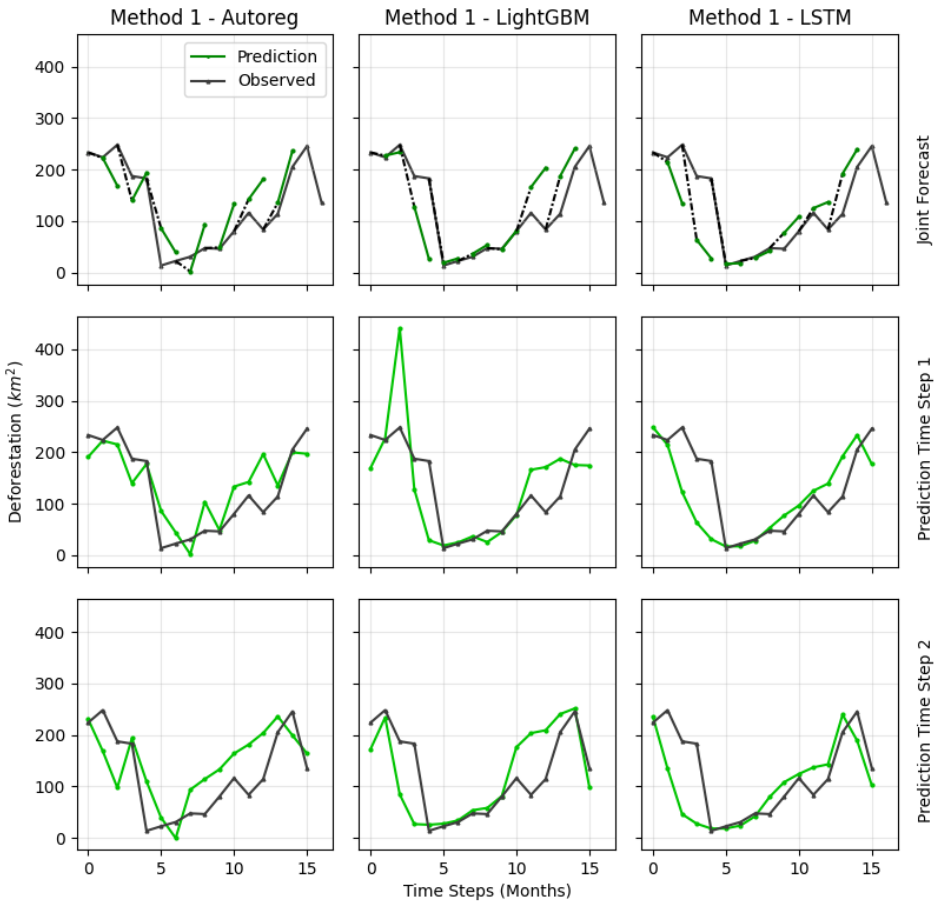


Figure 5.14: Best monthly models predictions for each algorithm used - Rondônia.

RO - Best Monthly Models

Step	1 - Autoregression			1 - LightGBM			1 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.571	36	46	0.571	52	75	0.750	46	66
2	0.571	52	60	0.667	45	65	0.667	45	65
Mean	0.571	44	53	0.619	48	70	0.708	45	66

Table 5.9: Comparative table of the best monthly models metrics by algorithm - RO

From Figure 5.15 and Table 5.10 the results of the best monthly model chosen can be viewed. Through this data, it is possible to observe more clearly that the predictions follow the behavior of the deforestation series, even though there are cases in which the peak points are not mapped. Furthermore, the data in the table indicate very balanced metrics for the two planned steps. It

is worth noting that the model's accuracy was perfect, that is, all predictions made above the adopted threshold were correct. Even though the model is not inadequate, weekly aggregated models will be used to study the possibility of improving the results, mainly to better map peak moments.

Figure 5.15: Best monthly model predictions - Rondônia.

RO - Metrics - Best Monthly Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	1	0.600	0.750	46	66
2	1	0.500	0.667	45	65
Mean	1	0.550	0.708	45	66

Table 5.10: Metrics table of the best monthly model by predicted time step - RO

The weekly aggregated data analysis will be carried out in the same way. First, the RMSE study of each model will be made to separate the best models of each algorithm applied. The results can be seen in Figure 5.16 and again the best models are obtained when method 1 is applied. One difference that can be noticed is that now no model is worse than the naive model considered. Therefore, a comparison of the best models can be made.

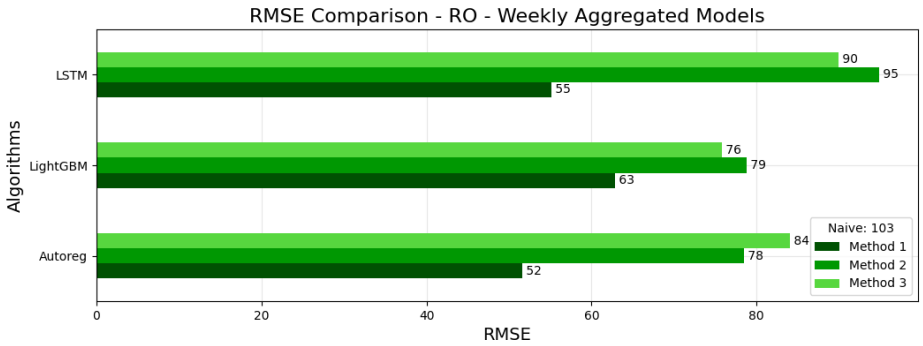


Figure 5.16: Comparison of the RMSE of each weekly aggregated model proposed for the state of Rondônia

Based on the examination of Figure 5.17 and Table 5.11 that display the comparison of metrics and graphs of the best weekly aggregated models, some conclusions can be drawn. As examined in the monthly model graphs, there is a false prediction of peak points during the test period. This can be seen mainly in the best LightGBM model, and the second step predicted by the autoregressive model. The LSTM model does not have the problem highlighted, even though it has smoothed predictions, which do not reflect the behavior of the series in question.

When observing the available metrics, it is evident that the second step of the LSTM model is the best, but the autoregressive model presents a better F1-Score and RMSE on average. Thus, combined with the fact that it can reach peak points, it can be considered a more promising model for modeling deforestation data in Rondônia. It can also be examined that the F1-Score of the monthly model using LSTM is greater than the autoregressive model applied to weekly aggregated data.

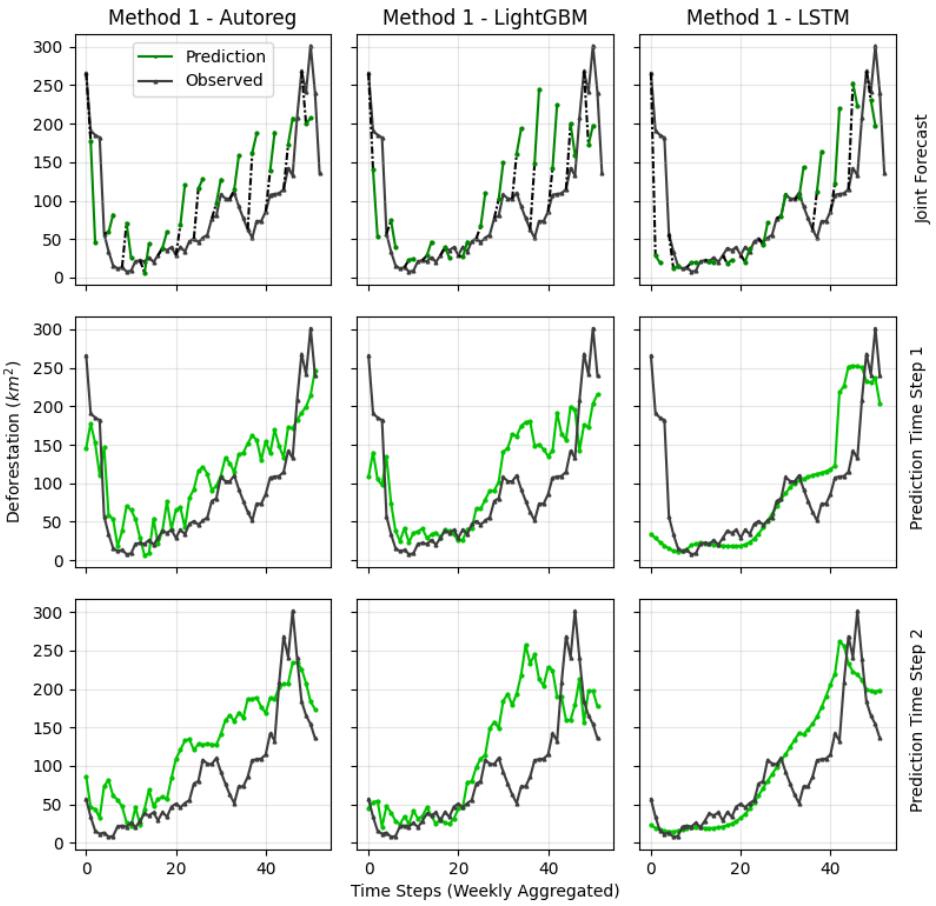


Figure 5.17: Best weekly aggregated models predictions for each algorithm used - Rondônia.

RO - Best Weekly Aggregated Models

Step	1 - Autoregression			1 - LightGBM			1 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.667	40	50	0.308	44	57	0.554	40	66
2	0.632	45	53	0.182	49	69	0.632	31	44
Mean	0.649	43	52	0.245	47	63	0.594	36	55

Table 5.11: Comparative table of the best weekly aggregated models metrics by algorithm - RO

In conclusion, Figure 5.18 and Table 5.12 show the results of the best weekly aggregated model defined. From the data, it is possible to observe that while the accuracy in classifying points is greater in the first step, recall is greater in the second. Thus, while in the first predicted time step all points classified above the adopted threshold are part of this group of points, in the

second step all points located above the threshold were identified. This implied a certain balance of the F1-Score for each predicted step. Furthermore, the regression metrics presented similar values, slightly worse in the data referring to the second temporal step.

Given what has been discussed, the weekly aggregated models do not appear to decisively improve the monthly models. When comparing the two best-selected models, although the weekly aggregated model can map peaks, the monthly model behaves better taking into account the behavior of the series. In addition, the monthly model has more consistent metrics in all aspects in contrast to the other model. Therefore, the monthly model can be defined as the best model for predicting deforestation in Rondônia.

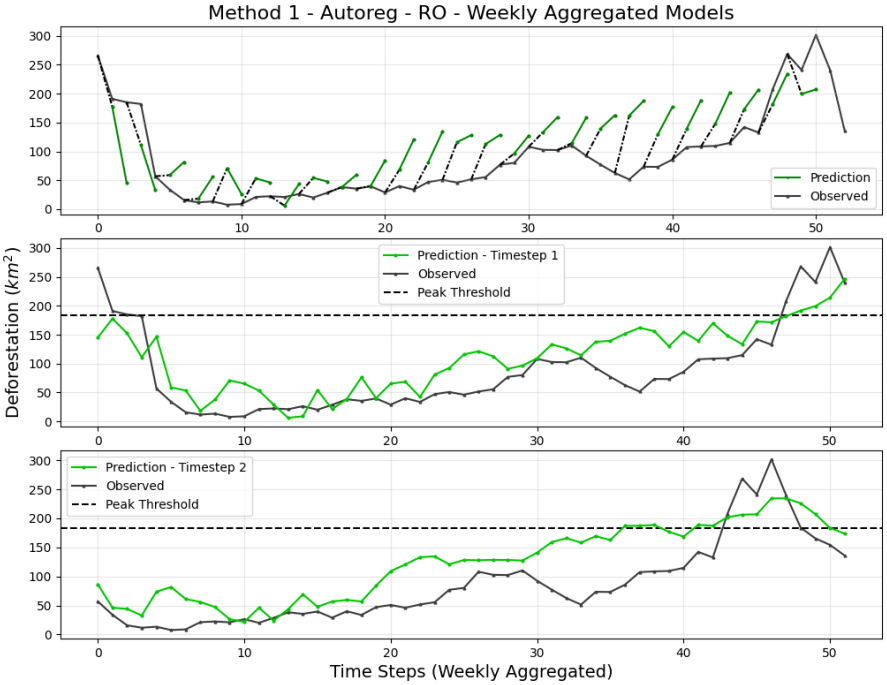


Figure 5.18: Best weekly aggregated model prediction - Rondônia.

