ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Hidden Markov Models and their application in modeling rainfall occurrence

By Konstantinos Grammenos

MSc Thesis

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
March 2025

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Hidden Markov Models and their application in modeling rainfall occurrence

Από Κωνσταντίνο Γραμμένο

Διατριβή

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα, Ελλάδα
Μάρτιος 2025

ACKNOWLEDGEMENTS

I would like to acknowledge and express my sincere gratitude to my supervisor, Prof. Panagiotis Besbeas. His consistent support and guidance throughout this MSc thesis made this trip an incredible experience for me. Furthermore, his patience, expertise, and dedication have provided me with the confidence and skills to navigate the challenges of this thesis.

In addition, I would like to extend my sincere appreciation to Dr. Dimitrios Karlis and Dr. Ioannis Ntzoufras for their exceptional guidance and dedication to the programm. Their commitment to fostering a rigorous and inspiring academic environment has been crucial to my growth.

A special thank you also to Dr. Stefanos Kechagias for his invaluable support and guidance throughout the MSc program. His encouragement and willingness to offer advice whenever needed have played a significant role in my career and my personal development.

Finally, I am deeply grateful to my family and friends for their unwavering support, patience, and understanding throughout these two years. Their constant encouragement and belief in me have been a source of strength, making this journey not only possible but also fulfilling

Abstract

Nowadays, the modeling of daily regional rainfall is an extremely significant area of research, especially in areas like Northeast Brazil, where the weather conditions are influenced by numerous and complicated oceanic and atmospheric phenomena. Accurate stochastic models are more than necessary so to identify the existing rainfall patterns and assist in water resource management and seasonal forecasts.

This thesis explores the use of Hidden Markov Models to model the daily rainfall occurrence over a 90-day season across 24 years (1975–2002) in the state of Ceará, Northeast Brazil. Firstly, the thesis begin with an introductory application to binary times series and specifically to the Old Faithful Geyser data, providing an real example on how hidden states can govern the eruption patterns. Building on this foundation, a series of homogeneous HMMs were applied in the rainfall data and 4 hidden states are identified. One pair of states represent wet versus dry conditions and the other one contains transitional states showing contrasting spatial distributions where one has higher probability of rain in northern stations and the other in southern ones.

The analysis goes on by incorporating external large-scale climate information and more specifically the GCM' s simulated seasonal mean rainfall anomaly. A Non-Homogeneous HMM (NHMM) is used to downscale daily precipitation occurrence at the ten stations and it showed that the covariate significantly affects the transition dynamics between rainfall states. Furthermore, additional models are developed but this time by incorporating the large-scale climate variable, into the observation process through a logistic regression framework. The best-performing model indicated that while each station has a different baseline rainfall pattern, large-scale climate variability exerts a coherent regional influence. Lastly, an alternative modeling approach was explored in which each year was modeled as an independent seasonal sequence, which as we observed, it may makes more sense than representing the data as a whole continuous sequence of 2160 days.

The findings of this thesis highlight the importance of the use of such models for the modeling of daily rainfall occurrence at the station level in terms of large-scale atmospheric patterns and can inform seasonal forecasts, guide climate impact assessments, and ultimately support decision-making in sectors sensitive to rainfall variability.

Περίληψη

Στην σημερινή εποχή, η μοντελοποίηση δεδομένων ημερήσιων βροχοπτώσεων αποτελεί ένα εξαιρετικά σημαντικό πεδίο έρευνας ιδίως σε περιοχές όπως η περιοχή της Ceará στην βορειοανατολική Βραζιλία, όπου οι καιρικές συνθήκες επηρεάζονται από ένα πλήθος πολύπλοκων καιρικών φαινομένων που έχουν να κάνουν με την θερμοκρασία του ωκεανού και την ατμόσφαιρα. Η ανάπτυξη κατάλληλων στοχαστικών μοντέλων είναι απαραίτητη για την αναγνώριση των υπαρχόντων μοτίβων στις βροχοπτώσεις και για την διαχείριση των υδάτινων πόρων μίας περιοχής και στην σωστή πρόβλεψη μελλοντικών βροχοπτώσεων.

Η παρούσα διατριβή διερευνά τη χρήση Κρυφών Μαρκοβιανών Μοντέλων για τη μοντελοποίηση ημερήσιων δεδομένων βροχόπτωσης κατά τη διάρκεια μιας περιόδου 90 ημερών για 24 χρόνια (1975–2002) στην πολιτεία Ceará της Βορειοανατολικής Βραζιλίας. Αρχικά, παρουσιάζεται μία εφαρμογή σε δυαδικές χρονοσειρές και συγκεκριμένα στα δεδομένα του θερμοπίδακα "Old Faithful", παρέχοντας ένα πραγματικό παράδειγμα για το πως αυτές οι κρυφές "καταστάσεις" μπορεί να διέπουν μοτίβα συμπεριφοράς των δεδομένων. Στην συνέχεια και με βάση αυτήν την πρώιμη εξερεύνηση, εφαρμόσαμε μια σειρά ομοιογενών Κρυφών Μαρκοβιανών Μοντέλων στα δεδομένα των βροχοπτώσεων και εντοπίσαμε 4 υποβόσκουσες "καταστάσεις". Ένα ζεύγος καταστάσεων αντιπροσωπεύει βροχερές έναντι ξηρών συνθηκών, ενώ όσον αφορά το εναπομείναθεν ζεύγος, παρατηρούμε μία αντιθεση στις 2 καταστάσεις, καθώς στη μία έχουμε υψηλότερη πιθανότητα βροχής στους σταθμούς που βρίσκονται βορειότερα και σαφώς χαμηλότερη σε όσους βρίσκονται νοτιότερα, ενώ στην άλλη το αντίστροφο.

Η ανάλυση συνεχίζεται με την ενσωμάτωση εξωτερικών κλιματικών δεδομένων μεγάλης κλίμακας και πιο συγκεκριμένα της προσομοιωμένης εποχικής μέσης απόκλισης βροχόπτωσης που προέρχεται από ένα μοντέλο Γενικής Κυκλοφορίας (GCM). Συνεπώς, αναπτύξαμε ένα μη Ομογενές Κρυφό Μαρκοβιανό Μοντέλο, με το οποίο συμπεράναμε ότι η κλιματική συνιστώσα που προέρχεται από το GCM μοντέλο επηρεάζει σημαντικά τη δυναμική μετάβασης μεταξύ των κρυφών "καταστάσεων". Επιπλέον, αναπτύξαμε πρόσθετα μοντέλα, ενσωματώνοντας την ίδια μεταβλητή αλλά αυτήν την φορά, στη διαδικασία παρατήρησης των δεδομένων, μέσω λογιστικής παλινδρόμησης, με το καλύτερο σε απόδηση να είναι το μοντέλο στο οποίο, ενώ κάθε σταθμός παρουσιάζει διαφορετική βασική πιθανότητα βροχής, υπάρχει παρόμοια επίδραση της μεταβλητής κλιματικών δεδομένων μεγάλης κλιμακας σε όλους τους καιρικούς σταθμούς. Τέλος, εξετάστηκε μια εναλλακτική προσέγγιση μοντελοποίησης των δεδομένων, στην οποία κάθε έτος μοντελοποιήθηκε ως ανεξάρτητη εποχική ακολουθία — κάτι που, όπως παρατηρήθηκε, έχει ελαφρώς καλύτερη απόδοση σε σχέση την αναπαράσταση των δεδομένων ως μία συνεχής ακολουθία 2160 ημερών.

Τα ευρήματα της παρούσας διατριβής αναδεικνύουν την σημασία της χρήσης παρόμοιων μοντέλων για τη μοντελοποίηση δεδομένων βροχοπτώσεων αλλά και την ανάγκη για ενσωμάτωση σε αυτά, κλιματικών δεδομένων μεγάλης κλίμακας. Τέλος, τα αποτελέσματα και τα συμπεράσματα που απορρέουν από την συγκεκριμένη ανάλυση μπορούν να συμβάλλουν σε προβλέψεις βροχοπτώσεων στην περιοχή, να βοηθήσουν στην εκτίμηση των επιπτώσεων τις κλιματικής αλλαγής αλλά και να υποστηρίξουν την λήψη αποφάσεων σε τομείς που επηρεάζονται άμεσα από την μεταβλητότητα των βροχοπτώσεων.

Table of Contents

Acronyms

AIC Akaike information criterion

BHMM Bernoulli Hidden Markov Model

BIC Bayesian information criterion

CI Confidence Interval

EM Expectation-Maximization

GCM General circulation model

HMM Hidden Markov Model

IRI International Research Institute for Climate and Society

ITCZ Intertropical Convergence Zone

MC Markov Chain

MLEs Maximum Likelihood Estimators

NHMM Non-Homogeneous Hidden Markov Model

PMF Probability mass function

SAMS South American Monsoon System

Se Standard error

SST Sea Surface Temperature

tpm transition probability matrix

List of Figures

List of Tables

Chapter 1

Introduction

Modeling rainfall occurrence is a fundamental component of hydrological and climate studies as rainfall is crucial for the ecosystems, the water resources but also for the local communities in an area. This is especially significant in the area of Northeast Brazil, in which there is extremely high interanual variability, as it is affected from plenty of numerous large-scale atmospheric and oceanic phenomena. In particular, the area is extremely sensitive to shifts in the Intertropical Convergence Zone (ITCZ) which influences the rainfall occurrences seasons. Also, changes in the atmospheric patterns and in the sea temperature can have a strong impact on these precipitation regimes.

In addition, the presence of climate change in recent years has intensified the situation in the area , having shifted rainfall patterns and caused extreme weather conditions. For instance, throughout May 2022 and extending into June, Eastern Northeast Brazil experienced catastrophic floods and landslides following exceptionally heavy rainfall which led thousands of people to leave their homes and evacuate.

The aim of this thesis is to investigate the use of Hidden Markov Models (HMMs) to model daily rainfall occurrence using a dataset from Northeast Brazil for illustration. The data taken from the work of Andrew W. Robertson, Sergey Kirshner, and Padhraic Smyth [1] correspond to ten weather cites in the state of Ceará in Northeast Brazil during the February–April wet season 1975–2002. We analyze data from 24 rain seasons, having omitted data from years 1976,1978,1984 and 1986 due to plenty of missing values. Each season resulting in a binary time series of 90 observations. We build on the analysis of Andrew W.

Robertson [1], extending their range of methods and exploring new modeling possibilities. The main goals of the thesis are:

- **To determine the optimal number of hidden states** to describe rainfall variability during the 24 seasons and intrerpet these states in terms of their meteorological significance. For instance, one of these states may be an extreme wet state with a hig probability of rain across all stations.
- **To assess the influence of external covariates such as the GCM's simulated seasonal average rainfall anomaly** on both the **transition probability matrix (tpm)** and the **observation process** to understand the impact of large-scale climate indicators on daily rainfall occurrence in the area of Northeast Brazil.
- **To compare different model specifications** to evaluate which model describes best our data and can capture the climate-driven variations in rainfall patterns in Northeast Brazil. By assessing the models that we developed, we can somehow improve the predictive power of the model. This is extremely useful in seasonal forecasting, water resource management, agricultural planning but also for enhancing response planing in case of extreme rainfalls.

Chapter 2

Hidden Markov Model Theory

2.1 Markov Chains and HMM explained

HMMs are a particular kind of independent mixture model. An independent mixture model is a statistical tool that combines multiple (two or more) probability distributions to describe complex data and is designed to account for heterogeneity within the population. Basically, it is a weighted combination of these distributions, where each distribution represents a different source from which the data may originate.

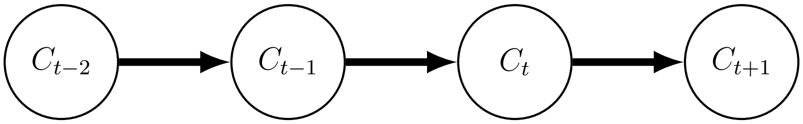
The probability distributions may be either discrete or continuous. In the HMM framework, these probability distributions are called state-dependent distributions. For example, suppose that an observation may come from a mixture of two Poisson distributions with means λ_1 and λ_2 respectively. In an independent mixture model, the choice of the distribution is independent between observations but in a HMM it depends on a Markov Chain (MC).

HMMs have found a wide use in science, as they have numerous applications in fields such as signal processing (e.g. speech, face, gesture recognition, see the tutorial of Rabiner [2]), bioinformatics where someone can find more information in the book of Timo Koski[3] which is published in 2001, in environmental studies (e.g, animal movement, see Pedersen [4]) and finance (e.g for modeling stochastic volatility in financial time series,see Langrock (2012) [5]). In general a HMM consists of two main components:

- A hidden, unobserved parameter process satisfying the Markov property, called a Markov chain (MC).
- A state-dependent process.