RO - Metrics - Best Weekly Aggregated Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	1	0.500	0.667	40	50
2	0.462	1	0.632	45	53
Mean	0.732	0.750	0.649	43	52

Table 5.12: Metrics table of the best weekly aggregated model by predicted time step - RO

5.2 Municipalities Results

The analysis of the results of the municipal models will be conducted in a similar way to what was seen in the state cases. The results from all selected municipalities, Lábrea, Apiacás, and Itaituba, will be displayed below along with a study of the values and graphs found. Predictions were calculated for two temporal steps in the monthly models used. As in state cases, the average of the predicted steps' metrics was also taken into account so that there are values for evaluating the model as a whole.

5.2.1 Lábrea (AM)

Initially, the RMSE analysis of the models applied to the municipality of Lábrea in the state of Amazonas can be done through Figure 5.19. Based on this, it is possible to observe that for each algorithm a different method stands out, with the lowest RMSE being found in method 1 applied to autoregression. The best models of each algorithm present a lower RMSE than the proposed naive model. In such a manner, a comparative study between the selected models can be conducted.

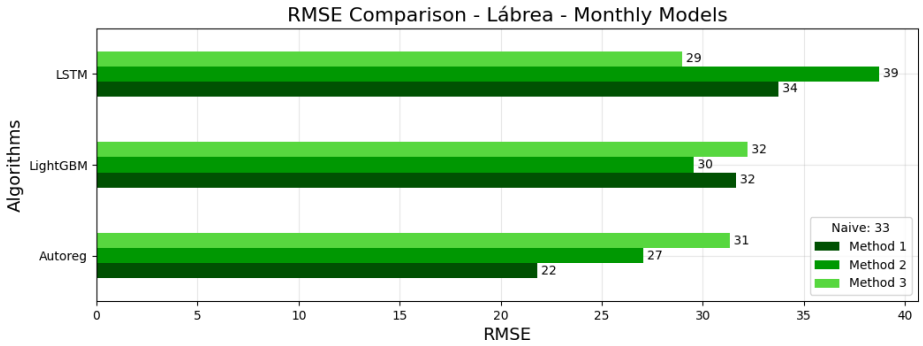


Figure 5.19: Comparison of the RMSE of each monthly model proposed for the municipality of Lábrea (AM).

Figure 5.20 shows the graphs of the chosen models. Through this figure and the metrics displayed in Table 5.13 it can be pointed out that although the three models create peaks that do not exist in the real series, they can subtly map the behavior of the series. This happens mainly with the predicted second step of each model and is proven when comparing the F1-Score of the second step with the first, except for the autoregressive model, there is a worsening in the classification of points. Furthermore, it is possible to notice that the MAE and RMSE of the models find their lowest values in the best autoregressive

model. Therefore, the final analysis will be done concerning the model that uses autoregression to predict the data.

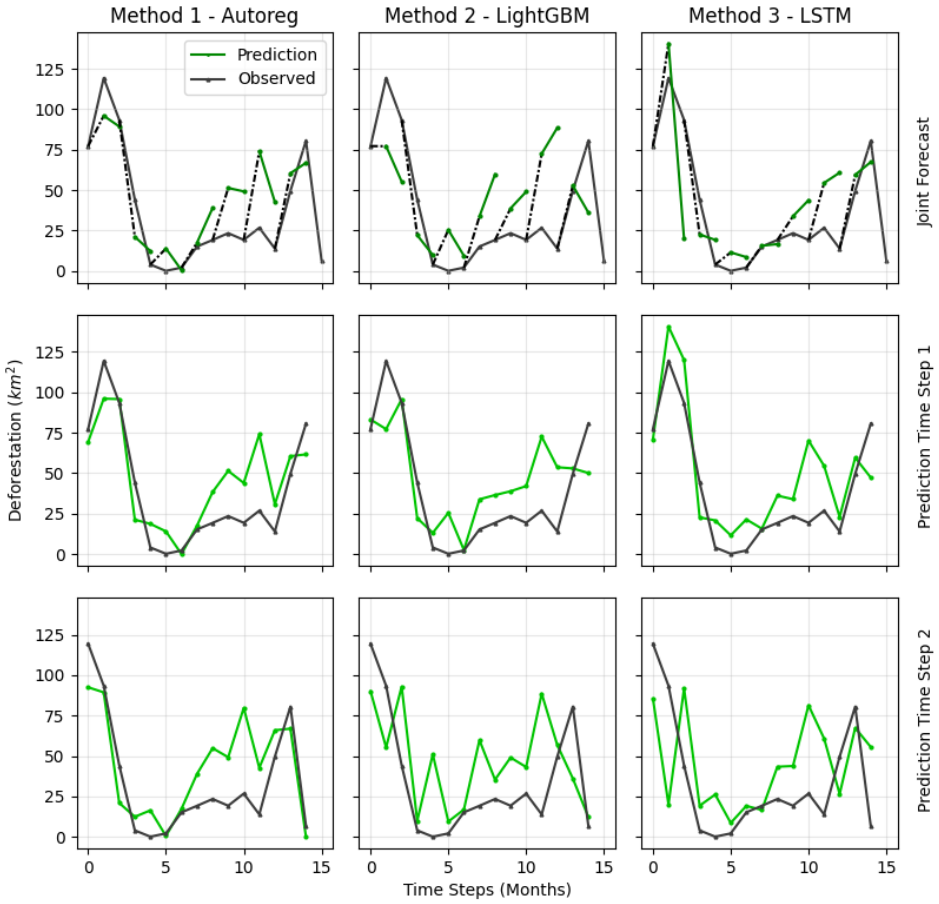


Figure 5.20: Best monthly models predictions for each algorithm used - Lábrea (AM).

Lábrea (AM) - Best Monthly Models

Step	1 - Autoregression			2 - LightGBM			3 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.750	17	21	0.750	20	25	0.750	19	22
2	0.750	19	23	0.333	28	35	0.571	29	36
Mean	0.750	18	22	0.542	24	30	0.661	24	29

Table 5.13: Comparative table of the best monthly models metrics by algorithm - Lábrea (AM)

The data relating to the best-selected model can be examined through Figure 5.21 and Table 5.14. Through them, it is observed that the prediction

errors became greater from the first to the second predicted step, as well as the data precision decreased, even though the recall increased. A point that can be reinforced about the model itself is that both MAE and RMSE are smaller than the variability of the data, calculated by the standard deviation of the series, equal to 41. This is important as it indicates that the predictions are less noisy, showing the factual predictive importance of the model.

Figure 5.21: Best monthly model predictions - Lábrea (AM).

Lábrea (AM) - Metrics - Best Monthly Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	0.750	0.750	0.750	17	21
2	0.600	1	0.750	19	23
Mean	0.675	0.875	0.750	18	22

Table 5.14: Metrics table of the best monthly model by predicted time step - Lábrea (AM)

5.2.2 Apiacás (MT)

The results from the models applied to the municipality of Apiacás in the state of Mato Grosso can be analyzed. The metrics comparison of the models created are shown in Figure 5.22. It is clear from this that method 2 obtained

the worst results for each algorithm. Method 3 stood out in the autoregressive and LightGBM models, while method 1 was the best in the models that used LSTM. All the best models have RMSE smaller than the naive model and the standard deviation of the deforestation series, respectively 24 and 19. This indicates that these three models are better than simple data variability. In this manner, to decide which is the best one of them, a comparative analysis of their graphs and metrics will be carried out.

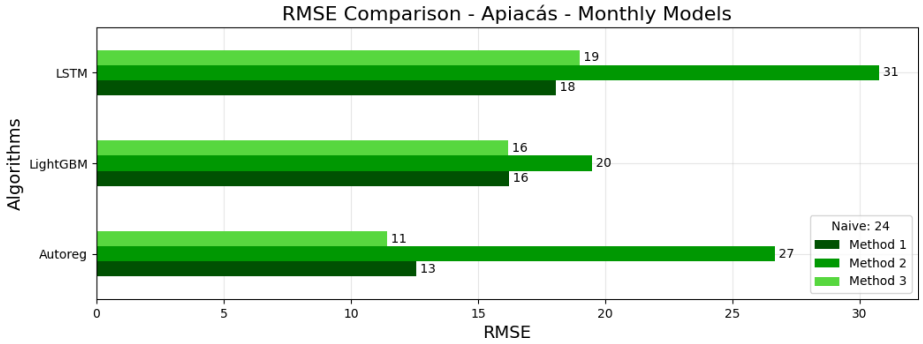


Figure 5.22: Comparison of the RMSE of each monthly model proposed for the municipality of Apiacás (MT).

By examining the graphs shown in Figure 5.23 some conclusions can be drawn regarding the choice of the best model. The most notable point is that the predictions made by the best model that uses LSTM cannot identify the peak points of the series of interest, especially in the second predicted step. The predictions made seem to accommodate periods of low deforestation. On the other hand, the autoregressive model managed to identify peak points throughout the test set, resulting in a better classification of predictions, as can be seen in Table 5.15. Therefore, the model chosen as the best was the autoregressive one due to its metrics and the ability to identify peak periods more precisely.

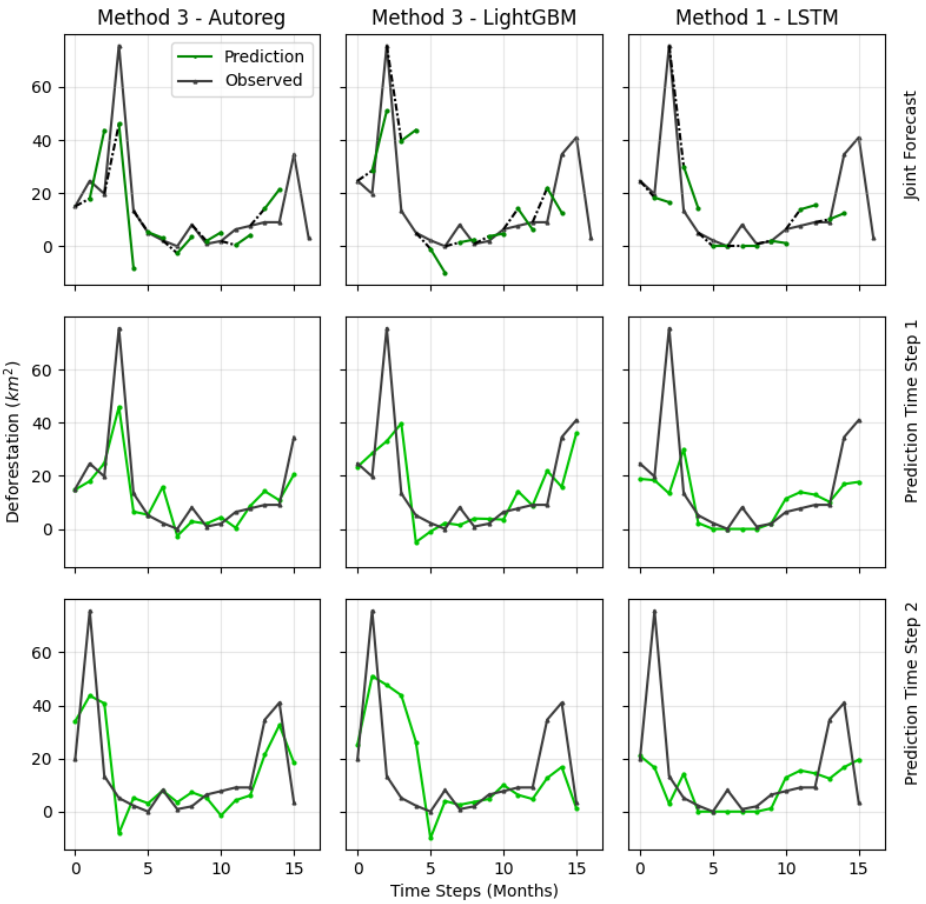


Figure 5.23: Best monthly models predictions for each algorithm used - Apiacás (MT).

Apiacás (MT) - Best Monthly Models

Step	3 - Autoregression			3 - LightGBM			1 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.909	6	10	0.833	10	14	0.800	10	18
2	0.800	10	13	0.600	13	18	0.667	11	18
Mean	0.855	8	11	0.717	11	16	0.733	10	18

Table 5.15: Comparative table of the best monthly models metrics by algorithm - Apiacás (MT)

Finally, a more detailed analysis of the best model can be carried out. Through Figure 5.24 and Table 5.16. It is possible to see that the classification of points above the adopted threshold was perfect for all steps. However, some additional points were erroneously classified as 'peak' points, which is reflected by the model's accuracy value. In general, the model obtained good results,

managing to fluctuate the forecast values between the two areas of interest, below and above the peak threshold, at the necessary times.

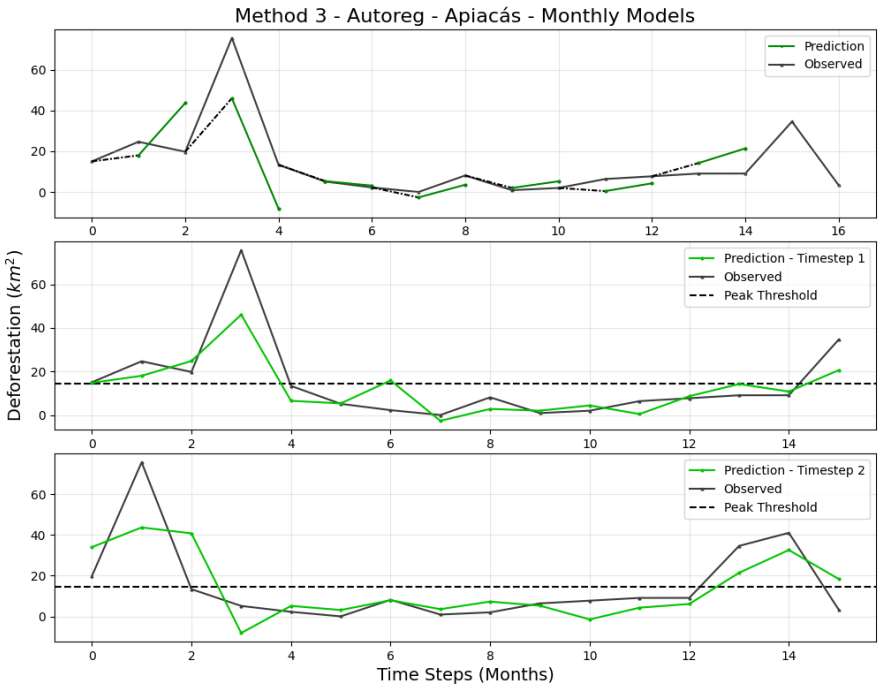


Figure 5.24: Best monthly model predictions - Apiacás (MT).

Apiacás (MT) - Metrics - Best Monthly Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	0.833	1	0.909	6	10
2	0.667	1	0.800	10	13
Mean	0.750	1	0.855	8	11

Table 5.16: Metrics table of the best monthly model by predicted time step - Apiacás (MT)

5.2.3 Itaituba (PA)

Finally, the results of the municipality of Itaituba in the state of Pará can be verified. Through Figure 5.25 the RMSE value of each applied model can be examined. It is clear that method 1 obtained better marks in the models that used autoregression and LSTM, while method 3 obtained better marks in those that used LightGBM. Once again all the best models from each algorithm performed better than the naive model. This way, a comparison between the three models can be made.

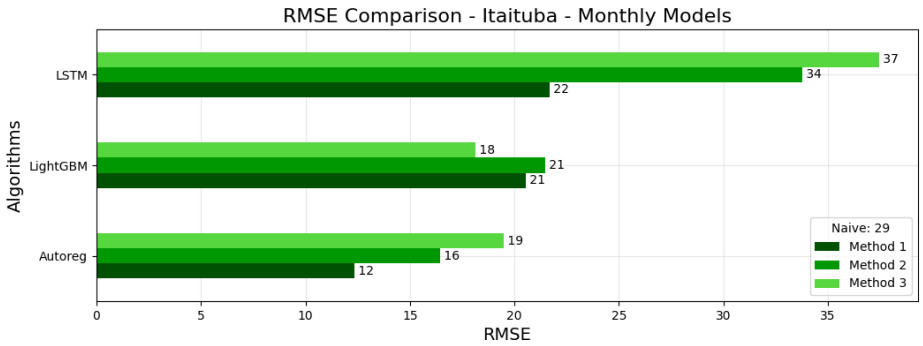


Figure 5.25: Comparison of the RMSE of each monthly model proposed for the municipality of Itaituba (PA).

Through Figure 5.26 it is possible to examine the comparison of the graphs of the best models. The first point that can be highlighted is that all models have behaviors that are close to the real values at peak moments. Based on the metrics attributed to the models, displayed in Table 5.17, the model that uses LSTM presents a better F1-Score. However, it is necessary to emphasize that the autoregressive model has better MAE and RMSE, in addition to having the most consistent metrics in the two steps foreseen. As the aim is to have models that have the most consistent predictions possible for all predicted steps, the autoregressive model was selected as the best model.

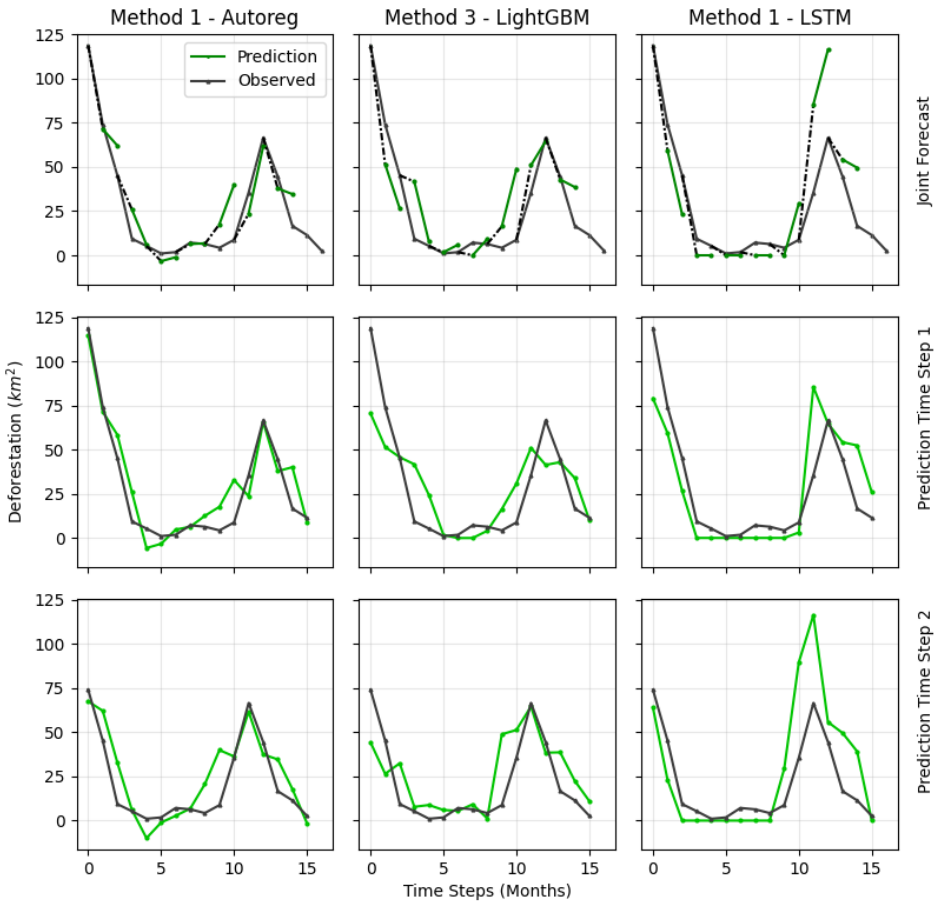


Figure 5.26: Best monthly models predictions for each algorithm used - Itaituba (PA).

Itaituba (PA) - Best Monthly Models

Step	1 - Autoregression			3 - LightGBM			1 - LSTM		
	F1	MAE	RMSE	F1	MAE	RMSE	F1	MAE	RMSE
1	0.769	9	12	0.857	14	20	0.833	14	20
2	0.769	10	13	0.667	12	17	0.727	17	23
Mean	0.769	9	12	0.762	13	18	0.780	15	22

Table 5.17: Comparative table of the best monthly models metrics by algorithm - Itaituba (PA)

Based on this choice, Figure 5.27 and Table 5.18 present more detailed information on the selected model. As already mentioned, the prediction of the two-time steps proved to be consistent, as they have similar RMSE and MAE and the same F1-Score. The point where they differ the most is when calculating precision and recall because while the first step is more precise,

the second has the highest recall. Calculating the standard deviation of the series reinforces the quality of the model. The standard deviation is equal to 28, while the average RMSE of the model is equal to 12. This implies that the model is more accurate than the variability of the deforestation series data, justifying its effectiveness and indicating that the relationships between the variables are being captured.

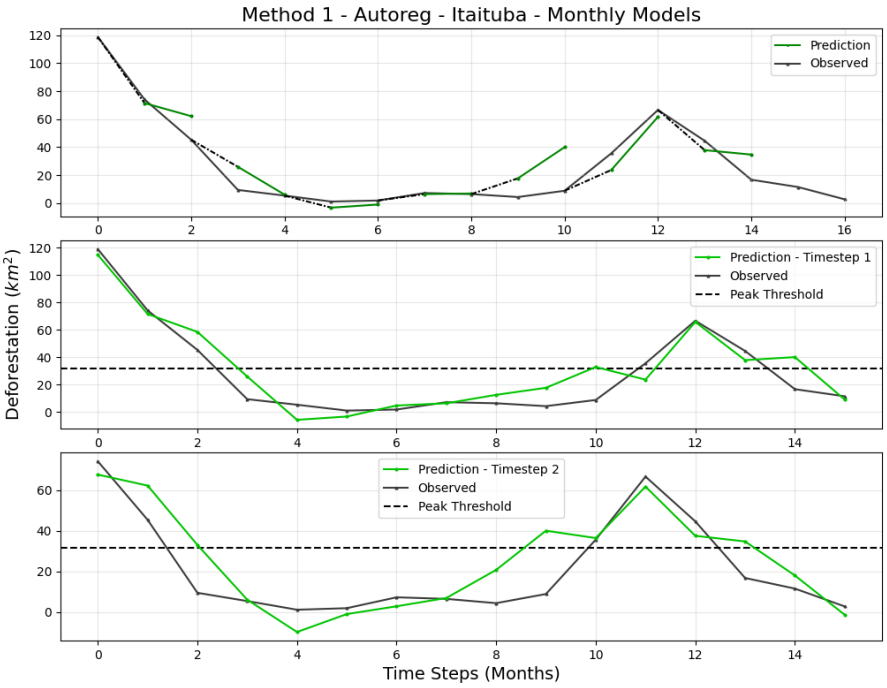


Figure 5.27: Best monthly model predictions - Itaituba (PA).