A MC is a stochastic process that describes a sequence of random variables which evolve over time and satisfy the Markov property. The state space of the Markov chain S is the set of values that each random variable can take. The first-order Markov property indicates that only the current state in the chain affects the future state. In other words, all the information required to predict the next state of a MC is fully captured by the most recent state of the process.

A first-order MC can be described as follows: let $C_{t-1}, C_t, C_{t+1}, \ldots$ represent the possible states which the process can occupy at any given time t, where C_t denotes the state of the Markov chain at time t. The random variables C_t are dependent on one another, as displayed in the directed graph below:



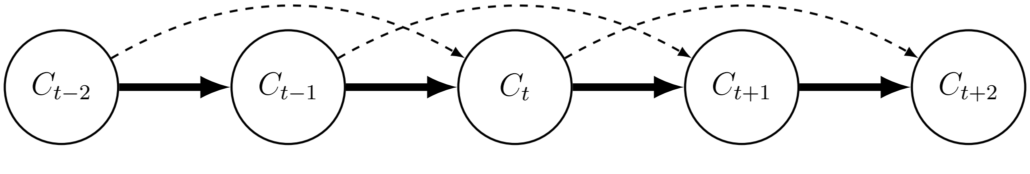
The Markov property can be mathematically expressed as:

$$P(C_t = c_t \mid C_{t-1} = c_{t-1}, C_{t-2} = c_{t-2}, \ldots) = P(C_t = c_t \mid C_{t-1} = c_{t-1}) \tag{2.1.1}$$

which means that the future state of the process depends only on its current state and not on the sequence of states that preceded it. The above type shows us a MC of first-order, but it is possible to have a higher order MC, such as a second or third order. A higher order MC ($order \geq 2$) is one in which the probability of the current state depends on more than 1 previous states. For example, in a second-order MC, the probability of the current state depends on the two most recent states, and this is expressed mathematically as:

$$P(C_t \mid C_{t-1}, C_{t-2}, \ldots) = P(C_t \mid C_{t-1}, C_{t-2}) \tag{2.1.2}$$

and we can see this dependency more clearly in the graph below:



In general a l-order Markov chain satisfies:

$$P(C_t \mid C_{t-1}, C_{t-2}, \ldots, C_{t-l}, \ldots) = P(C_t \mid C_{t-1}, C_{t-2}, \ldots, C_{t-l}), \tag{2.1.3}$$

where ($l \geq 1$) represents the order of the Markov chain. Now, the way the chain moves from one state to another has to do with a set of conditional probabilities, which are called transition probabilities. If these probabilities do not depend on the specific time t, meaning they remain the same regardless of time, then the MC is called homogenous otherwise the MC is a non-homogenous one. We will assume that our MC is homogenous and we will denote the transition probability from state i to state j as:

$$\gamma_{ij} = \Pr(C_{t+1} = j \mid C_t = i) \tag{2.1.4}$$

So, the transition probabilities of a homogenous MC which has m states can be summarized as a $m \times m$ matrix Γ, which is called (one step) transition probability matrix (tpm). Of course, as you may understand there are some constraints about the elements of the matrix Γ, such as that the $\gamma_{ij}(t)$ must be either positive or 0 and each row of the matrix must sum up to 1. In order to understand better the concept of the transition probabilities we will demonstrate an example. Imagine that the weather of a day depends only on the weather of the previous day and that it can be either sunny, cloudy or rainy. Suppose that the transition probabilities are as set in the table below:

Day t	sunny	cloudy	rainy
sunny	0.8	0.1	0.1
cloudy	0.4	0.4	0.2
rainy	0.2	0.3	0.5

So that, for example if the weather today is sunny there is 0.8 probability that the weather tomorrow will be again sunny. In our example, there are 3 states and the transition probabilities do not depend on t, so the MC is a 3-state homogenous MC.

Now, a significant and well-known property of a finite state-space homogeneous MC is the one of the Chapman–Kolmogorov equations and is expressed mathematically as:

$$\Gamma(t + u) = \Gamma(t)\Gamma(u), \tag{2.1.5}$$

which basically implies that:

$$\Gamma(t) = \Gamma(1)^t \tag{2.1.6}$$

This property is discussed in detail by Zucchini, MacDonald k [6]. This means that the transition probability matrix of t-step transition probabilities is the t^{th} power of $\Gamma(1)$, the matrix of one-step transition probabilities. $\Gamma(1)$ is equivalent to the tpm Γ which for m hidden states is given by:

$$\Gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{pmatrix}, \quad \text{with } \gamma_{ij} > 0 \text{ and } \sum_{j=1}^{m} \gamma_{ij} = 1 \quad \text{for all } i = 1, \ldots, m$$

Another feature of homogeneous Markov Chains is that the unconditional probabilities of the process being in a specific state at a specific time t are expressed like this:

$$\mathbf{u}(t) = (P(C_t = 1), P(C_t = 2), \ldots, P(C_t = m)), \quad t \in \mathbb{N}, \tag{2.1.7}$$

where $\mathbf{u}(t)$ is a 1xm row vector containing the probabilities of being in state i, i=1,..m at time t. So, as we understand from the above, if t=1 then the vector $\mathbf{u}(1)$ indicates the initial state distribution of the MC. Now, to compute the distribution at time $t+1$, given the one in time t, we postmultiply by Γ:

$$\mathbf{u}(t+1) = \mathbf{u}(t)\Gamma \tag{2.1.8}$$

Yet, an important question is if the system has a stable-state behavior through time. If it does, then we say that the MC has a stationary distribution δ, which is basically a row vector with non-negative elements with satisfying properties:

$$\delta\Gamma = \delta \quad \text{and} \quad \sum_{i=1}^{m} \delta_i = 1$$

The first equation tells us that if we apply Γ to δ, the distribution remains the same, while the second ensures that δ is a probability distribution meaning the probabilities for all states i (from 1 to m) sum up to 1. Now, computing in practice the stationary distribution is done by deleting one equation from the system of $\delta\Gamma = \delta$ and replace it by the $\sum_{i=1}^{m} \delta_i = 1$. For example, if $\delta = \left(\frac{15}{32}, \frac{9}{32}, \frac{8}{32}\right)$, this means that the probability that the MC will be at state 1 at any time is $\frac{15}{32}$. In general, the vector δ is a stationary distribution of a MC with tpm Γ if and only if:

$$\delta\left(I_m - \Gamma + U\right) = \mathbf{1} \tag{2.1.9}$$

where 1 is a 1xm vector with ones, I_m is a $m \times m$ identity matrix, and U is a $m \times m$ matrix of ones. Of course, all of the above do not hold in the case of non-homogeneous MC, and there is no stationary distribution in that case, as the transition probabilities are not constant over time and so the tpm depends on each time point t.

Finally, note that for a MC to have one and only stationary distribution, it must also be irreducible. By irreducible we mean that we can go from any state to any other state at any time t.

2.2 The likelihood of a HMM

In this section, we describe how to calculate the likelihood of a HMM efficiently. The likelihood function of a statistical model is defined as the joint distribution of the observed data. Suppose that observations $x_1, x_2, \ldots, x_T$ are generated from an m-state HMM with an initial distribution δ, tpm Γ, and state dependent distributions p_i. Here, p_i denotes the probability of observing x_t given that the hidden state at time t is i. Furthermore, it can be shown that the likelihood of how likely this sequence is produced by the HMM can be computed with $O(m^2)$ operations. In particular, the likelihood of a non-stationary first-order HMM is given by:

$$L_T = \delta \, P(x_1) \, \Gamma \, P(x_2) \, \Gamma \, P(x_3) \cdots \Gamma \, P(x_T) \, \mathbf{1}', \tag{2.2.1}$$

while the likelihood of a stationary first-order HMM can be computed by:

$$L_T = \delta \, \Gamma \, P(x_1) \, \Gamma \, P(x_2) \, \Gamma \, P(x_3) \cdots \Gamma \, P(x_T) \, \mathbf{1}' \tag{2.2.2}$$

In this notation, $P(x)$ is defined as the mxm diagonal matrix with diagonal elements $p_i(x)$, and $\mathbf{1}'$ is a column vector with m ones, where m is the number of states of the HMM. These formulations are discussed in detail by Zucchini [6].

2.3 Maximization of the likelihood

The two main approaches of maximizing the likelihood of a HMM are by direct numerical maximization of the likelihood or by Expectation-Maximization (EM) algorithm which is called by many as the Baum-Welch algorithm. The first method has several significant advantages over the second as it has the flexibility of easily fitting alternative models by adapting existing model software. On the other hand, direct numerical maximization has two important limitations that has to be considered:

1. **Numerical underflow problem**: This derives from the fact that the likelihood of the HMM during the maximization process can take extreme values, often approaching 0, causing computational issues. We have experienced this problem reportedly during the work of this thesis for the models that we will descibe in the later sections.

2. **Constraints problem**: During the process of maximizing the likelihood there are plenty of constraints that need to be taken into account. Firstly, the elements of the tpm have to be non-negative and each row of the matrix must sum up to 1, in order to ensure that the transition probabilities are valid. In addition, depending on the distribution of the observation process, there may be some state-dependent distribution constraints. For example, if the distribution is Poisson, the state-dependent means (λ_i) must be non-negative since they represent rates of events. Correspondingly, for a Bernoulli distribution observation process, the success probabilities have to lie between $[0, 1]$.

In this work, to satisfy these constraints, we apply suitable parameter transformations before the maximization of the likelihood. More specifically, we map the constrained parameter space to an unconstrained space with free parameters. In this way, we would facilitate the optimization process of the likelihood. These transformations were implemented based on the theoretical foundation and the code provided in Zucchini, Walter and MacDonald [6].

2.4 Global Decoding

In applications of HMMs, a common quantity of interest is to know the most probable sequence of states for the MC, given the observation sequence of our data. This problem is know as the " global decoding" problem and it was used a lot in speech recognition applications like Rabiner describes in his tutorial [2] and in the article of Juang and Rabiner (1991) [7]. The most widely used algorithm for the decoding problem is the dynamic programming Viterbi algorithm. Instead of trying to find the sequence of states $c_1, c_2, \ldots, c_T$ by maximizing the $P(C_t = i \mid X^{(T)} = x^{(T)})$ for each i and for each t, the Viterbi algorithm maximize this conditional probability:

$$P(C^{(T)} = c^{(T)} \mid X^{(T)})$$

, where $X^{(T)}$ represents the observation sequence. In this way the time complexity is decreased significantly from $O(|S|^T)$ in the first case to $O(T \times |S|^2)$, where T is the sequence length and S represents the state space of the MC.

2.4.1 Viterbi Algorithm

As discussed above, the Viterbi algorithm estimates the most likely sequence of states, given the observations. This specific sequence of state is also called as Viterbi path. We begin by defining:

$$\xi_{1i} = \Pr(C_1 = i, X_1 = x_1) = \delta_i p_i(x_1), \tag{2.4.1}$$

where C_t is the state at time t, X_t is the observed data at time t, δ_i is the initial probability of state i, and $p_i(x_1)$ is the emission probability of observing x_1 given state i. For $t = 2, 3, \ldots, T$, we define:

$$\xi_{tj} = \max_i(\xi_{t-1,i}\, \gamma_{ij})\, p_j(x_t), \tag{2.4.2}$$

where ξ_{tj} is the probability of the most likely path ending in state j at time t. We also store the state that maximizes this expression:

$$\psi_{tj} = \arg\max_i(\xi_{t-1,i}\, \gamma_{ij}), \tag{2.4.3}$$

where ψ_{tj} is a backtracker to the state at time $t-1$ that leads to state j at time t.

At the final time step, we determine the final state of the most probable path as:

$$C_T = \arg\max_i \xi_{Ti} \tag{2.4.4}$$

To recover the full state sequence, we use the backtrackerr matrix to trace the path backwards:

$$C_t = \psi_{t+1,C_{t+1}} \quad \text{for } t = T-1, T-2, \ldots, 1 \tag{2.4.5}$$

In the next sections and more specifically in section 2.3 we will dive deeper into how we practically implement the Viterbi algorithm using R.

2.5 Hidden Markov Models for multivariate observations

Until now, we have focused on the case of univariate observations, but there may be cases where we are interested in fitting HMMs to multivariate observations and we will describe an example of such a HMM in the later sections. These models are extremely useful especially when we have observations from multiple variables at the same time, that is multivariate time-series data. Examples of such a case can be precipitation data from multiple weather sites and medical data with multiple measurements simultaneously. In this framework, there are two kind of assumptions that can be adopted dto simplify model-fitting. Those are:

1. **Longitudinal Conditional Independence**: Consider that there is a multivariate vector of observations at time t, X_t= (X_{t1},X_{t2},...,X_{tn}). By this assumption the X_{t1},X_{t2},...,X_{tn} are mutually independent conditional on the sequence of hidden states. In the modeling of X_t, there may be different state-dependent distributions for each component of the vector, as they may represent observations of different type. For instance, X_{t1} may indicate precipitation occurrence at a location and the state-dependent distributions of X_{t1} may be Bernoulli distributions, while X_{t2} may represent daily temperature at this location and the distribution be a normal one. In general, this is the mathematical type of the state-dependent probabilities of a multivariate vector under this assumption:

   $$P(\mathbf{X}_t) = \Pr(\mathbf{X}_t = \mathbf{x} \mid C_t = i) \tag{2.5.1}$$

   So, as we understand those probabilities can vary in time.