Itaituba (PA) - Metrics - Best Monthly Model

Step	Precision	Recall	F1-Score	MAE	RMSE
1	0.714	0.833	0.769	9	12
2	0.625	1	0.769	10	13
Mean	0.670	0.917	0.769	9	12

Table 5.18: Metrics table of the best monthly model by predicted time step - Itaituba (PA)

5.3 Final Discussion

Through the results shown in the previous sections, some general analyses of the results obtained can be made. As observed in most cases, the first step generally obtained better results than the second. This is due to the difficulty

in anticipating the event and the increase in uncertainty with the prediction step. Table 5.19 summarizes the best models found for each location studied. Two points stand out when analyzing it, which will be discussed below.

Local	Method	Algorithm	Data
AM	2	Autoregressive	Weekly Aggregated
MT	3	Autoregressive	Monthly
PA	1	Autoregressive	Monthly
RO	1	LSTM	Monthly
Lábrea	1	Autoregressive	Monthly
Apiacás	3	Autoregressive	Monthly
Itaituba	1	Autoregressive	Monthly

Table 5.19: Summary of the results obtained for the chosen locations.

In most cases, the autoregressive models showed better results for predicting the steps of interest in the series. The other suggested algorithms, LightGBM and LSTM, did not present such satisfactory results compared to the autoregressive models. The reason for this can be attributed to the fact that there is little data availability, especially for LSTM, considering that it is a neural network and there are many parameters and hyperparameters to be optimized, the optimal results may not have been found.

Specifically for LightGBM, there is a certain difficulty within the algorithm in capturing changes in the relationships of variables, an essential characteristic in a time series forecasting problem. Although the addition of correlated variables and the series of interest's lags may help combat this problem, it may not have been enough to achieve optimal results. On the other hand, the autoregressive models used were developed to specifically decode the temporal relationship of the data and, based on the results obtained, this was achieved. As a result, its predominance over other algorithms can be explained.

Another point that draws attention to the table is that for most of the locations considered, monthly data was enough to find good results. It should be noted that the weekly aggregated models were implemented solely in the states of Amazonas and Rondônia. Finally, of the methods used, method 1 obtained the best results more often, followed by method 3. It is possible that the variables external to the deforestation series of the locations do not help as much as expected for methods 2 and 3, indicating that the choice of variables that explain other properties of the region may be necessary.

6
Conclusion

The research presented, based on autoregression, LSTM, and LightGBM algorithms, proposes ways to forecast deforestation data in regions of the Legal Amazon. The forecasts were created based on data from the DETER-B project from August 2016 to November 2023. Two forms of data grouping were applied: monthly and weekly. Models that used monthly data always predicted two months. While models that used weekly data predicted two steps of four weeks added together, models that used this organization of data were called weekly aggregated models.

In addition to the algorithms and data organizations considered, three different methodologies were proposed for their applications. The first methodology was based on the variation in the temporal interval considered for deforestation predictions. The second method considered the correlation of the series relating to climate and deforestation variables as a determining factor for choosing the features of interest to make the model predictions. The latter used PCA on the most correlated variables to find the features that best explained the space of sampled values, applying them to predict the values of interest.

The predictive capacity of the models was evaluated based on regression and classification metrics. Initially, to select the best model, the RMSE was compared, as this metric gives greater weight to larger errors. Thus, it helps to address one of the challenges encountered during the execution of the study, predicting outliers. Since deforestation values are typically low in most series, predicting anomalies becomes more challenging. Therefore, the use of RMSE combined with the properties of algorithms to identify temporal patterns, mainly in autoregressive models and LSTM networks, makes it possible to predict abnormal moments in the series.

Furthermore, the selected ranking metrics played a key role in choosing the best models. It is crucial to ensure that the models not only predict deforestation values accurately but also correctly identify periods of high and low deforestation throughout the series. The combination of both types of metrics results in more reliable results over time. However, both predicted time steps must be consistent in terms of their results so that better planning related to preservation can be carried out.

In general, the predictive models of deforestation in the Amazon region proposed in this research are vital to protect the environment, help in the

creation of public policies, and mitigate climate change. They direct conservation efforts, promote economic sustainability, and help monitor and hold accountable illegal activities. Furthermore, they contribute to the preservation of biodiversity, essential for global ecological balance.

Regarding the results, for most of the locations studied, the objectives of interest were achieved, that is, the behavior of the series was able to be predicted, as well as the identification of peak points. The metrics revealed that the state of Pará presented the best model according to the classification metrics, having a precision of 91% and recall of 90% on average. Meanwhile, the state of Rondônia, despite achieving 100% accuracy, had an average recall of just 55%, the lowest among all models. Therefore, it was unable to identify many peak points, but there is greater confidence in it when it indicates the possible presence of extreme events in the future due to its perfect precision. The same happens with the state of Mato Grosso, where the precision is ideal, while the recall is 71%.

The state of Amazonas shows the opposite pattern compared to RO and MT, with a perfect recall but an average precision of 70%. This indicates that many more points are being predicted as peak points, implying lower reliability in the model. Furthermore, AM was the only state that obtained an RMSE greater than the standard deviation of its data, indicating that the predictions made for this state were not as effective as those made for the other regions. It is worth mentioning that the less accurate predictions may be related to the substantial increase in deforestation rates in the region, explaining the difficulty in predicting the outliers in the series.

Concerning the selected municipalities, there was greater difficulty with the accuracy of classifying the series' peak points. The municipalities of Lábrea, Apiacás, and Itaituba presented respectively 67%, 75%, and 67% precision, while their recall metrics were 87%, 100%, and 91%. Despite achieving lower-than-expected accuracy, the three models effectively projected fluctuations in deforestation rates and accurately predicted extreme points.

The objectives proposed throughout Chapter 1 can be considered complete, as models were generated to forecast deforestation values for states and municipalities in the Brazilian Legal Amazon. Such predictions were analyzed and classified according to their results and it can be seen that the proposed models managed to capture deforestation trends and patterns. The peak moments of the series, for the most part, managed to be studied and reached during the forecasts, comprising another objective highlighted at the beginning of this work. In this way, reliable models were achieved that could identify atypical periods and temporal patterns concerning the increase of deforestation in

the selected regions. The implementation of weekly aggregated models and the use of classification metrics combined with regression metrics should also be highlighted, as they represent new contributions to the area in question, as well as the use of meteorological and temporal features, in the case of models that used LightGBM.

It is observed that in most cases the best models were found when algorithms linked to autoregression were applied. The worst results of models that used LSTM and LightGBM may have been caused by the lack of data for training and validation. The DETER-B project is relatively new and using data from other projects may not help solve this problem, as the accuracy of previous projects was considerably worse. Another factor that may have limited the results is the set of available variables, which opens up a discussion on what future projects could incorporate.

Given everything that has been said, the use of geospatial and socioeconomic variables, such as altitude (52), type of relief (52), number of inhabitants (53), and presence of agribusiness (54) can be studied aiming to better explain the relationships between the model variables. Furthermore, the increase in the number of data seems to be crucial for models to achieve better results, given their possible training with various hyperparameters. In this way, the use of generative artificial intelligence (55) can be used to expand deforestation series, or generate data based on a highly correlated variable.

Other models can also be tested, applying probabilistic methods can be made to make predictions in smaller regions that are still at risk of deforestation (14) (56). The inclusion of practices to improve the quality of climate data can also be carried out through kriging (57), obtaining values continuously over the entire territory of the Legal Amazon, or through satellite data. Furthermore, the study of possible data bias can be carried out since deforestation data are indirect measurements from satellite images. Finally, the consideration of other ways of dividing the training, validation, and test sets can be adopted so that the model can learn and be more adaptable to any situation, we can cite as an example the k-Fold Cross Validation (58).

7
References

[1] RIVERO, S.; ALMEIDA, O.; ÁVILA, S. ; OLIVEIRA, W.. **Pecuária e desmatamento: uma análise das principais causas diretas do desmatamento na amazônia**. Nova economia, 19:41–66, 2009. 1

[2] GABARDO, G.; SARZEDAS, C. G. ; DA SILVA, H. L.. **Queimadas na amazônia brasileira: Brasil em chamas**. A educação ambiental em uma perspectiva interdisciplinar. Disponível em:< https://downloads. editoracientifica. org/articles/200800872. pdf> Acesso em, 4, 2021. 1

[3] SILVA MONTEIRO, A. L.; BARRETO, P. G.; PANTOJA, F. L. D. S.; GERWING, J. J. ; OTHERS. **Impactos da exploração madeireira e do fogo em florestas de transição da amazônia legal**. Scientia forestalis, 2004. 1

[4] DOMINGUEZ, D.; DEL VILLAR, L. D. J.; PANTOJA, O. ; GONZÁLEZ-RODRÍGUEZ, M.. **Forecasting amazon rain-forest deforestation using a hybrid machine learning model**. Sustainability, 14(2):691, 2022. 1

[5] MEDSKER, L. R.. **Hybrid neural network and expert systems**. Springer Science & Business Media, 2012. 1

[6] HOCHREITER, S.; SCHMIDHUBER, J.. **Long short-term memory**. Neural computation, 9(8):1735–1780, 1997. 1, 2.1, 2.5.3

[7] RUMELHART, D. E.; HINTON, G. E. ; WILLIAMS, R. J.. **Learning representations by back-propagating errors**. nature, 323(6088):533–536, 1986. 1

[8] LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W. ; JACKEL, L. D.. **Backpropagation applied to handwritten zip code recognition**. Neural computation, 1(4):541–551, 1989. 1, 2.1

[9] FODOR, G.; CONDE, M. V.. **Rapid deforestation and burned area detection using deep multimodal learning on satellite imagery**. arXiv preprint arXiv:2307.04916, 2023. 1

[10] SAHA, S.; BHATTACHARJEE, S.; SHIT, P. K.; SENGUPTA, N. ; BERA, B.. **Deforestation probability assessment using integrated**

machine learning algorithms of eastern himalayan foothills (india). Resources, Conservation & Recycling Advances, 14:200077, 2022. 1

[11] CORTES, C.. **Support-vector networks**. Machine Learning, 1995. 1

[12] NEAL, R. M.. **Bayesian learning for neural networks**, volumen 118. Springer Science & Business Media, 2012. 1

[13] BISHOP, C. M.; NASRABADI, N. M.. **Pattern recognition and machine learning**, volumen 4. Springer, 2006. 1, 2.1, 2.4, 2.6.4, 2.6.5, 2.6.6, 2.6.7, 2.6.8, 2.7

[14] SILVA, A. C.; FONSECA, L. M.; KÖRTING, T. S. ; ESCADA, M. I. S.. **A spatio-temporal bayesian network approach for deforestation prediction in an amazon rainforest expansion frontier**. Spatial statistics, 35:100393, 2020. 1, 6

[15] MAYFIELD, H.; SMITH, C.; GALLAGHER, M. ; HOCKINGS, M.. **Use of freely available datasets and machine learning methods in predicting deforestation**. Environmental modelling & software, 87:17–28, 2017. 1

[16] RASMUSSEN, C. E.. **Gaussian processes in machine learning**. In: SUMMER SCHOOL ON MACHINE LEARNING, p. 63–71. Springer, 2003. 1

[17] BROCKWELL, P. J.; DAVIS, R. A.. **Introduction to time series and forecasting**. Springer, 2002. 1, 2.5.1

[18] BAŞARAN, N.; MATCI, D. K. ; AVDAN, U.. **Using multiple linear regression to analyze changes in forest area: the case study of akdeniz region**. International Journal of Engineering and Geosciences, 7(3):247–263, 2022. 1

[19] KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q. ; LIU, T.-Y.. **Lightgbm: A highly efficient gradient boosting decision tree**. Advances in neural information processing systems, 30, 2017. 1, 2.1, 2.5.2

[20] MCCARTHY, J.; MINSKY, M. L.; ROCHESTER, N. ; SHANNON, C. E.. **A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955**. AI Magazine, 27(4):12, Dec. 2006. 2.1

[21] WATT, J.; BORHANI, R. ; KATSAGGELOS, A. K.. **Machine learning refined: Foundations, algorithms, and applications**. Cambridge University Press, 2020. 2.1

[22] BLOCK, H.-D.. **The perceptron: A model for brain functioning. i**. Reviews of Modern Physics, 34(1):123, 1962. 2.1

[23] ABDI, H.; VALENTIN, D. ; EDELMAN, B.. **Neural networks**. Número 124. Sage, 1999. 2.1

[24] MAN, D.; VISION, A.. **A computational investigation into the human representation and processing of visual information**. WH San Francisco: Freeman and Company, San Francisco, 1:1, 1982. 2.1

[25] KESELJ, V.. **Book review: Speech and language processing by daniel jurafsky and james h. martin**. Computational Linguistics, 35(3), 2009. 2.1

[26] VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł. ; POLOSUKHIN, I.. **Attention is all you need**. Advances in neural information processing systems, 30, 2017. 2.1

[27] SZELISKI, R.. **Computer vision: algorithms and applications**. Springer Nature, 2022. 2.1

[28] BOX, G. E.; JENKINS, G. M.; REINSEL, G. C. ; LJUNG, G. M.. **Time series analysis: forecasting and control**. John Wiley & Sons, 2015. 2.1

[29] GOODFELLOW, I.; BENGIO, Y. ; COURVILLE, A.. **Deep learning**. MIT press, 2016. 2.1, 2.5.3

[30] GERS, F. A.; SCHMIDHUBER, J. ; CUMMINS, F.. **Learning to forget: Continual prediction with lstm**. Neural computation, 12(10):2451–2471, 2000. 2.1, 2.5.3

[31] CHUNG, J.; GULCEHRE, C.; CHO, K. ; BENGIO, Y.. **Gated feedback recurrent neural networks**. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, p. 2067–2075. PMLR, 2015. 2.1

[32] MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C. ; HUNG BYERS, A.. **Big data: The next frontier for innovation, competition, and productivity**. 2011. 2.1

[33] LIN, Y.; GUO, H. ; HU, J.. **An svm-based approach for stock market trend prediction**. In: THE 2013 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), p. 1–7, 2013. 2.2

[34] MENDES, P. H. R.; MALVEZZI, W. R.. **Geração de músicas polifônicas utilizando redes neurais artificais**. Programa de Iniciação Científica-PIC/UniCEUB-Relatórios de Pesquisa, 2019. 2.2

[35] JOLLIFFE, I. T.. **Principal component analysis for special types of data**. Springer, 2002. 2.3, 2.5.4

[36] GOLUB, G. H.; REINSCH, C.. **Singular value decomposition and least squares solutions**. In: HANDBOOK FOR AUTOMATIC COMPUTATION: VOLUME II: LINEAR ALGEBRA, p. 134–151. Springer, 1971. 2.3, 2.5.4

[37] SCHUMPETER, J. A.. Journal of the American Statistical Association, 31(196):791–795, 1936. 2.5.1

[38] ZHU, R.. **Gradient-based sampling: An adaptive importance sampling for least-squares**. In: Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I. ; Garnett, R., editors, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, volumen 29. Curran Associates, Inc., 2016. 2.5.2

[39] CHAI, T.; DRAXLER, R. R.. **Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature**. Geoscientific Model Development, 7(3):1247–1250, 2014. 2.6.1, 2.6.2

[40] DE MYTTENAERE, A.; GOLDEN, B.; LE GRAND, B. ; ROSSI, F.. **Mean absolute percentage error for regression models**. Neurocomputing, 192:38–48, 2016. Advances in artificial neural networks, machine learning and computational intelligence. 2.6.3

[41] LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P.. **Grid search, random search, genetic algorithm: A big comparison for nas**, 2019. 2.7.1, 2.7.2

[42] ALVES DE OLIVEIRA, B. F.; BOTTINO, M. J.; NOBRE, P. ; NOBRE, C. A.. **Deforestation and climate change are projected to increase heat stress risk in the brazilian amazon**. Communications Earth & Environment, 2(1):207, 2021. 3.1

[43] ELLWANGER, J. H.; KULMANN-LEAL, B.; KAMINSKI, V. L.; VALVERDE-VILLEGAS, J.; VEIGA, A. B. G.; SPILKI, F. R.; FEARNSIDE, P. M.; CAESAR, L.; GIATTI, L. L.; WALLAU, G. L. ; OTHERS. **Beyond diversity loss and climate change: Impacts of amazon deforestation on infectious diseases and public health**. Anais da Academia Brasileira de Ciências, 92:e20191375, 2020. 3.1

[44] ROCHA, L. R. L.. **A correlação entre doenças respiratórias e o incremento das queimadas em alta floresta e peixoto de azevedo norte do mato grosso-amazônia legal**. Revista Brasileira de Políticas Públicas, 6(1):246–254, 2016. 3.1, 3.3

[45] TERRABRASILIS, INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Limites dos Estados**. `https://terrabrasilis.dpi.inpe.br/geonetwork/srv/eng/catalog.search#/metadata/a8e0661a-5f47-4928-bc79-ee9ad1aa21ab`. Access: 24/06/2024. 3.2

[46] TERRABRASILIS, INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Limites dos Municípios**. `https://terrabrasilis.dpi.inpe.br/geonetwork/srv/eng/catalog.search#/metadata/94002b1c-d537-4140-a0ef-f5c84e219a62`. Access: 24/06/2024. 3.2

[47] DINIZ, C. G.; SOUZA, A. A. D. A.; SANTOS, D. C.; DIAS, M. C.; LUZ, N. C. D.; MORAES, D. R. V. D.; MAIA, J. S.; GOMES, A. R.; NARVAES, I. D. S.; VALERIANO, D. M.; MAURANO, L. E. P. ; ADAMI, M.. **Deter-b: The new amazon near real-time deforestation detection system**. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(7):3619–3628, 2015. 3.2.1

[48] TERRABRASILIS, INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Aviso, degradação e exploração madeireira na Amazônia Legal à partir de 2016**. `https://terrabrasilis.dpi.inpe.br/geonetwork/srv/por/catalog.search#/metadata/f2153c4a-915b-48a6-8658-963bdce7366c`. Access: 24/06/2024. 3.2.1

[49] INSTITUTO NACIONAL DE METEOROLOGIA (INMET). **Dados Históricos de Meteorologia do INMET**. `https://portal.inmet.gov.br/dadoshistoricos`. Access: 24/06/2024. 3.2.2

[50] BARONA, E.; RAMANKUTTY, N.; HYMAN, G. ; COOMES, O. T.. **The role of pasture and soybean in deforestation of the brazilian amazon**. Environmental Research Letters, 5(2):024002, 2010. 3.3

[51] KINGMA, D. P.; BA, J.. **Adam: A method for stochastic optimization**. arXiv preprint arXiv:1412.6980, 2014. 4.3.3

[52] MA, C.; PU, R.; DOWNS, J. ; JIN, H.. **Characterizing spatial patterns of amazon rainforest wildfires and driving factors by using remote sensing and gis geospatial technologies**. Geosciences, 12(6):237, 2022. 6

[53] PFAFF, A. S.. **What drives deforestation in the brazilian amazon?: Evidence from satellite and socioeconomic data**. Journal of environmental economics and management, 37(1):26–43, 1999. 6

[54] DE ANDRADE VASCONCELOS, P. G.; ANGELO, H.; DE ALMEIDA, A. N.; MATRICARDI, E. A. T.; MIGUEL, E. P.; DE PAULA, M. F.; GONCALEZ, J. C.; JOAQUIM, M. S. ; OTHERS. **Determinants of the brazilian amazon deforestation**. African Journal of Agricultural Research, 12(3):169–176, 2017. 6

[55] ZHANG, C.; KUPPANNAGARI, S. R.; KANNAN, R. ; PRASANNA, V. K.. **Generative adversarial network for synthetic time series data generation in smart grids**. In: 2018 IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS, CONTROL, AND COMPUTING TECHNOLOGIES FOR SMART GRIDS (SMARTGRIDCOMM), p. 1–6. IEEE, 2018. 6

[56] SAHA, S.; SAHA, M.; MUKHERJEE, K.; ARABAMERI, A.; NGO, P. T. T. ; PAUL, G. C.. **Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, reptree: A case study at the gumani river basin, india**. Science of the Total Environment, 730:139197, 2020. 6

[57] CRESSIE, N.. **The origins of kriging**. Mathematical geology, 22:239–252, 1990. 6

[58] SEKULIĆ, A.; KILIBARDA, M.; HEUVELINK, G. B.; NIKOLIĆ, M. ; BAJAT, B.. **Random forest spatial interpolation**. Remote Sensing, 12(10):1687, 2020. 6

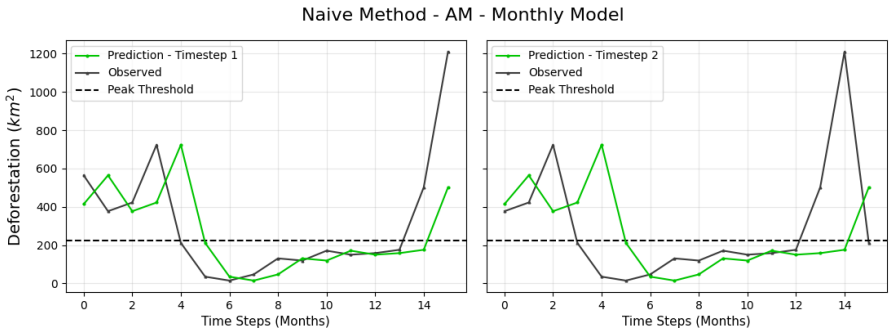
A
Appendices

A.1
Extra Results Graphs

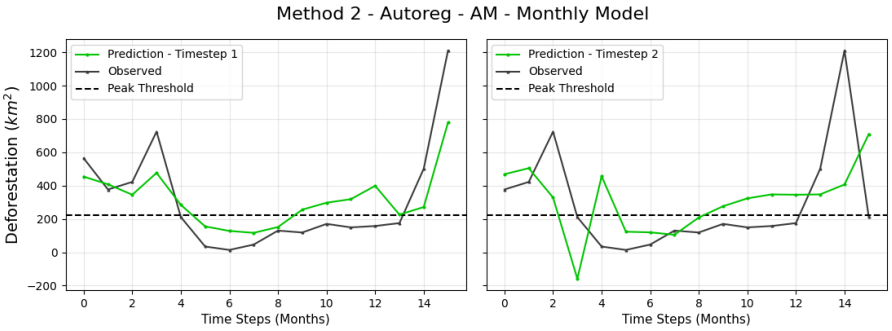
Since not all graphs of the best models were included in the results chapter, they will be presented below. It is important to note that only the graphs of the top-performing models for each algorithm will be shown. The displayed graphs illustrate the predictions for the first and second steps.