2. **Contemporaneous Conditional Independence**: A second simplifying assumption in the HMM analysis of multivariate observations is that of contemporaneous condition independence, examples of which can be found in Robertson [1], Zucchini and Guttorp [8] amongst others. In the first paper, which will analyze thoroughly in Chapters 4 and 5 , the data are a multivariate vector $\mathbf{R}_t = (R_{t1}, R_{t2}, \ldots, R_{tq})$ at time t, where each vector is the observed rainfall occurrence sequence at time t from q weather stations. Under this assumption, each of these vectors is independent of each other at each time step. More specifically, it assumes that the components of the multivariate random vector R_t are independent conditional on the hidden state C_t at time t, so the state-dependent probability is the product of the marginals as we see in this type:

$$P(\mathbf{R}_t) = \prod_{j=1}^{q} P(R_{tj} = r_{tj} \mid C_t = i) \tag{2.5.2}$$

Here, we have to denote that these two assumptions can appear together in an application but usually there is one of the two and not both. Finally, though these assumptions make the modelling an easier and less computationally expensive, they do not remove the serial dependence between time steps and the mutual dependence between individual components. For further details and examples of HMMs for multivariate data, see MacDonald and Zucchini (1997) [9].

2.6 Non-Homogeneous Hidden Markov Models (NHMMs)

While we have already described the case and the form of the log-likelihood of a Homogeneous HMM, we have not described the Non-homogeneous case. NHHMs are widely used in fields like climate studies and hydrology for predicting rainfall occurrence patterns (see Hughes (1994) [10]), in finance for predicting interest rate regimes (see the paper of Meligkotsidou and Delaportas [11]), in speech processing for modeling phoneme transitions based on external acoustic features like in the paper of Taylor (2005) [12] and in ecology for tracking animal movements by incorporating environmental factors as covariates like Langrock (2012) [5]). In a NHMM, the model takes into account the time trend that may be present in the data, as now the transition probabilities depend on time and they are not constant over time. More specifically the transition probabilities between the hidden states depend on some covariates which of course differ depending on the application and the data. Of course, we need to ensure that for all possible values of covariates, the row-sum and the positivity constraints of the tpm are satisfied. Also, as anyone understands, by incorporating the covariates in the tpm, our model becomes more complex as we now have to estimate more parameters, but may be a more realistic one and also provide a better and more meaningful intepretation to our hidden states and be more able to capture real-world phenomena and complexities. Hughes [13] was one of the first who described the NHMM structure and plenty of different parametrizations. He used a NHMM approach to model precipitation patterns but the methodology he uses can be implemented to a wider class of applications and problems. To understand it better let's consider a MC with 2 states where

$$P(C_t = 2 \mid C_{t-1} = 1) = \gamma_{1t}, \quad P(C_t = 1 \mid C_{t-1} = 2) = \gamma_{2t},$$

and γ_{1t} and γ_{2t} depend on a covariate x_t through a logistic regression framework:

$$\gamma_{1t} = \frac{\exp(\beta_1 x_t)}{1 + \exp(\beta_1 x_t)}, \quad \gamma_{2t} = \frac{\exp(\beta_2 x_t)}{1 + \exp(\beta_2 x_t)}$$

, where β_1 and β_2 are the coefficients which show us the effect that the covariate x_t has on the transition probabilities. This is an example on how the covariates can influence the tpm. In the later sections we will illustrate an example on real data of an NHMM where the transition probabilities are influenced by covariates within a multinomial logistic regression framework.

Next, one more factor that someone should take into account when tries to fit a NHMM is the fact that because of the seasonality that there is in the model, there is no such a thing like stationary distribution. In a NHMM the initial state probabilities, that is the probability of the system to start (at t=1) in each state is denoted as:

$$\pi_i = P(S_1 = i), \quad i = 1, \ldots, k,$$

where $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ is a $k \times 1$ probability vector satisfying the constraint: $\sum_{i=1}^{k} \pi_i = 1$. So, in a NHMM we have also to estimate the initial state probabilities combined with the parameters appearing in the transition probabilities, and the parameters determining the state-dependent probabilities. Covariates can also influence the initial state probabilities which may be modeled with various ways by incorporating them and one example could be by using a multinomial logistic framework as we will see in the later sections. So now we have to estimate the parameters associated with the initial state distribution, the parameters that refer to the tpm and also the parameters regarding the state-dependent probabilities. So, for the estimation of the set of these parameters we use either EM or a numerical optimization method like optim or nlm. If we choose the EM algorithm, during the M-step we have two parts that can be optimized separately. One part invloves the state-dependent probabilities that is the emission probabilities and the other part involves the parameters that have to do with the transition probabilities and the covariates. Usually, the part of the transition probabilities has no closed-form solution and we need to implement a numerical optimization method like optim or a conjugate gradient method, while the first part has and it is less complex. However, using EM for estimating the optimal parameters can be extremelly complex and that is why in this thesis we will use a direct numerical optimization method to find them.

Furthermore, the likelihood of a NHMM is this one:

$$L_T = \boldsymbol{\delta} P(\mathbf{R}_1)\Gamma_1 P(\mathbf{R}_2)\Gamma_2 \cdots \Gamma_{T-1} P(\mathbf{R}_T)\mathbf{1}', \tag{2.6.1}$$

where Γ_t is the tpm at time t. As we understand for each time step t, there is a mxm tpm, where m is the number of states. Also R_t here are the observations at time t and δ is the initial state distribution not the stationary distribution.

In conclusion, Non-Homogeneous HMMs are much more computationally expensive as the transition probabilities must be recomputed at each time step and for a big dataset this can increase a lot the run times. In addition, at each case we have to assess if this modeling formulation and the increase in the parameters makes sense and it is worth it. We can evaluate this by comparing the famous model selection metrics AIC and BIC for which we will say more in the next section.

2.7 Model Selection

In the fitting process of numerous HMMs either homogeneous or non-homogeneous ones and in statistical modeling in general, we need to assess which model is the best choice among others. For example and we will see it later in the next sections when we fit HMMs with various number of states we have to evaluate which HMM has the best statistical fit. To do this, we need some model selection criterion. There are many of them in the literature as Celeux, G. and Durand [14] describe but we will denote the two most famous and most used. The first is called Akaike Information Criterion (AIC) and is one which belongs to the frequentist approach and aims to find the model which is closest to the true model. This criterion mathematically is expressed as:

$$AIC = -2\log L + 2p, \tag{2.7.1}$$

where $\log L$ stands for the log-likelihood of the fitted model and p for the number of parameters that need to be estimated in this model. The firs part in AIC decreasing as the model becomes more complex (for example for a HMM with more states) because as we understand the likelihood also decreases, while the second part is like a penalty for how complex is our model. The second approach is the Bayesian which wants to find the model which is most likely to be the true one and is given by:

$$BIC = -2\log L + p\log T, \tag{2.7.2}$$

where T denotes the number of observations in our dataset. It differs from AIC only in the penalty term as it grows more for large sample sizes ($T > e^2$) and because of this it favors models with fewer parameters.

Throughout all the analysis and models that we developed in this thesis, we used these two key model selection criteria to evaluate and compare them and decide which is the best one for our data.

Chapter 3

HMM for Binary Data

3.1 Old Faithful Geyser Data

The remaining chapters of this thesis will illustrate the use of HMMs for binary data from a range of real applications. In this chapter, we consider data from the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. The raw data, which are available in a data frame in library MASS in R as Geyser, consist of 299 pairs of observations collected between 1st of August and 15 of August 1985. Each pair corresponds to the waiting time (in minutes) between 2 successive eruptions and the duration time (in minutes) of the subsequent eruption. The data have been analyzed by several authors, including Azzalini and Bowman [15], who proposed a way to dichotomize the data, based on the fact that both variables exhibit a bimodal distribution.

Following the dichotomy proposed by Azzalini and Bowman, we provide a HMM analysis of binary data corresponding to the duration times of the Old Faithful Geyser. This served as a stepping stone for our subsequent analysis of multivariate rainfall occurrence data in Northeast Brazil. For example, the approach of this chapter could be used to analyze the rainfall occurrence of each station individually but in this approach we may face interpretability problems. In addition, the Geyser dataset provides a more controlled and well-understood framework, where the states are more easily interpretable and it has been extensively studied by numerous authors. Some of the applications of this dataset are the one from Aston and Martin

(2007) who applied advanced statistical analysis to discover and understand eruption patterns [16], the one from Cook and Weisberg [17] where they searched for influencial points and model fitting but also in the paper of Langrock (2012) [18], where he analyzed the data using flexible latent-state models,while also capturing the serial dependence in the data and of course also in the book of Mcdonald and Zucchini [6].

To understand the data better we made some graphs, so below we can see the distributions of the eruptions duration and of the waiting time. In the histogram of the eruptions duration, we see this separation of the data and it seems like having two groups of data, one with short eruptions, that is below 3 minutes and one with long eruptions, which are identified as those with durations equal to or greater than 3 minutes. Similarly, in the histogram of the waiting time we see that we could also dichotomize these data at the threshold of 68 minutes.

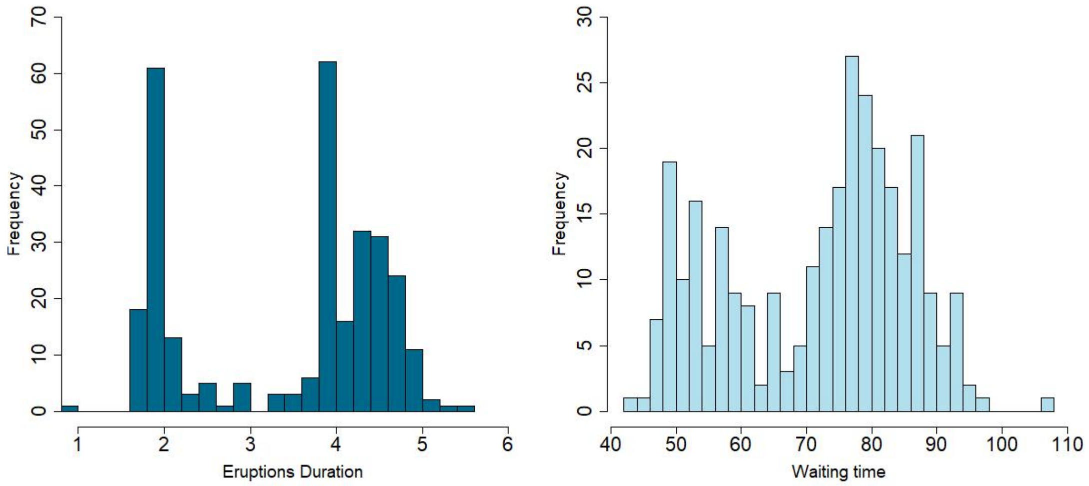


Figure 3.1: Histograms of eruption durations and waiting time for the Old Faithful Geyser data

So, after this dichotomy that we described above, the final form of the data is a vector of 299 binary values. In the final data, a value of 1 represents a long eruption, that is a large burst, while a value of 0 indicates a short eruption. The new representation of the data captures their initial structure and simplify the analysis for us, without losing any information about the eruption behavior of the Old Faithful Geyser. Below we can see the table of the new form of the data with the binary values.

So now that my data are binary, I can consider a long eruption (value 1) as success and a short one as failure (value 0) and use the Bernoulli distribution. In the context of the Bernoulli distribution, the probability p

Table 3.1: Short and long eruption durations of Old Faithful Geyser (299 obs.)

1	0	1	1	1	0	1	1	0	1	0	1	0	1	1	0	1	0	1	1	0	1	0	1	0	1	0	1	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	1	1
0	1	0	1	0	1	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	1	0	1	0	1	0	1
0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	1
0	1	0	1	0	1	0	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	1	0	1	0	1	0	1	1
0	1	0	1	1	1	1	0	1	0	1	0	1	0	1	1	1	0	1	0	1	0	1	1	0	1	1	0	1	1
1	0	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	0	1	1
0	1	1	1	0	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0
1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	

represents the likelihood of a success, where:

P(long)= $P($X = 1$)$ =p and P(short)= $P($X = 0$)$ =1 - p

The corresponding PMF of the Bernoulli distribution is:

$$P(X = x) = p^x \cdot (1 - p)^{1-x}, \quad x \in \{0, 1\} \tag{3.1.1}$$

In order to see the transition between the states 0,1 we did a frequency table, where the rows represent the previous state and the columns the current one. We also illustrated the probabilities of transition from one state to another. We can see both of the tables below.