A.1.1
States

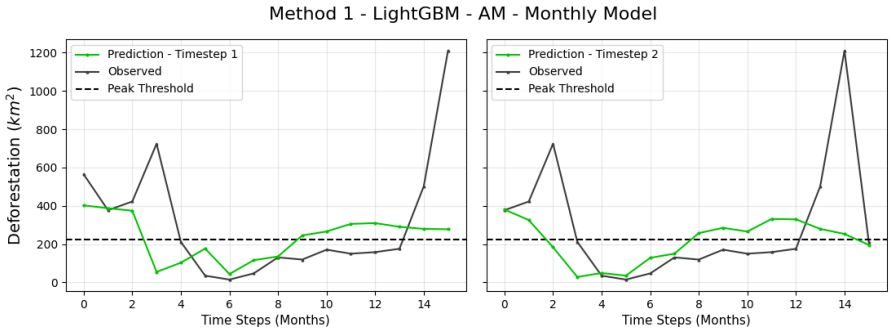
A.1.1.1 Amazonas



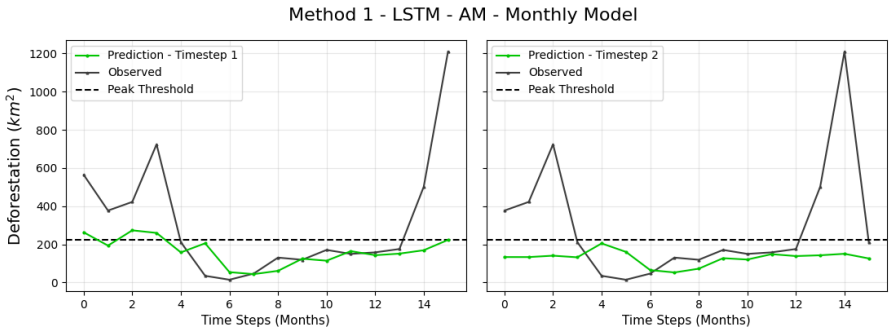
(a) Naive Model - Monthly - AM



(b) Best Autoregressive Model - Monthly - AM

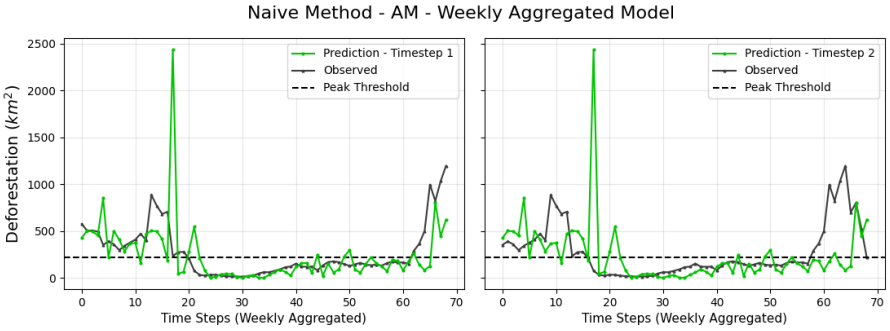


(c) Best LightGBM Model - Monthly - AM

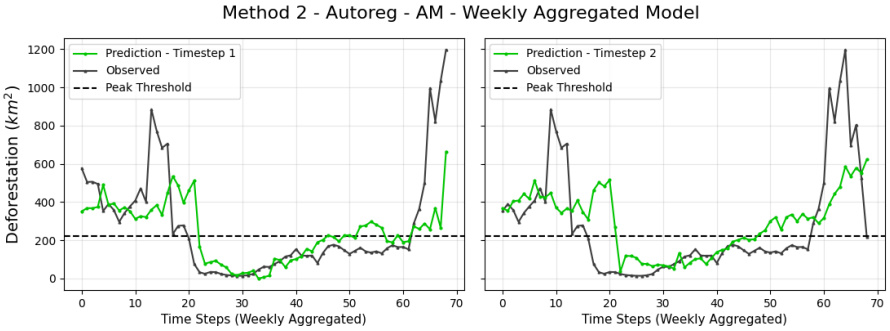


(d) Best LSTM Model - Monthly - AM

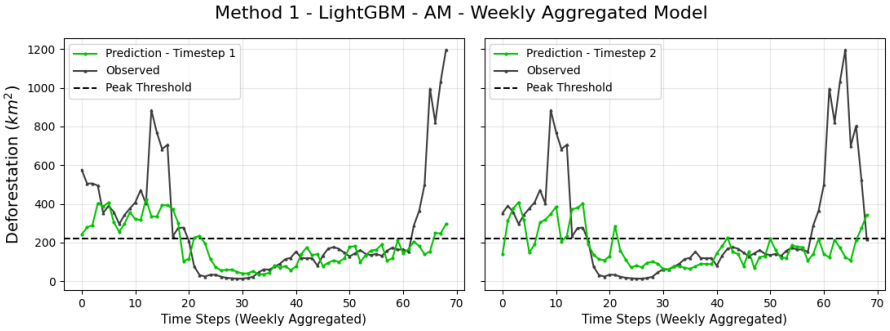
Figure A.1: Graphs of best results by algorithm of monthly models and naive model - AM.



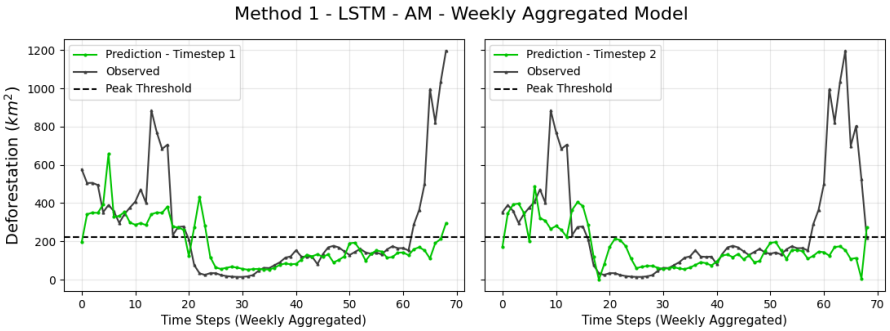
(a) Naive Model - Weekly Aggregated - AM



(b) Best Autoregressive Model - Weekly Aggregated - AM



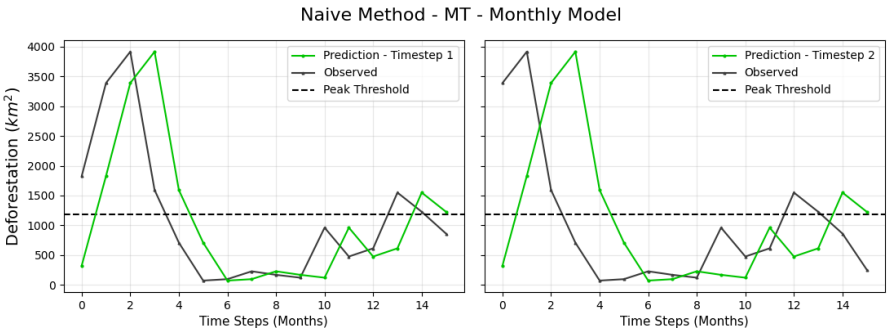
(c) Best LightGBM Model - Weekly Aggregated - AM



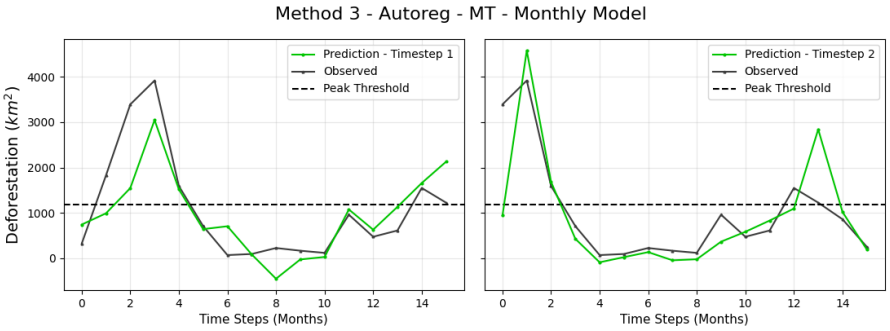
(d) Best LSTM Model - Weekly Aggregated - AM

Figure A.2: Graphs of best results by algorithm of weekly aggregated models and naive model - AM.

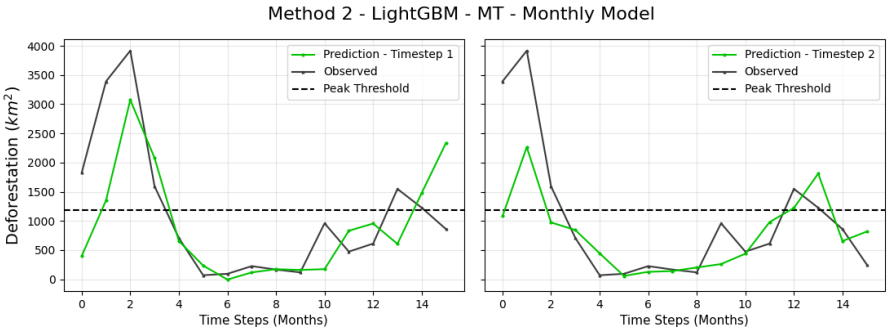
A.1.1.2 Mato Grosso



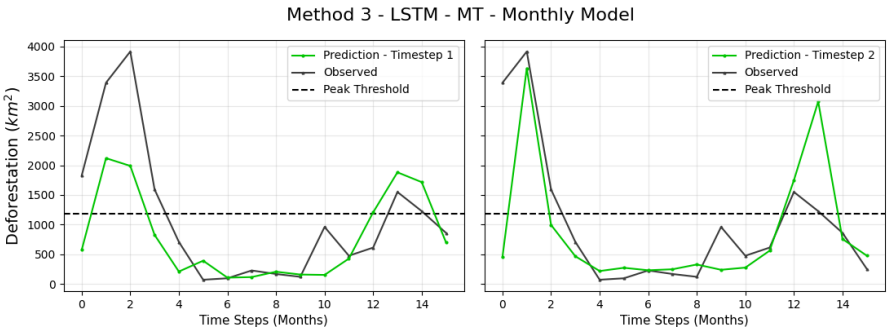
(a) Naive Model - Monthly - MT



(b) Best Autoregressive Model - Monthly - MT



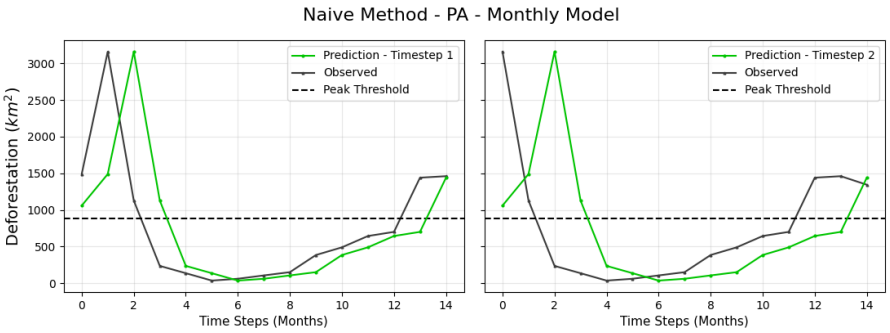
(c) Best LightGBM Model - Monthly - MT



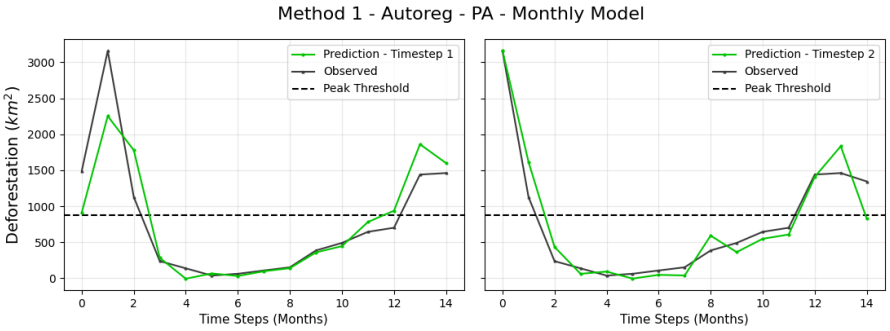
(d) Best LSTM Model - Monthly - MT

Figure A.3: Graphs of best results by algorithm of monthly models and naive model - MT.

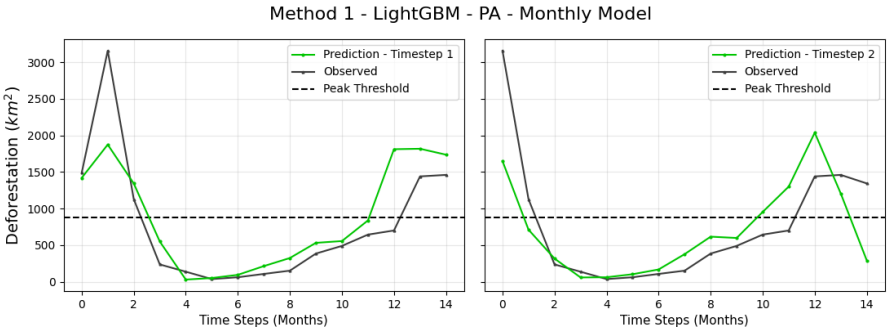
A.1.1.3 Pará



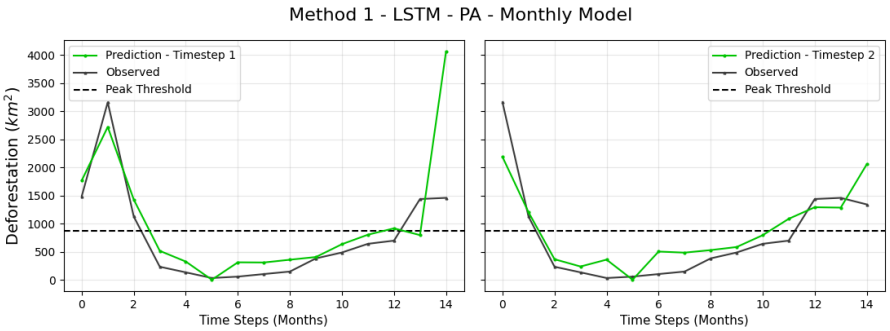
(a) Naive Model - Monthly - PA



(b) Best Autoregressive Model - Monthly - PA



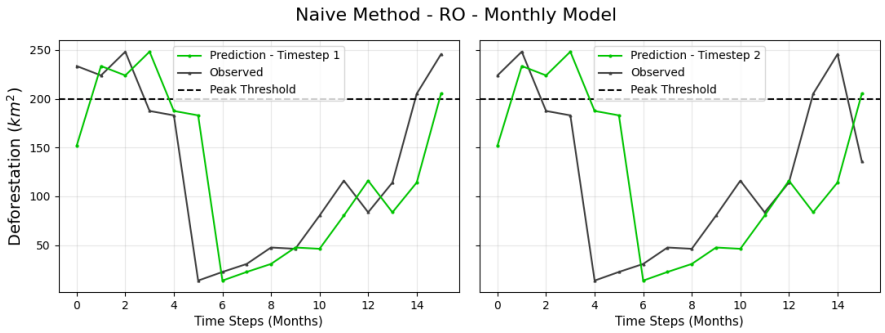
(c) Best LightGBM Model - Monthly - PA



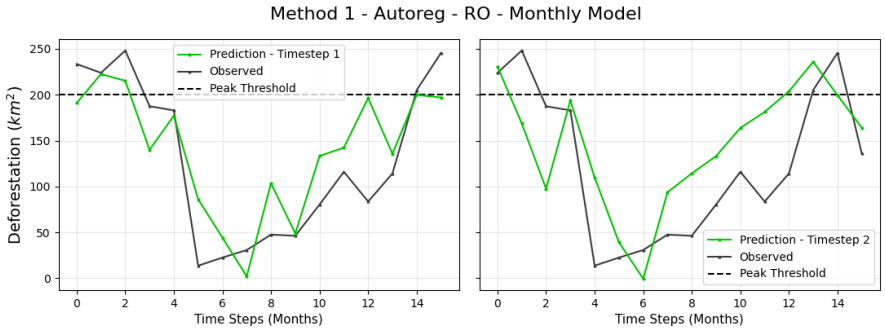
(d) Best LSTM Model - Monthly - PA

Figure A.4: Graphs of best results by algorithm of monthly models and naive model - PA.

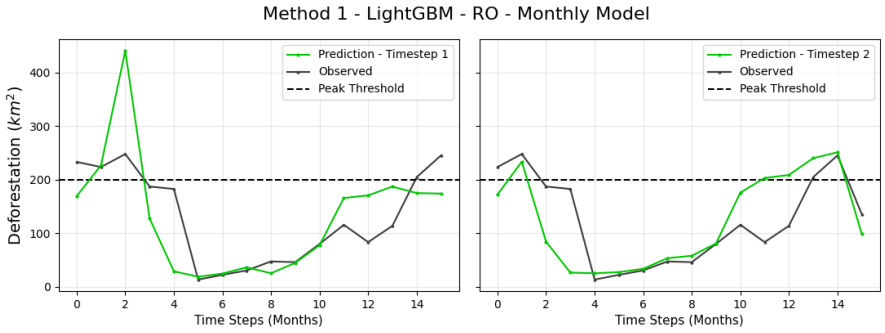
A.1.1.4 Rondônia



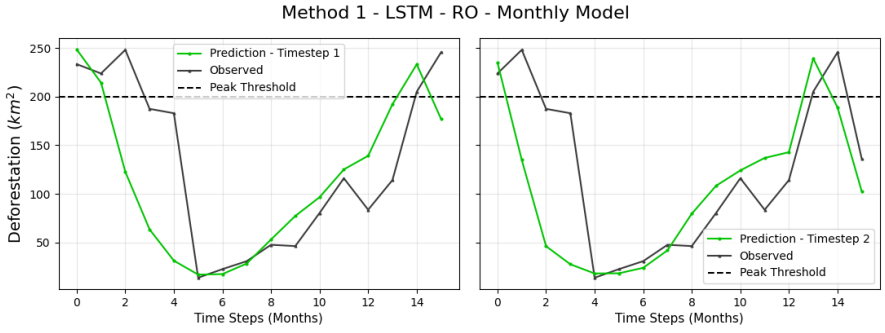
(a) Naive Model - Monthly - RO



(b) Best Autoregressive Model - Monthly - RO

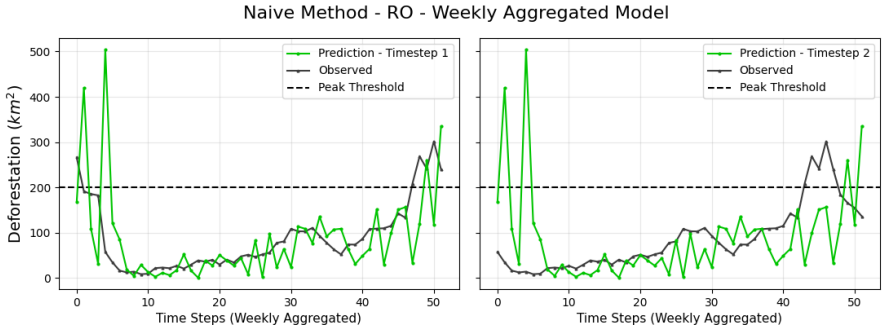


(c) Best LightGBM Model - Monthly - RO

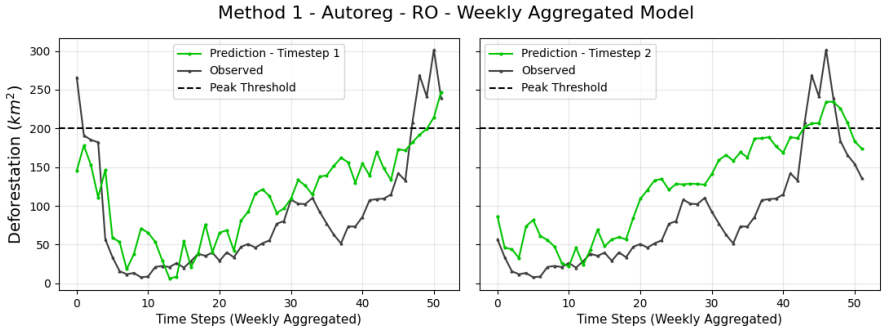


(d) Best LSTM Model - Monthly - RO

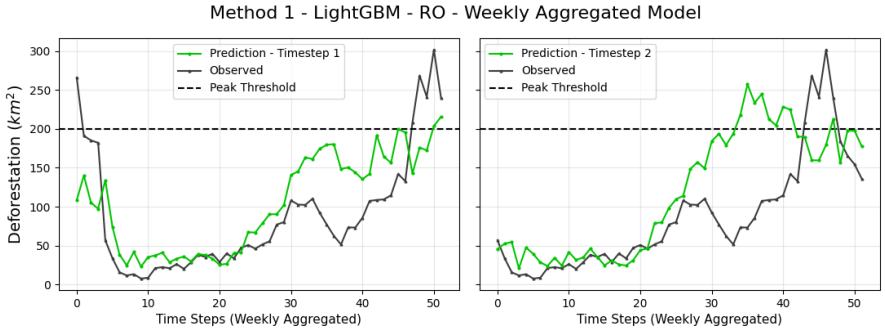
Figure A.5: Graphs of best results by algorithm of monthly models and naive model - RO.



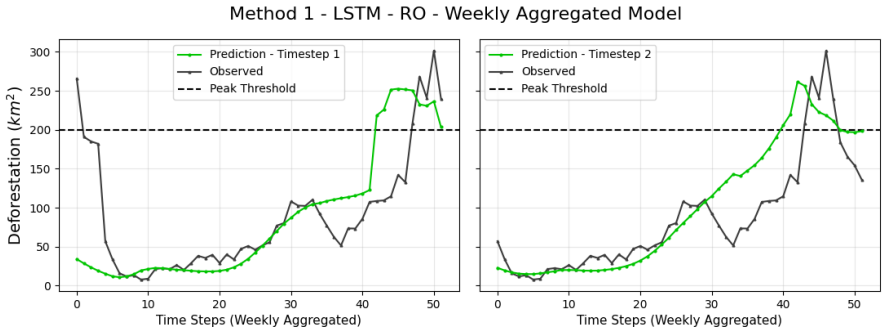
(a) Naive Model - Weekly Aggregated - RO



(b) Best Autoregressive Model - Weekly Aggregated - RO



(c) Best LightGBM Model - Weekly Aggregated - RO

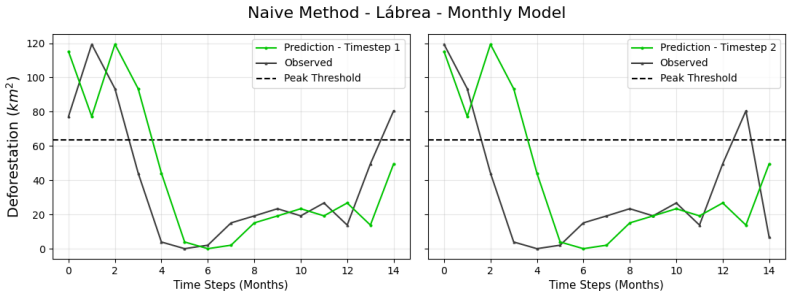


(d) Best LSTM Model - Weekly Aggregated - RO

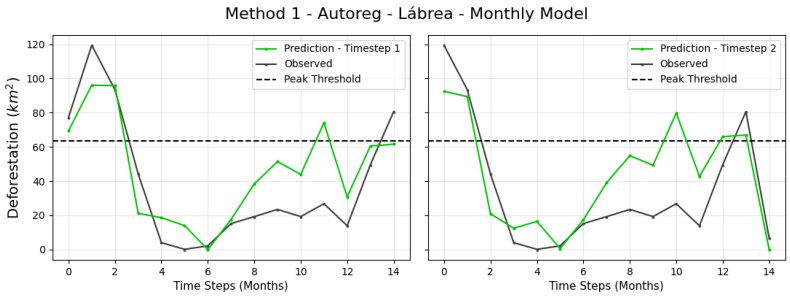
Figure A.6: Graphs of best results by algorithm of weekly aggregated models and naive model - RO.