Table 3.2: Frequency Table

X_{t-1}	$X_t = 0$	$X_t = 1$	Total
$X_{t-1} = 0$	0	104	104
$X_{t-1} = 1$	105	89	194

Table 3.3: Transition Probabilities

X_{t-1}	$X_t = 0$	$X_t = 1$
$X_{t-1} = 0$	0.00	1.00
$X_{t-1} = 1$	0.54	0.46

What we understand from the above tables is that after a short duration eruption the next eruption is always a long one, as the probability of transitioning from $X_{t-1} = 0$ to $X_t = 1$ is always 1. This indicates a deterministic pattern following short eruptions.

3.2 First-order stationary Binary Hidden Markov Models (BHMMs)

Given the binary structure of our data that we described above we decided to fit multiple first-order BHMM with several hidden states k=1,2,3,4 in order to identify the exact number of the underlying states but also what these states represent. Our binary data of the Old Faithful Geyser, which basically is a vector X_t of 299 observations, is assumed to follow a Bernoulli distribution, with the probability of success depending on the underlying hidden state C_t. So, given the simplicity of the data we will begin by assuming that the underlying factor which affects the eruption duration has only two levels, that is state 1 and state 2. For this reason our first model that we will fit will be a first-order stationary BHMM with 2 states. Now, the hidden states that are modeled as a two-state first-order Markov chain have the following components:

- The initial state distribution $\delta = (\delta_1, \delta_2)$, representing the probability of the Markov chain starting in state 1 or state 2, respectively.
- The transition probability matrix $\Gamma_{2\times2}$, which shows us the conditional probabilities of transitioning between states, such that:

  $$\Gamma_{2\times2} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}$$

  where $\gamma_{ij} = P(C_t = j|C_{t-1} = i)$ for $i, j \in \{1, 2\}$.
- The emission probabilities ϵ_{ij} which represent the probabilities of observing $X_t = j$ where j=0 or j=1 given the hidden state $C_t = i$. For our case:

  $$\epsilon_{ij} = P(X_t = j \mid C_t = i), \quad \forall i \in \{1, 2\}, \quad j \in \{0, 1\} \tag{3.2.1}$$

  Specifically, for the two hidden states:

  - $\epsilon_{10} = 1 - p_1$ and $\epsilon_{11} = p_1$, where p_1 is the probability of observing $X_t = 1$ given that the current state is 1.
  - $\epsilon_{20} = 1 - p_2$ and $\epsilon_{21} = p_2$, where p_2 is the probability of observing $X_t = 1$ given current state is 2.

For our case, which is the BHMM, the likelihood function is based on the observed binary sequence $\mathbf{X} = (X_1, X_2, \dots, X_T)$ and the hidden states $\mathbf{C} = (C_1, C_2, \dots, C_T)$. So, the likelihood L_T of the model can be expressed as:

$$L_T = P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \sum_{C_1, C_2, \dots, C_{299}=1}^{m} P(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, \mathbf{C}^{(T)} = \mathbf{c}^{(T)}) \tag{3.2.2}$$

$$= \sum_{c_1, c_2, \dots, c_{299}=1}^{2} (\delta_{c_1} \gamma_{c_1, c_2} \gamma_{c_2, c_3} \cdots \gamma_{c_{298}, c_{299}}) (p_{c_1}(x_1) p_{c_2}(x_2) \cdots p_{c_{299}}(x_{299})) \tag{3.2.3}$$

$$= \delta P(x_1) \Gamma P(x_2) \Gamma P(x_3) \cdots \Gamma P(x_T) \mathbf{1}' \tag{3.2.4}$$

So, in our process of fitting the first-order stationary two state BHMM,the parameter vector to be estimated is defined as $\hat{\theta} = (p_1, p_2, \gamma_{12}, \gamma_{21})$. As we know, p_1 and p_2 are the Bernoulli probabilities and γ_{12} and γ_{21} are the transition probabilities between the two hidden states. Because our model is stationary, the number of parameters that we need to estimate is $m + m(m-1)$, while for the non-stationary case this number would increase by $m - 1$, meaning that we would have to compute $m + m(m-1) + m - 1$ parameters, with m being the number of the hidden states. For the 2 state first-order BHMM this number is 4 and these are the parameters described above in the vector ϑ. In practice what we did to estimate this set of parameters ϑ was to maximize the log-likelihood using the Newton-Raphson optimization algorithm via the `nlm` function in R. Furthermore, in order to prevent the numerical underflow problem during the maximization process of the log likelihood we implemented several transformations:

1. **Logit Transformation for Emission Probabilities:** The emission probabilities (λ), which are the Bernoulli probabilities, were transformed into the ”working parameters” using the `qlogis` function:

$$\text{logit}(\lambda) = \log\left(\frac{\lambda}{1-\lambda}\right) \tag{3.2.5}$$

In this way, we transformed the probabilities to free parameters. Also, the reverse transformation (`plogis`) was applied to map the working parameters back to probability scale.

2. **Transformation of transition probabilities:** To ensure that the transition probabilities are properly constrained and rows of the transition matrix sum to 1, the off-diagonal elements of the transition probability matrix are normalized relative to the diagonal elements and then logarithmically transformed. So, the analytical process is as follows:

First, all the elements of each row are divided by the diagonal element of the row they belong to. This normalization ensures that the rows of the transition probability matrix sum to 1. The normalized transition probability matrix is:

$$\Gamma_{ij} = \begin{pmatrix} 1 & \frac{\gamma_{12}}{\gamma_{11}} & \frac{\gamma_{13}}{\gamma_{11}} \\ \frac{\gamma_{21}}{\gamma_{22}} & 1 & \frac{\gamma_{23}}{\gamma_{22}} \\ \frac{\gamma_{31}}{\gamma_{33}} & \frac{\gamma_{32}}{\gamma_{33}} & 1 \end{pmatrix}$$

Then, we take the logarithms of all elements of the previous matrix to ensure numerical stability and finally, the unconstrained working parameters of the tpm are:

$$\tau_{ij} = \begin{pmatrix} 0 & \log\left(\frac{\gamma_{12}}{\gamma_{11}}\right) & \log\left(\frac{\gamma_{13}}{\gamma_{11}}\right) \\ \log\left(\frac{\gamma_{21}}{\gamma_{22}}\right) & 0 & \log\left(\frac{\gamma_{23}}{\gamma_{22}}\right) \\ \log\left(\frac{\gamma_{31}}{\gamma_{33}}\right) & \log\left(\frac{\gamma_{32}}{\gamma_{33}}\right) & 0 \end{pmatrix}$$

When the optimization process ends, we get back our transition probabilities by this type:

$$\gamma_{ij} = \frac{\exp(\tau_{ij})}{1 + \sum_{k \neq i} \exp(\tau_{ik})}, \quad \forall i \neq k \tag{3.2.6}$$

The above tranformation of the tpm is for a 3×3 tpm , but the logic is the same for any number of hidden states m. After implementing these transformations, we fitted BHMMs for $m = 2, 3, 4$ hidden states. Each of these models followed the same theoretical framework and estimation process, with the number of parameters varied according to m. In the next section, we will evaluate the fitted models based on metrics like AIC and BIC and also describe a bit more the cases with 3 and 4 states.

3.3 Model Selection and results

After fitting the BHMM we calculated with the help of the nlm method in R the MLEs for the 2 state stationary first-order BHMM. So, the MLEs for ϑ are $\hat{\theta} = (p_1, p_2, \gamma_{12}, \gamma_{21}) = (0.225, 1, 1, 0.827)$., where $p_1 = 0.225$ is the Bernoulli probability of observing a long duration eruption (X_t=1) when we are in state 1, while $p_2 = 1$ shows us that it is certain that we will have a long eruption if we are in state 2. Correspondingly, if we are in state 1, the probability of observing a short duration eruption (X_t=0) is 1-p_1=1-0.225=0.775. Also, for the transition probabilities, the estimated parameters are $\hat{\gamma}_{12} = 1$ and $\hat{\gamma}_{21} = 0.827$, so the probability transition matrix is:

$$\hat{\Gamma} = \begin{pmatrix} 1-\gamma_{12} & \gamma_{12} \\ \gamma_{21} & 1-\gamma_{21} \end{pmatrix} = \begin{pmatrix} 1-1 & 1 \\ 0.827 & 1-0.827 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0.827 & 0.173 \end{pmatrix}$$

From the tpm we understand that if we are in state 1 at any time t, is certain that we will move to state 2 at the next step $t+1$. That is after a short eruption, the next eruption will always be forced by high geological activity meaning we will have a long eruption. Now, because we have said before that our BHMM model is stationary we also calculated the stationary distribution which will show us what is the probability of being in each state at any time t. So δ is this one: $\delta = \begin{pmatrix} 0.453 & 0.547 \end{pmatrix}$ and with these probabilities and the number of observations in our data we will can estimate the frequencies of the short and long duration eruptions like this:

$$\hat{X}_0 = 299 \cdot (\delta_1(1-\hat{p}_1) + \delta_2(1-\hat{p}_2)) = 299 \cdot (0.453 \cdot 0.775 + 0.547 \cdot 0) = 104.97 \approx 105,$$

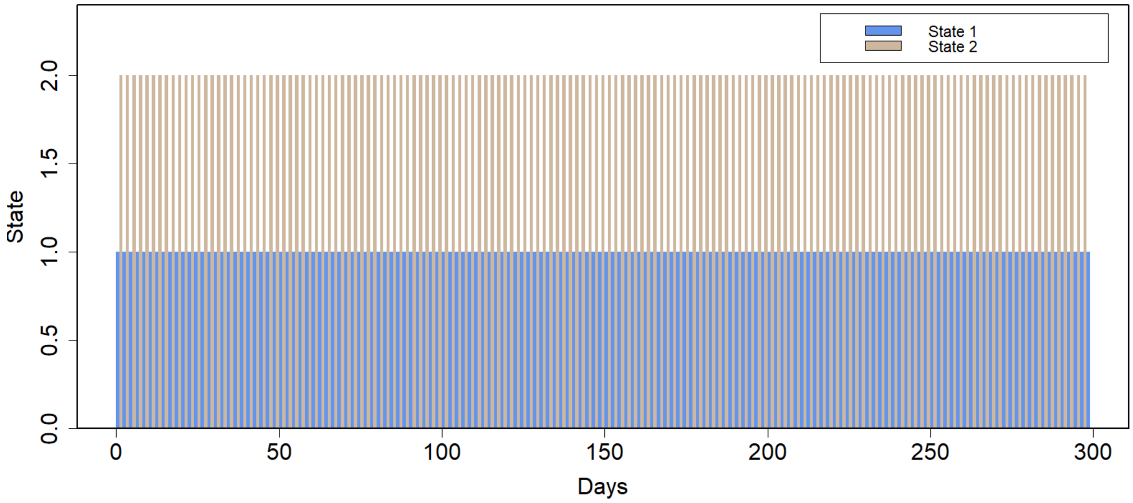
while the frequency of the long duration eruptions is:

$$\hat{X}_1 = 299 \cdot (\delta_1\hat{p}_1 + \delta_2\hat{p}_2) = 299 \cdot (0.453 \cdot 0.225 + 0.547 \cdot 1) = 194.03$$

As we see, the above frequencies are almost equal to those of the Table 3.2, which are the observed ones. Next, what we did was to calculate the most likely sequence of states that could have generated the observations, using the well-known Viterbi algorithm. The first step in the implementation of the Viterbi in R was to intialize a matrix ξ in which as we said also in the introduction, we will store the probabilities of the most likely paths to end in each of the states (two in this case) at each day t. For the first day, the forward probabilities are calculated using the formula $\xi_{1i} = \delta_i p_i(x_1)$, where $p_i(x_1)$ here is the Bernoulli probability, that is the probability of observing a long eruption given state i. Of course, we normalized these probabilities to ensure that their sum is equal to 1. Then, for each next day, we compute these probabilities using $\xi_{tj} = \max_i(\xi_{t-1,i}\,\gamma_{ij})\,p_j(x_t)$ and what we did is to consider all the possible paths that go to each state and take the one with the highest probability. Again we normalized these probabilities. Once I have calculated the forward probabilities for all the 298 days,at the final time step T, the last day, the day 299 the most likely state is identified as $i_{299} = \arg\max_i \xi_{299i}$. Identifying the state in the last day is crucial because then we will do a backtracking process to estimate the state sequence. Next step in our code is to initialize the vector `iv` where we will store the most likely states. Finally the algorithm recursively determines the most likely states for each day $t = 298, 297, \ldots, 1$ using $i_t = \arg\max_i(\gamma_{i,i_{t+1}} \cdot \xi_{ti})$,where ξ_{ti} is the forward probability for state i at time t and lastly we return the vector `iv`, which contains the most probable state sequence (Viterbi path) for our 2-state model that describes the best the Old Faithful Geyser data. In our case the Viterbi path will have the values of 1 for state 1 and 2 for state 2 which represent a long and a short duration eruption correspondingly.

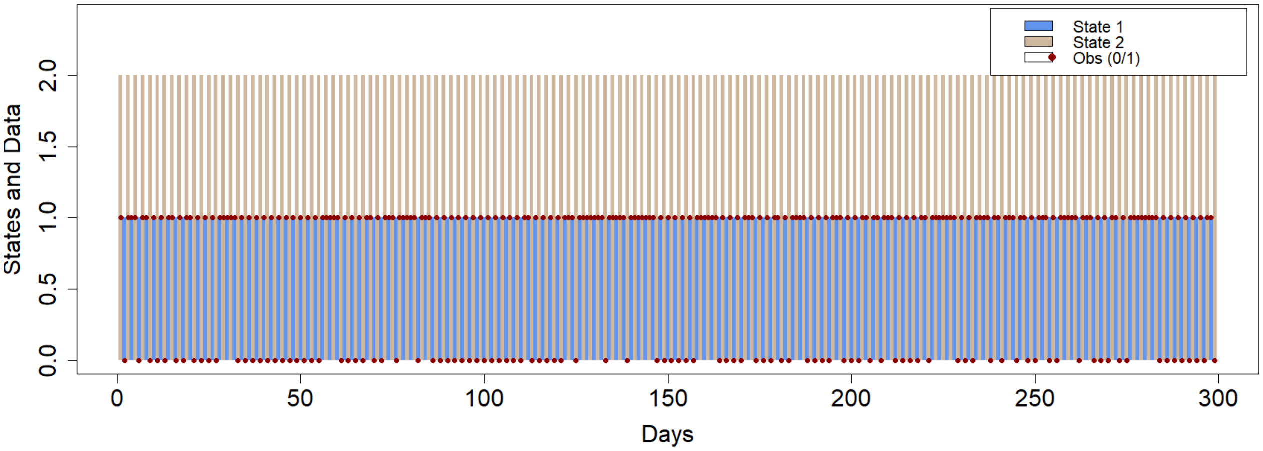
So, after running successfully the Viterbi algorithm, we constructed two plots. The first one displays the sequence of hidden states predicted by the algorithm over the 299 days and each vertical line corresponds to the most probable state at this time. Lines of blue color indicate the state 1 while the bisque color show us the second state. As we see in the plot below, the states change almost in each step.

Figure 3.2: Viterbi plot for Old Faithful Geyser data for the first-order 2 state BHMM



The second plot is again a Viterbi plot which shows the most likely sequence of states but unlike the first one this shows also the observed data, which of course can be either 0 or 1 at each time step. So the x axis represent the days and the y axis indicates the states and the observations. The observed data can be seen as red bullet points in the plot. With this new Viterbi plot, we can easily see and compare the observed eruption duration with the ones from the fitted first-order 2 state BHMM.

Figure 3.3: Viterbi plot with data for Old Faithful Geyser data for the first-order 2 state BHMM.



Using the same logic and again the `nlm` function for the optimization of the likelihood we also fitted first-order stationary BHMMs with 3 and 4 hidden states to evaluate if those models give a better fit to our data and further assure the number of hidden states that are in my data. As regards the model with the 3 states the number of parameters increase to 9 as $3 + 3(3 - 1) = 9$ and our parameter vector ϑ is this one $\hat{\theta} = (p_1, p_2, p_3, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{23}, \gamma_{31}, \gamma_{32})$. Through the Newton Raphson method we found that $\hat{\theta}_3 = (0.97, 1, 0.1, 0.52, 0, 0.09, 0.74, 0, 1)$. So, the Bernoulli probabilities are $\hat{p}_1 = 0.1, \quad \hat{p}_2 = 1, \quad \hat{p}_3 = 0.97$, the tpm Γ is given as:

$$\Gamma = \begin{pmatrix} 0.48 & 0.52 & 0 \\ 0.09 & 0.17 & 0.74 \\ 0 & 1 & 0 \end{pmatrix}$$

and the stationary distribution δ is: $\delta = (\delta_1, \delta_2, \delta_3) = (0.086, 0.525, 0.390)$.

In our effort to find the best model for our data we fitted also a BHMM with 4 hidden states. Now the number of parameters need to be estimated are 16. This of course increases the computational complexity of our model. The parameter vector is $\hat{\theta} = (p_1, p_2, p_3, p_4, \gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{21}, \gamma_{23}, \gamma_{24}, \gamma_{31}, \gamma_{32}, \gamma_{34}, \gamma_{41}, \gamma_{42}, \gamma_{43})$ and this is estimated again through nlm method. The MLEs for the Bernoulli probabilities of the 4-state BHMM are: $\hat{p}_1 = 1, \quad \hat{p}_2 = 1, \quad \hat{p}_3 = 0.556, \quad \hat{p}_4 = 0$. The estimated tpm, which in this case is a 4x4 matrix, is this one ($\hat{\Gamma}$) is:

$$\hat{\Gamma} = \begin{pmatrix} 0.106 & 0.185 & 0 & 0.71 \\ 0 & 0 & 0.702 & 0.298 \\ 0 & 1 & 0 & 0 \\ 0.84 & 0.16 & 0 & 0 \end{pmatrix}$$

and the stationary distribution (δ) of the model is: $\delta = (\delta_1, \delta_2, \delta_3, \delta_4) = (0.244, 0.291, 0.204, 0.260)$. Finally, the following table summarizes the most important information for each of the models fitted. In this table, we can see for each model, the number of estimated parameters, the negative log-likelihood ($-\log L$),the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

Table 3.4: First-order stationary BHMMs

Model	**Number of Parameters**	**-logL**	**AIC**	**BIC**
First-order 2 state BHMM	4	127.31	262.62	277.42
First-order 3 state BHMM	9	126.84	271.68	304.99
First-order 4 state BHMM	16	123.89	279.79	338.99

What we observe from the Table 3.4 is that as the number of states and the parameters increasing, the value of the log-likelihood is decreasing, but the AIC and BIC values are increasing, resulting to ”worse” models that the one with the 2 states. So, we conclude that the hidden factor that affects the duration of Old Faithful’s Geyser eruptions has only two levels.

3.4 First-order non-stationary Binary Hidden Markov Models (BHMMs)

Now, because of the inherent variability in the eruption patterns, we thought that we should also fit non-stationary BHMMs which may capture better the Old Faithful Geyser data . In the non-stationary models, there is no stationarity assumption, meaning that the system does not converge to a specific distribution.

As anyone can understand, this may be closest to plenty of real world problems and scenarios. So, in this case, there is no stationary distribution δ and also the number of the parameters that we should estimate in a model increases by $m-1$ so they are $m+m(m-1)+m-1$.

Of course, also here in the non-stationary case we used the same functions in order to transform our parameters, that is the bernoulli probabilities $lambda$ are transformed using the `qlogis` function in order to be out of constraints and then again transformed them back to probabilities using the `plogis` and the transition probabilities $Gamma(t)$ are parameterized in the form that we described above. Lastly, now we can not derive δ from Γ as the tpm is not constant over time and we have to estimate the initial state distribution.

3.5 Results and model selection for first-order non-stationary BHMMs

After fitting the non-stationary models we would like to evaluate their performance and their results. So, first of all we fitted the model with 2 states in which we have to estimate 5 parameters and the parameter vector is $\theta = (p_1, p_2, \gamma_{12}, \gamma_{21}, \delta_2)$. Now, we have to estimate 1 more parameter compared to the same stationary model, and this parameter is δ_2 as we have to explicitly estimate the parameters of δ as model parameters. So, after fitting the model with 2 states, we found these MLEs using again a Newton-Raphson method and the results are: $\hat{p}_1 = 1, \quad \hat{p}_2 = 0.225$, the tpm is:

$$\Gamma = \begin{bmatrix} 0.171 & 0.829 \\ 1 & 0 \end{bmatrix}$$

and the initial state distribution δ is: $\delta = [1, 0]$, which shows us that at time t=0 which is the start point, the MC was in the first state. Following this model, we fitted one with 3 states where the parameter vector is $\hat{\theta} = (p_1, p_2, p_3, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{23}, \gamma_{31}, \gamma_{32}, \delta_2, \delta_3)$ and the MLEs which we found are: $\hat{p}_1 = 0.104, \quad \hat{p}_2 = 1, \quad \hat{p}_3 = 1,$, which means that for states 2 and 3 a long eruption is certain, while when the system is in state 1, the probability of a high duration eruption is pretty low. The tpm Γ is:

$$\Gamma = \begin{bmatrix} 0.174 & 0.083 & 0.743 \\ 0.541 & 0.459 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

, while the initial state distribution is: $\delta = [1, 0, 0]$, which indicates that the MC starts in state 1, suggesting that the Geyser starts with a short eruption. Finally, we fitted a non-stationary model with 4 states, where the parameters need to be estimated increase significantly to 19. So, the parameter vector is this one: $\theta = (p_1, p_2, p_3, p_4, \gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{21}, \gamma_{23}, \gamma_{24}, \gamma_{31}, \gamma_{32}, \gamma_{34}, \gamma_{41}, \gamma_{42}, \gamma_{43}, \delta_2, \delta_3, \delta_4)$ and the MLEs are: $\hat{\theta} = (0.599, 1, 0.013, 0, 0, 1, 0.112, 0.658, 0.092, 0, 1, 0, 0.681, 0, 0.319, 1, 0, 0)$. So, the tpm is

$$\Gamma = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0.112 & 0.139 & 0.658 & 0.092 \\ 0 & 1 & 0 & 0 \\ 0.681 & 0 & 0.319 & 0 \end{bmatrix}$$

and the initial state distribution is: $\delta = [0, 1, 0, 0]$, meaning that at t=0 the MC is in state 2 so there is a long eruption. Again as also in the stationary case, we compared those 3 models based on AIC and BIC and the value of the log likelihood. We can see the results in the table below:

Table 3.5: First-order Non-stationary BHMMs

Model	**Number of Parameters**	**-logL**	**AIC**	**BIC**
First-order 2 state Non-Stationary BHMM	5	126.71	263.42	281.92
First-order 3 state Non-Stationary BHMM	11	126.21	274.43	315.13
First-order 4 state Non-Stationary BHMM	19	123.51	285.01	355.32

Firstly, we observe that the number of parameters need to be estimated is increasing a lot as we go from 2 to 4 states, something that we expected. Furthermore, the model that seems to has the better fit is the HMM with the two states as this is the one with the lowest BIC and AIC values and small difference in the log-likelihood. Previously, we observed that the same is true for the stationary case. So. comparing these two models, the stationary 2-state HMM and the non-stationary one with the same number of states, we see extremely small differences between the log-likelihood, AIC and BIC, indicating that non-stationarity assumption does not lead to any significant better fit, which means that the process influencing the eruptions remain stable over time.

Chapter 4

Modeling rainfall occurrence over Northeast Brazil using a baseline Homogeneous HMM

4.1 Rationale for using a baseline HMM

Rainfalls in Northeast Brazil follows a complex pattern due to the peculiarity of the area and the numerous phenomena that influence its weather. For this reason, we decided to try to model the rainfall occurrences with a much more difficult way than some simple statistical model. So, we fitted various Hidden Markov Models with our main objective to be the identifying of the hidden states that better describe the rainfall occurrences and of course the interpretation of those. For example, we would like to learn if a state corresponds to a dry or a wet one and of course the probability of rain in each of those weather stations given the state. By knowing this, we could predict future rainfall patterns in the area. Firstly, we began the modeling with a simple Homogeneous HMM which will work as a baseline model throughout our analysis. We started with this model so to identify the hidden rainfall states that may exist and assess their stability over time.

4.2 Study area and rainfall data

The dataset consists of 2160 daily observations across 10 weather stations in the region of Ceará in Northeast Brazil. Specifically the stations are: (1) Acopiara (317 m), (2) Aracoiaba (107 m), (3) Barbalha (405 m), (4) Boa Viagem (276 m),(5) Camocim (5 m), (6) Campos Sales (551 m), (7) Caninde (15 m), (8) Crateus (275 m), (9) Guaraciaba Do Norte (902 m), and (10) Ibiapina (878 m). Observations are binary, that is either 1 or 0 with 1 representing rain and 0 no rain. More specifically, the 2160 days refer to the period February–March–April of each year for the years of 1975 to 2002. These 3 months consist the peak rainy season for this area in Brazil. The years of 1976,1978,1984 and 1986 were omitted because of plenty of missing values. So, as we understand our data consist of a 90 day period for each year for 24 years and this is how at the end we have 2160 daily observations. The data were provided by FUNCEME (Fundação Cearense de Meteorologia e Recursos Hídricos). Below, we illustrated two maps, one which shows the area of Brazil with highlighted the state of Ceará and one with a detailed view of this state with the 10 weather stations of our study with red.