A.1.2 Municipalities

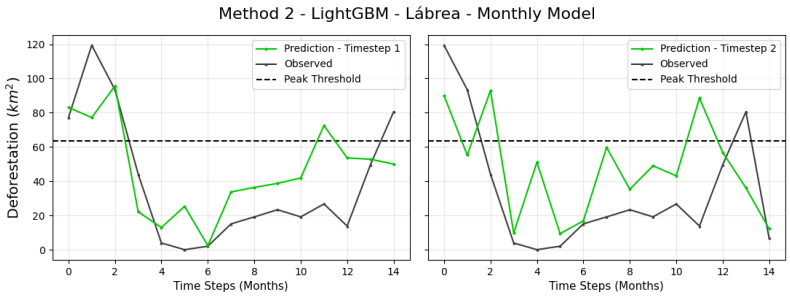
A.1.2.1 Lábrea (AM)



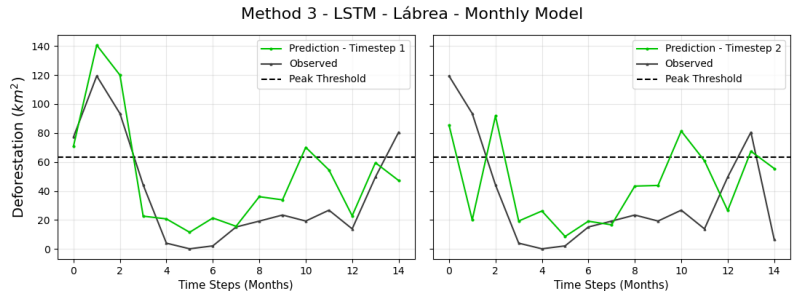
(a) Naive Model - Monthly - Lábrea



(b) Best Autoregressive Model - Monthly - Lábrea



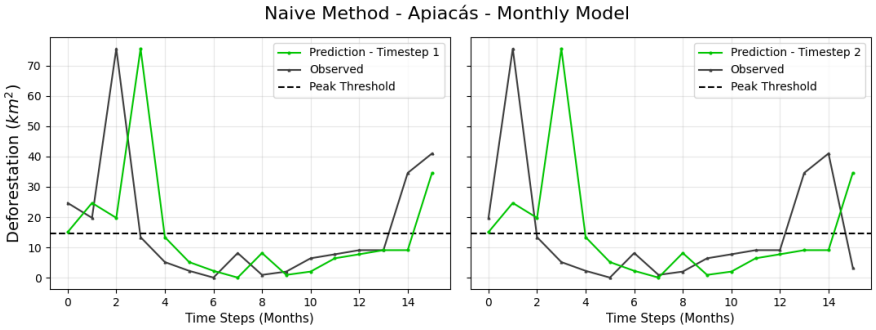
(c) Best LightGBM Model - Monthly - Lábrea



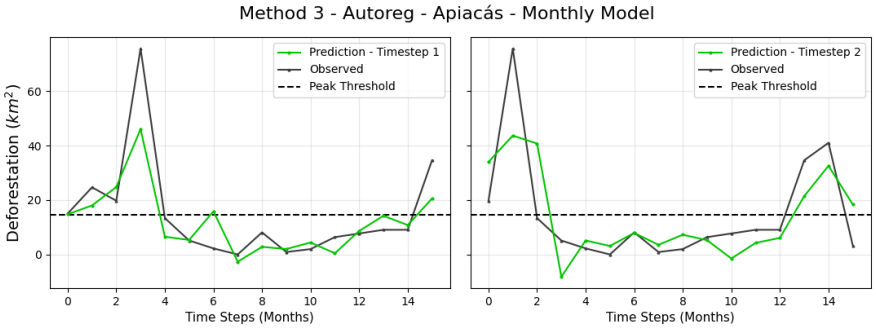
(d) Best LSTM Model - Monthly - Lábrea

Figure A.7: Graphs of best results by algorithm of monthly models and naive model - Lábrea (AM).

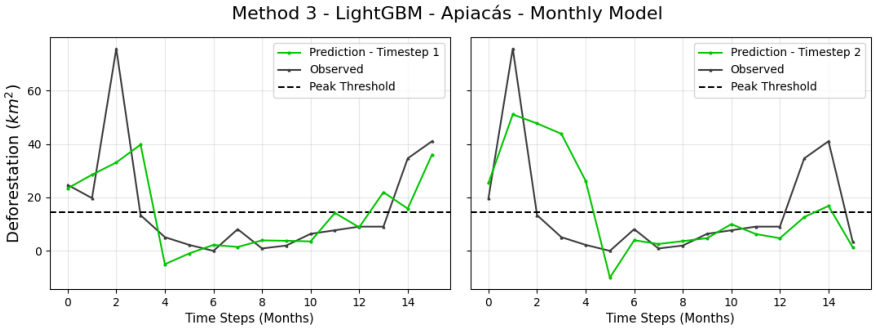
A.1.2.2 Apiacás (MT)



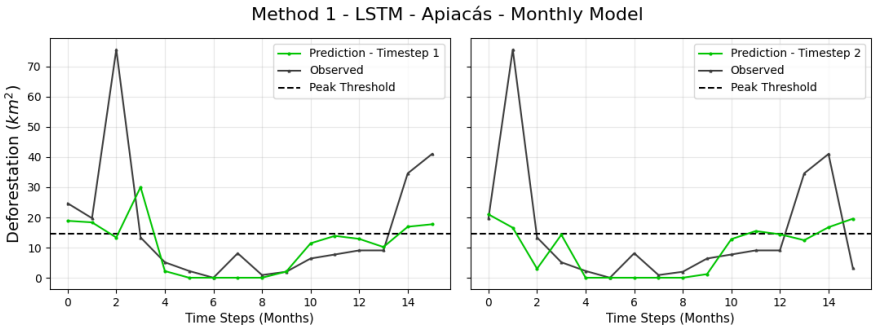
(a) Naive Model - Monthly - Apiacás



(b) Best Autoregressive Model - Monthly - Apiacás



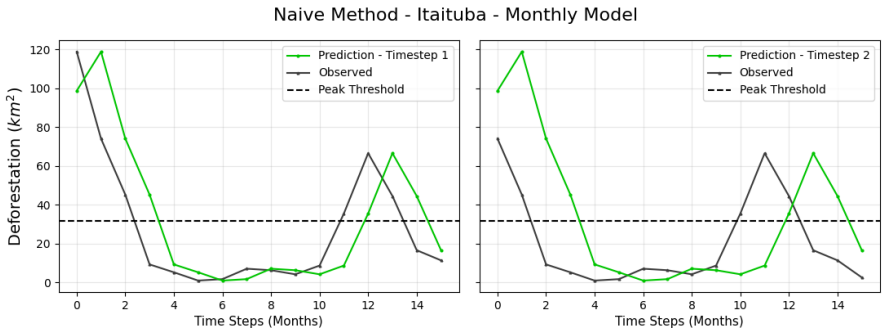
(c) Best LightGBM Model - Monthly - Apiacás



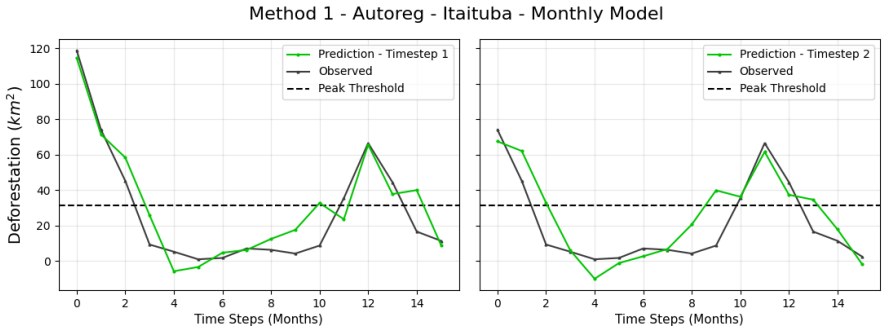
(d) Best LSTM Model - Monthly - Apiacás

Figure A.8: Graphs of best results by algorithm of monthly models and naive model - Apiacás (MT).

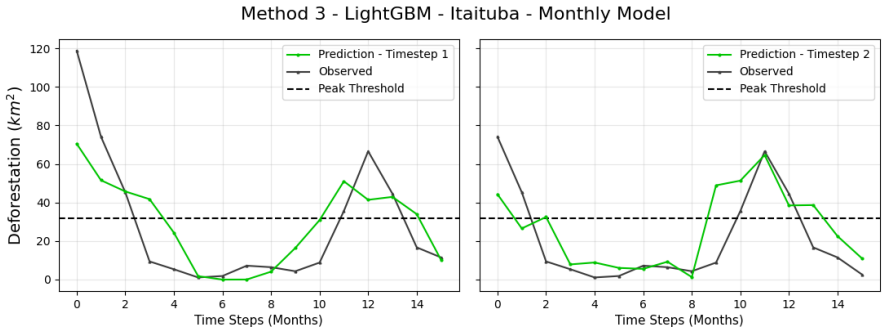
A.1.2.3 Itaituba (PA)



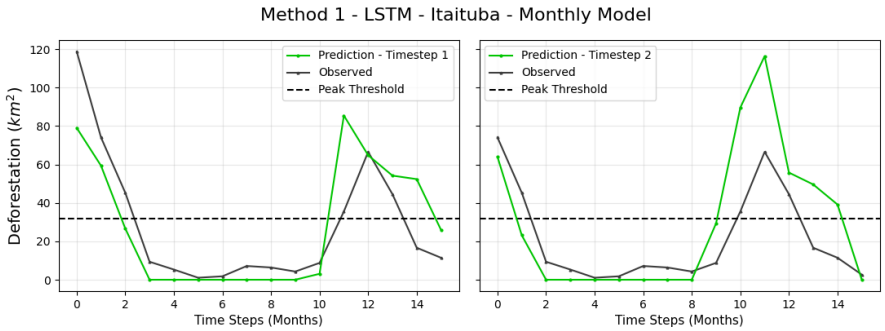
(a) Naive Model - Monthly - Itaituba



(b) Best Autoregressive Model - Monthly - Itaituba



(c) Best LightGBM Model - Monthly - Itaituba



(d) Best LSTM Model - Monthly - Itaituba

Figure A.9: Graphs of best results by algorithm of monthly models and naive model - Itaituba (PA).