Figure 4.1: Geographic Location of Ceará in Brazil

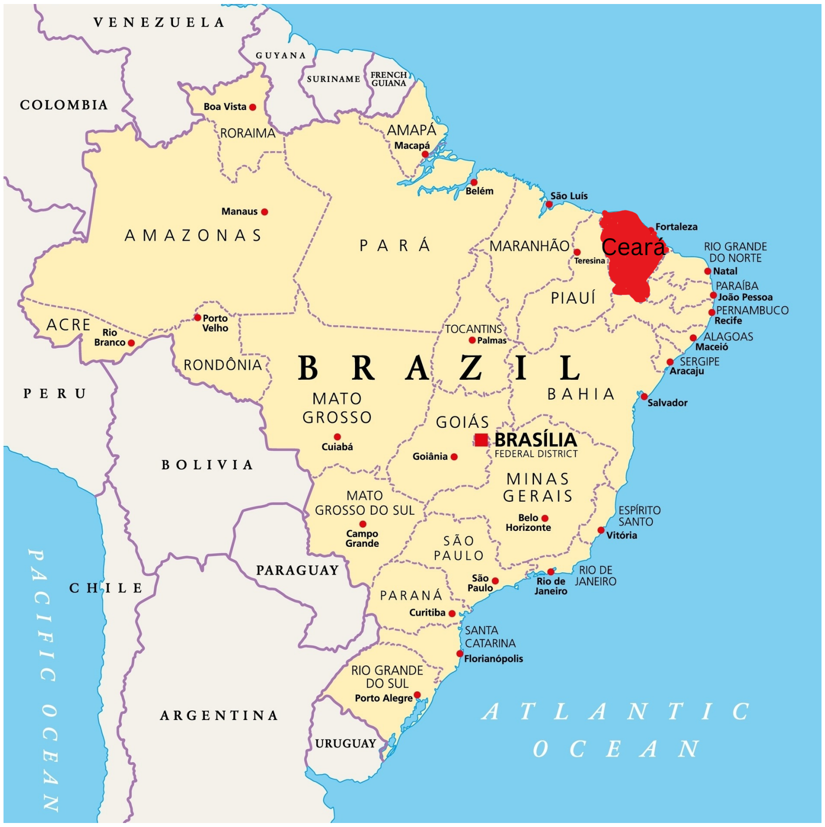
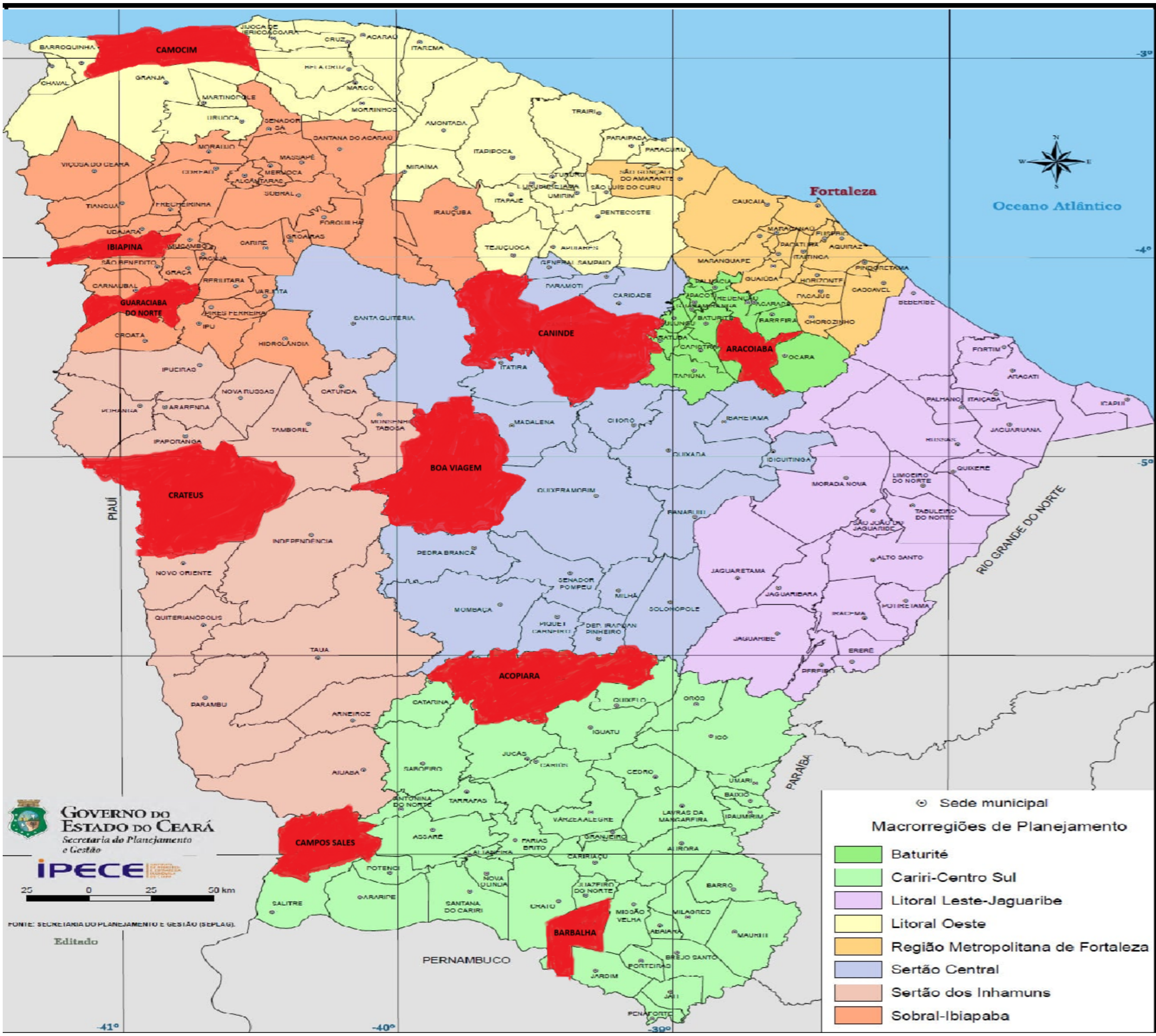


Figure 4.2: Ceará Map with Highlighted the 10 Weather Stations



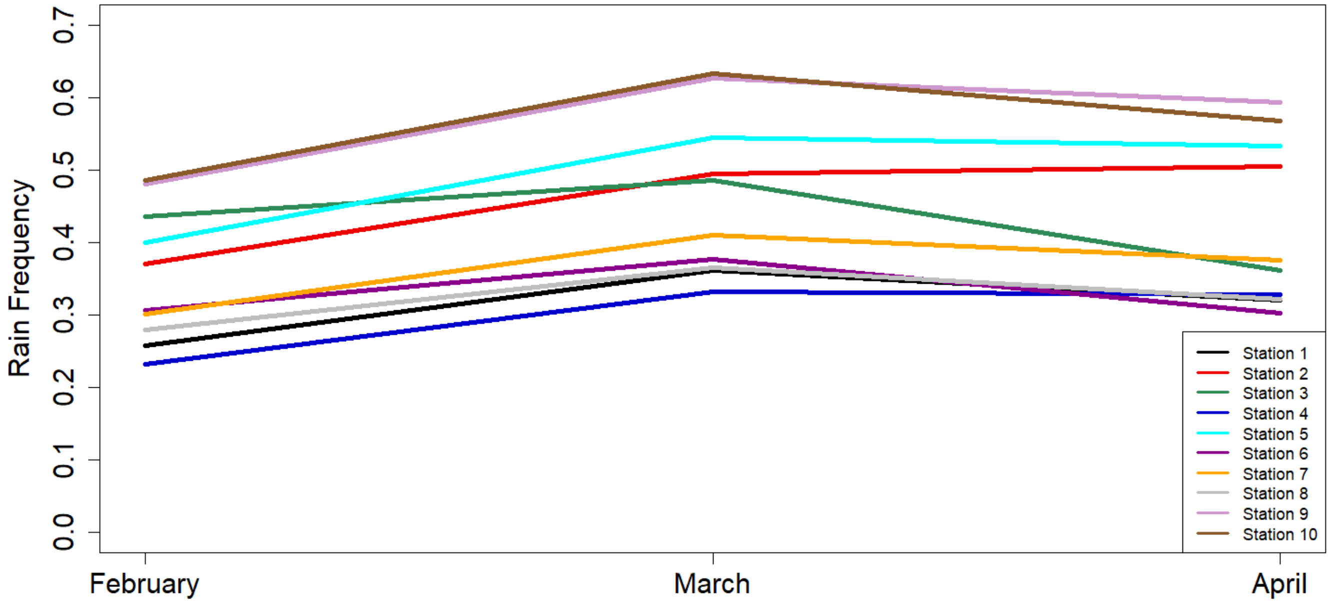
To learn more about our data we calculated some descriptive statistics. For instance in the the Table 4.1 below we can see the frequencies of rain for the total of days at each of the 10 weather stations. By this table, we understand that there is a notable discrepancy between the cites with some of them like in station 2, Aracoiaba (107 m), station 9, Guaraciaba Do Norte (902 m), and 10, Ibiapina (878 m) having significantly high number of rainy days while others like station 4, Boa Viagem and station 1, Acopiara show substatianly lower rainfall frequencies.

Table 4.1: Frequency table of days with no rainfall and rainfall across the 10 stations

Rainfall	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7	Station 8	Station 9	Station 10
Rain ($R_t = 1$)	677 (31.34%)	987 (45.69%)	924 (42.78%)	642 (29.72%)	1064 (49.26%)	710 (32.87%)	782 (36.20%)	696 (32.22%)	1224 (56.67%)	1215 (56.25%)
No Rain ($R_t = 0$)	1483	1173	1236	1518	1096	1450	1378	1464	936	945

Following the above table, we also did a plot to illustrate the rain frequencies of each of the 10 stations on a monthly basis throughout the 24 seasons-years. Each line represents a different station and in this way we can highlight both monthly and spatial (station-level) variability in rainfall occurrences. From the Figure 4.3 it is very clear that March is the wettest month with the biggest proportion of rainy days and that rain frequency decreases from March to April for most of the stations. This is not true only for the stations 4,5 and 9 which correspond to the areas of (4) Boa Viagem (276 m),(5) Camocim (5 m) and (9) Guaraciaba Do Norte (902 m) correspondingly. For these districts we observe that the rain frequency is increasing from February until March and then it remains relatively stable until April. Furthermore, stations of (1) Acopiara (317 m), (4) Boa Viagem and (8) Crateus (275 m) have the lowest rain frequencies while (9) Guaraciaba Do Norte (902 m), and (10) Ibiapina (878 m) have the highest of all may due to their high altitude.

Figure 4.3: Monthly Rain Frequency for the 10 Stations during the Rainy Season



Additionally to exploring the distribution of the monthly rain frequencies over the 24 years we wanted to find out the longest consecutive rainy and dry periods across all 2160 days for each cite and in this way understand better the rainfall patterns of our data. So from the Table 4.2 we observe that the 2 stations with the longest wet periods are station 10, Ibiapina, (878 m) which experienced 28 consecutive days of rainfall and station 9, Guaraciaba Do Norte (902 m) with 23 consecutive rainy days. This finding aligns with the picture that we see in the graph 4.3. Meanwhile, all the other stations experience pretty similar consecutive wet periods, ranging between 10 and 17 days. Now as regards the dryness, we observe that station 1, Acopiara (317 m) and station 4, Boa Viagem have an extremely long dry period of 37 days both. Notably, station 7 Caninde (15 m) has the biggest dry period of 50 days, which indicates extreme drought conditions in the area.

Table 4.2: Maximum Consecutive Rainy and Dry Days for Each Station over the 2160 days

Metric	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7	Station 8	Station 9	Station 10
Max Consecutive Rainy Days	8	13	17	10	15	13	13	13	23	28
Max Consecutive Dry Days	37	33	26	37	22	28	50	37	21	22

4.3 The Homogeneous Hidden Markov Model

Now, we will describe how we fitted a Homogeneous HMM (HHMM) in the multivariate binary rainfall data for the 10 weather stations in the Northeast Brazil. As we have said before, by homogeneous we mean that the transition probabilities between states remain constant over time. First of all, our data has the form of a multivariate random vector $R_t = (R_t^1, R_t^2, \ldots, R_t^{10})$, where each element in this vector indicates the rainfall occurrence across each of the 10 stations at day t. This notation is similar to the one that Hughes uses in their paper of the Non-Homogeneous HMM [19]. Of course each R_t^i takes values of 0 for no rain and 1 for rain. Also, we will symbolize as S_t the hidden state of our HMM. So, first we need to define that our HMM will be a first-order HMM which means that the current state S_t at time t depend only on the previous hidden state S_{t-1} at time $t-1$ and not on other previous hidden states. This is expressed mathematically as:

$$P(S_t \mid S_{1:t-1}) = P(S_t \mid S_{t-1}) \tag{4.3.1}$$

Second, the multivariate observations R_t are conditionally independent from all other variables, given the weather state at time t and only depend on the current hidden weather state S_t. So, each observation R_tj at day t and station j is conditionally independent of the observations at other stations, given the hidden state S_t at day t. This assumption is called contemporaneous conditional independence (see section 1.4) and mathematically this is translated as:

$$P(\mathbf{R}_t \mid S_{1:t}, \mathbf{R}_{1:t-1}) = P(\mathbf{R}_t \mid S_t) \tag{4.3.2}$$

These two assumptions are shown and explained also in the paper of Hughes and Guttorp 1994[10]. So, as we understand and compared to the analysis that we did on the Old Faithfull Geyser data, which was implemented in a univariate level, here our study involves a multivariate binary time series. This adds complexity and means that we need to extend our methodology and fit a range of multivariate binary HMMs to appropriately model our data. Yet, before going to the fitting process, we have to refer that there are two ways that someone can fit those multivariate HMMs. One way is to fit a dependent multivariate distribution in each of the hidden states of the model and the second one which we will use is by assuming contemporaneous conditional independence. This term is also discussed in a similar case of ours in the paper of Zucchini and Guttorp (1991) [20]. By this term, we mean that the rainfall occurrence at station i at time t is conditionally independent from the rainfall occurrence at any other station, given the hidden state S_t. So for example, given the state, the probability that a rainfall will occur at stations 1,2,7,9,10 and not at stations 3,4,5,6,8 will be the product of the marginal Bernoulli probabilities:

$$\pi_{i1}\pi_{i2}(1-\pi_{i3})(1-\pi_{i4})(1-\pi_{i5})(1-\pi_{i6})\pi_{i7}(1-\pi_{i8})\pi_{i9}\pi_{i10}$$

Thus, the state-dependent probabilities are:

$$P(\mathbf{R}_t = \mathbf{r} \mid S_t = s) = \prod_{m=1}^{10} P(R_t^m = r^m \mid S_t = s) = \prod_{m=1}^{10} p_{smr}, \tag{4.3.3}$$

where r is the binary value in our data (0 or 1) which indicates if there was a rainfall or not in that day, m is the number of stations-columns in our dataset and the s represents the state. So, by this notation, where we account for station-specific Bernoulli probabilities p_{smr} is the probability of having a rainfall occurrence or not (r=1 or r=0) at a station m when we are in state s. To be more specific and to understand better the observation process we will illustrate the example of one station. Let' s say that we have identified 4 states (s=1,2,3,4) and that in station 1 and at day 1 the process is in state 1 and r=0 which means no rain, then the probability of this day will be $1-p_{s=1,m=1,r=1}$, where p_{111} is the Bernoulli probability of rain at station 1 given that the process is in state 1. The next day t=2, the process transits to state 2 and there is a rainfall on that day (r=1) then the probability will be $p_{s=2,m=1,r=1}$,which is the Bernoulli probability of rain of station 1 given that we are in state 2. We will follow the same logic for the 2160 days of our dataset, so at the end we will have to calculate the product of all daily probabilities over the sequence of observed rainfall and the corresponding hidden states. Each station has a different set of Bernoulli probabilities π_1,π_2,π_3,π_4 if we assume 4 states, so because we have 10 weather stations, we will have to calculate 10xk, in that case 40 Bernoulli probabilities, each one indicating the probability of rain at each station given the state.

As regards now the likelihood $L(\Theta;\mathbf{R})$ of the homogeneous HMM it has the form of:

$$L(\Theta;\mathbf{R}) = \boldsymbol{\delta} P(\mathbf{R}_1)\Gamma P(\mathbf{R}_2)\Gamma \cdots \Gamma P(\mathbf{R}_T)\mathbf{1}', \tag{4.3.4}$$

where δ is the initial state distribution (a kx1 raw vector) and here it is also the stationary distribution because our model assumes stationarity, Γ is the tpm (kxk matrix) and is constant,1' is a column vector of ones which is used to sum over the states and $P(\mathbf{R}_1)$ for example indicates the likelihood of observing the rainfall sequence across all 10 weather stations at day 1 for each hidden state. The type of $P(\mathbf{R}_1)$ is given by:

$$P(\mathbf{R}_1) = \text{diag}\left(P(\mathbf{R}_1 \mid S_1 = 1), P(\mathbf{R}_1 \mid S_1 = 2), \ldots, P(\mathbf{R}_1 \mid S_1 = k)\right), \tag{4.3.5}$$

where $P(\mathbf{R}_1 \mid S_1 = s)$ is the probability of observing the rainfall sequence for all stations at day 1, given the hidden state s. This rainfall sequence is basically a vector of zeros and ones depending if it rained or not this specific day, at each station. Each diagonal element in this matrix, is computed as shown in Equation 4.3.3. Same logic for the $P(\mathbf{R}_2)...P(\mathbf{R}_T)$.

Finally, we have to say that by assuming contemporaneous conditional independence the model becomes simpler and also computationally less expensive.

4.4 Parameter estimation and results of the Homogeneous HMM with shared Bernoulli probability across stations

First of all, although it does not seem a really logical and realistic assumption, we assumed that there is the same Bernoulli probability across all 10 stations. We did this so to test and confirm the station-specific rain probabilities assumption. So, under the shared across stations rain probability assumption, we fitted several multivariate homogeneous and stationary HMMs with a variety of hidden states ranging from 2 to 5 so to identify the exact number of hidden states that better describes the rain patterns in the area and evaluate the interpretability of the states. This type of model is called as model "Mo" in most cases and especially in capture-recapture applications. In the Mo model each individual and in the most examples each animal in the population has the same probability of being captured during each sampling occasion, so there is no heterogeneity and those probabilities remain constant throughout all samplings. Someone can learn more about this model in capture-recapture studies described in the book of McDonald and Amstrup[21]. So, more specifically, in our example now the Equation 4.3.3 is being simplified because of $p_{sm} = p_s$ for all m

to this one:

$$P(R_t = r \mid S_t = s) = \prod_{m=1}^{10} p_s^{r^m} (1 - p_s)^{1-r^m} \tag{4.4.1}$$

From the equation (4.4.1), we understand that the same Bernoulli probability p_s applies to all stations and in this way the number of Bernoulli probabilities reduces from $m \times$ stations (e.g., if 4 states to $4 \times 10 = 40$) to just m, that is one per state. This reduces significantly the computational burden as now we have significantly less parameters to estimate in the log likelihood optimization but as we said this is not a valid assumption, as the stations refer to weather sites far away from each other and with different characteristics.

We implemented this model by constructing our own functions to transform the parameters into working parameters so as to prevent the numeric underflow, as we did when fitting the Old Faithful Geyser data. More specifically, we used the qlogis to transform the vector of the Bernoulli probabilities to a working free parameter and plogis to transform it back to a valid probability bounded from 0 to 1 and we implemented the transformation for gamma as described in the Section 3.2. Furthermore, we made a function in which we calculated the log-likelihood of the observed data by using the forward algorithm. The first step in this function was to transform the estimated parameters to natural scale by calling the function that does the transformations we described earlier. Then, the forward probabilities are initialized at $t = 1$, where the probability of observing the rainfall occurrence vector at day 1, that is R_1 given a hidden state s is computed by this product:

$$P(R_1 \mid S = s) = \prod_{j=1}^{10} p_s^{r_{1,j}} (1 - p_s)^{1-r_{1,j}} \tag{4.4.2}$$

, where $r_{1,j}$ is the rainfall occurrence (1 for rain and 0 for no rain) at station j at day 1 and p_s is the probability of rain occurring at any station, given state $S = s$.Since we have 10 weather stations, we know that if there was a rain in k of them, this simplifies to:

$$P(R_1 \mid S = s) = p_s^k (1 - p_s)^{10-k} \tag{4.4.3}$$

Using δ and the Bernoulli probabilities computed above, the forward probabilities are initialized as:

$$\alpha_1(s) = \delta(s) P(R_1 \mid S = s), \quad \forall s \in \{1, \ldots, m\} \tag{4.4.4}$$

To avoid numerical underflow, we also implemented a normalization to ensure that the forward probabilities remain in a numerically stable range. Next for each subsequent time step $t \geq 2$, the forward probabilities are recursively updated. First, the probability of the rainfall occurrence vector at day t, R_t given state s is computed as:

$$P(R_t \mid S = s) = p_s^k (1 - p_s)^{10-k} \tag{4.4.5}$$

and the forward probabilities for the next steps are computed as:

$$\alpha_t(s) = \sum_{s'} \alpha_{t-1}(s') \Gamma P(x_t \mid S = s), \quad \forall s \tag{4.4.6}$$

As we understand, this computation involves products of many small probabilities over time so normalization is needed again. At the end, after processing all observations up to $t = T = 2160$ we will have the total log-likelihood of the observed sequence $\log P(X \mid \theta)$. After constructing the log-likelihood, we estimated the optimal parameters of the HMM by maximizing this function by using the `nlm` optimization algorithm. To do this, we used for all our parameters initial values drawn from the uniform distribution $U(0.1, 0.9)$.

So, more specifically, the total number of parameters need to be estimated is: $m \cdot 1 \ + m(m-1)$, where the first part refers to the Bernoulli probabilities and the second to the probabilities of the tpm. More analytically, below we can see the number of parameters estimated, the max log-likelihood and the BIC for each model:

	Number of Parameters	**Log-Likelihood**	**BIC**
State 2	4	-13126.52	26283.75
State 3	9	-12907.07	25883.23
State 4	16	-12866.53	25855.9
State 5	25	- 12862.09	25916.13

Table 4.3: Homogeneous multivariate HMMs with shared Bernoulli probabilities across stations

From the Table 4.3, we observe as we expected that as the number of states increases, the log-likelihood and the BIC metric is being decreased. However, this happens until one point and specifically when the number of hidden states becomes 5, where we see an increase on BIC and just a slight decrease in the likelihood. Therefore, we conclude that the 4-state model is the most appropriate for our analysis. The tpm for the 4-state model is presented in the table 4.4 and we observe that all states have high enough self-transition probability indicating some stability. In addition, transitions to state 2 are extremely rare to non-possible across all states, as shown by the smaller or zero probabilities in the second column. Finally, if the process is in state 2 there is no probability of going to state 4.

	State 1	**State 2**	**State 3**	**State 4**
State 1	0.64	0.11	0.01	0.24
State 2	0.43	0.56	0.01	0
State 3	0.07	0	0.67	0.26
State 4	0.25	0	0.17	0.58

Table 4.4: Transition probabilities for the 4-state HMM-shared Bernoulli probabilities across stations

Finally, the estimated shared Bernoulli probabilities across stations but different for each state are $p_1 = 0.825, \quad p_2 = 0.053, \quad p_3 = 0.572, \quad p_4 = 0.285$. As we observe, state $S_t = 4$ has the highest probability, suggesting that this states correspond to a ”wet” state, while $S_t = 2$ has the lowest probability ($p_2 = 0.0534$), likely representing a ”dry” regime. In addition, the initial state probabilities (δ) are as follows: $\delta = \{0.38, 0.09, 0.19, 0.34\}$ From (δ), we understand the process at the day 1 is most probable to be in state 1 or in state 4. We fitted also the same models in one extra way, by using the depmixS4 package in R and we found the same results as with our own method.

4.5 Parameter estimation and results of the Homogeneous HMM with station-specific Bernoulli probabilities

While, the previous simple approach helped us to explore the hidden states of rain in Northeast Brazil, now we would like to extend this approach to a more realistic one, in which the rainfall occurrence probabilities will be different for each station (station-specific) and not shared so we will have a matrix of:

$$\text{number of Hidden States} \times \text{number of Stations}.$$

This case is described in the Equation 4.3.3, as we now define a distinct probability $p_{s,j}$ for each state s and each station j. In this way we account for the spatial variability that may exist in rain patterns across the 10 different weather cites. Of course also here we followed the same methodology that we described in the Section 4.4.

So, first we fitted the models to the whole data and we derived some metrics like the maximum log-likelihood and BIC. We can observe those and the number of parameters that we needed to estimate for each model in the table below:

	Number of Parameters	**Log-Likelihood**	**BIC**
State 2	22	-12538.59	25253.76
State 3	36	-12272	24835.76
State 4	52	-12165.15	24752.58
State 5	70	-12120.42	24809

Table 4.5: Homogeneous first-order stationary multivariate HMMs with station-specific probabilities

First of all, we have to denote that by comparing the log-likelihood and the BIC values of the Table 4.5 with the ones from the Table 4.3 where we assumed shared rain probability across stations, we conclude and confirm that rainfall occurrence probabilities vary across different stations, as this model leads to a much better statistical fit.

Now, to better understand how the number of parameters derives, we will see a case of one model. For example for the model of the 2 hidden states, the parameters that we need to estimate their MLEs are 22 as:

1. the transition matrix has 2 rows as it is a 2x2 matrix and 1 free parameter per row because each row sums up to 1, so we have two parameters which we have to estimate.
2. for each hidden state I have to calculate one Bernoulli probability for each station, so because we have 10 stations and 2 states, we need to calculate the MLEs of 20 Bernoulli probabilities. Thus, in total we have to calculate 22 parameters.

With the same logic someone can derive the number of the parameters need to be estimated for the models with more than 2 states. So, from the results in Table 4.5, we observe that as the number of states increases the log-likelihood decreases substantially indicating a better fit. Yet, we see the same trend for the BIC values but this stops with the 4-state model. More specifically, BIC value in the 5-state model is slightly increased indicating no further improvement. So, we understand that there are 4 hidden states that govern the rainfall dynamics through the area of Northeast Brazil.

Now, while these metrics are useful to evaluate the performance of the fitted HMMs, we wanted to check a more robust evaluation strategy and thats why we did a cross-validation fitting process like the paper of Andrew W. Robertson, Sergey Kirshner, and Padhraic Smyth.[1]. So, what we did was to split our data to 4 folds where each fold is consisted of 6 consecutive years, that means 540 days. Remember that we have 2160 days in our dataset where for each year there are 90 days representing the days from February until the end of the March. We fitted the different models from K=2 to K=5 states in the training data, that is the 3 out of the 4 folds leaving out one different fold at each time and then we calculated some normalized metrics regarding the log-likelihood of the test data and the BIC. More specifically, as regards the normalized log-likelihood, what we did was to calculate the total log-likelihood for the test folds across all iterations and then divide this sum by the total number of binary events in our data, which is $N = 24 \times 90 \times 10 = 21,600$. Now, as regards the normalized BIC, we compute the BIC as in the Equation (2.7.2), where Ł is the total log-likelihood of the model evaluated on the test data, p is the number of parameters in the model and T is the number of days in the training data. Finally we normalized it as:

$$N_{BIC} = -\frac{BIC}{2 \times N} \tag{4.5.1}$$

As we observe the value of these metrics for the multivariate HMMs with 2, 3, 4, and 5 states, in the Table 4.6, we understand that also from this evaluation strategy the result is that we choose again the 4-state model because it exhibits the optimal trade-off in terms of both normalized metrics.

	Normalized Log-likelihood	**Normalized BIC**
2-state model	0.586	0.589
3-state model	0.575	0.581
4-state model	0.572	0.580
5-state model	0.570	0.582

Table 4.6: Normalized Metrics after Cross Validation

After our conclusion that the 4-state model is the best choice among the others, we derived first the initial state distribution (stationary) and of course the transition probabilities from one state to another. So, the stationary distrubition for the 4-state model is this one:

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4) = (0.2948, 0.2599, 0.2899, 0.1552)$$

What we understand is that the process spends almost equal time in states 1,2 and 3, while state 4 seems to be the least persistent state. This aligns with the results of the transition probabilities, as the state 4 is the one with the smallest self-transition probability among all others.

So, now, regarding the transition probabilities in the Table 4.7 we see that there are very high self-transition probabilities for states 1,2 and 3 which means that if the process is one of these three states it will probably remain like this indicating a stable weather pattern. In contrast, state 4 may be the only one with less persistence and seems to be an intermediary state with equal probability of going to state 1 and state 2 and a bit smaller one for going to the third state.

	State 1	**State 2**	**State 3**	**State 4**
State 1	0.70	0.02	0.18	0.10
State 2	0.03	0.68	0.17	0.13
State 3	0.17	0.15	0.61	0.07
State 4	0.21	0.21	0.12	0.46

Table 4.7: Transition probabilities for the 4-state HMM.

After reviewing the transition probabilities for the 4 states, it is the time to check the Bernoulli probabilities of rainfall occurrence for each of the 10 stations. So, from the Table 4.8 we observe very high probabilities of rain during state 1, across all stations and especially in station 2 Aracoiaba (107 m), station 9 Guaraciaba Do Norte (902 m) and station 10 Ibiapina (878 m). Contrarily, in state 2 all we see are dry conditions with extremely small probability of a rainfall almost in all stations except the station 5 which is the area of Camocim and where we observe a moderate probability of rain. This station is pretty close to the coast and benefits from moist air masses that come from the Atlantic Ocean. Finally, regarding state 3 we observe the highest probability of rain in northern stations such as in Aracoiaba, Camocim, in Guaraciaba Do Norte and Ibiapina area and lower probabilities in most of the south ones like in Acopiara, Barbalha and in the area of Campos Sales. For state 4 we see a similar pattern in the probabilities of rainfall but somehow inverted with what happening in state 3, as we observe high probabilities in most of the south stations and low probabilities in the north stations.

State	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7	Station 8	Station 9	Station 10
1	0.64	0.79	0.77	0.61	0.68	0.64	0.69	0.63	0.90	0.87
2	0.04	0.08	0.11	0.03	0.21	0.06	0.05	0.03	0.11	0.10
3	0.22	0.53	0.20	0.26	0.61	0.11	0.41	0.31	0.67	0.69
4	0.32	0.31	0.75	0.22	0.39	0.60	0.17	0.26	0.51	0.50

Table 4.8: Bernoulli probabilities of rainfall for each state and station

Next, what we did was to calculate and plot the most probable sequence of states (Viterbi path) for the optimal 4-state model by constructing a Viterbi algorithm. So, in Figure 4.4 we can observe the estimated state for the model for each of the 90 days for each year from the years of 1975 until 2002, starting from the first of February. We observe that some years (such as 1985) have blocks dominated mainly by one state (state 1), indicating consistent wet weather conditions, while other years show frequent alternation between states.

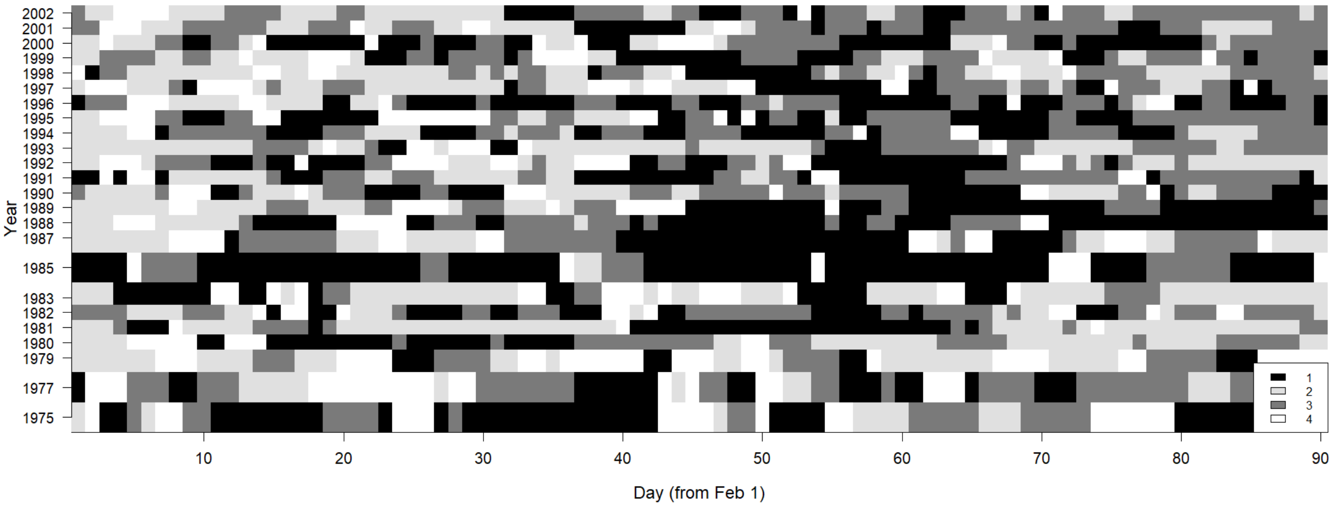


Figure 4.4: Estimated State Sequence

Lastly, we would like to see the state variability for our final HMM with the 4 states. So, in the Figure 4.5 we observe the estimated number of days (February to April) that each state has in each year from 1975 to 2002. What we see is an intense variability mainly for states 1 and 2 and their frequencies seem in opposition for the majority of the years. This fact is one more sign that these two states represent contrasting weather patterns, something which is also confirmed by the high probabilities of rain for state 1 and the extremely low probabilities of a rainfall occurrence in state 2 across all 10 weather stations. As regards state 3, we observe an upward trend during the 1980s, while for the state 4 there is relatively low interannual variability, with its frequency remaining almost stable across the years.

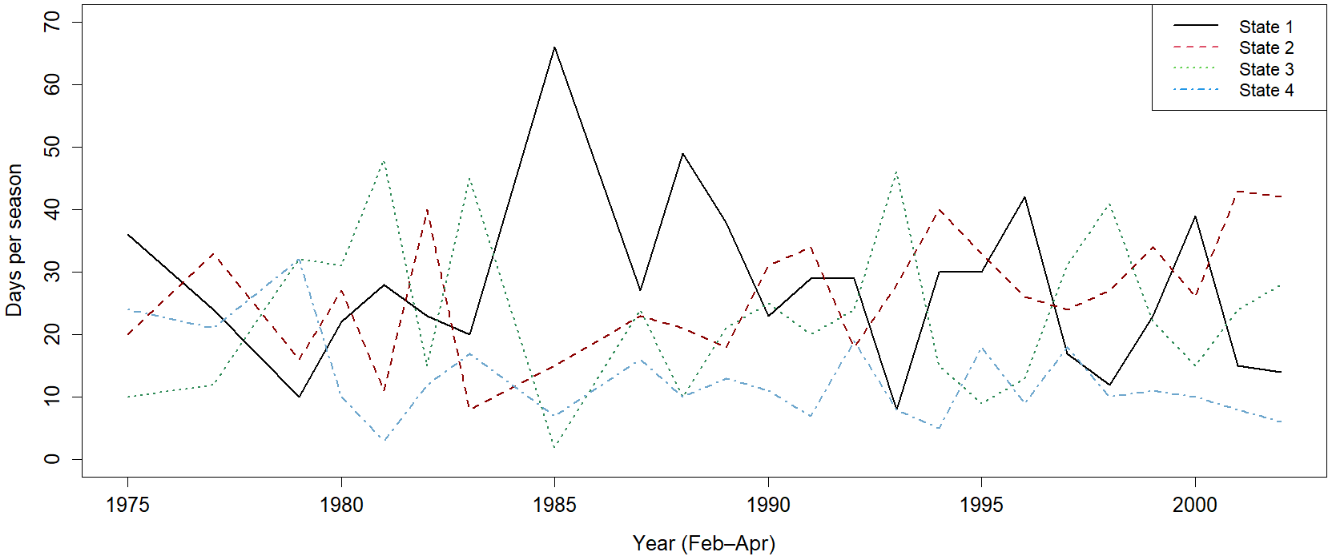


Figure 4.5: Interannual variability of the Homogeneous HMM state-occurrence frequency

Chapter 5

Modeling rainfall occurrence over Northeast Brazil by introducing time-dependency

5.1 Non-Homogeneous Hidden Markov Model (NHMM)

After identifying the 4 states of the rainfall patterns in our data we decided to introduce some time-dependency in the process so to include climate variability information and to improve our model's ability to simulate and predict rainfall occurrences in the area. In Hidden Markov Models, there are two ways to add time-dependency and heterogeneity in a model by incorporating external information. The first approach is the one, which will analyze in this section and it describes how the transition probabilities between states for each time step can vary based on the different values of this external input. The second way is to allow the external variables-covariates to influence the observation process and in this way change the emission probabilities, in our case the Bernoulli probabilities of rainfall occurrence at each station.

Firstly, we would like to extend our research and link the hidden weather states to large-scale atmospheric patterns, using the seasonal mean precipitation simulated by a General Circulation Model (GCM) as an external variable. So, we will describe how we incorporated this variable as a covariate into the tpm and how we fitted a NHMM in our rainfall data. In this way the transitions between states will no longer be purely stochastic but we could interpret them using this external information. In general NHMMs which incorporate any extra atmospheric information are extremely useful in climate change studies and those precipitation models which take advantage of atmospheric variables are known as ”weather state models” or ”downscaling” models.These type of models are also described more in the papers of Hay [22] and the paper of Bárdossy and Plate [23]. Finally, with this model and in general with such ”downscaling” models, we can forecast-generate multiple sequences of rainfall at multiple weather cites.

Now, first of all, we have to denote the 2 assumptions of the NHMM that we constructed. These are:

1. The first assumption is that the current state S_t depends only on the previous state S_{t-1} and the values of the covariates X_t and not on all the history and the previous values of all the states and covariates. This can be expressed mathematically as:

$$P(S_t|S_{1:t-1}, \mathbf{X}_{1:T}) = P(S_t|S_{t-1}, \mathbf{X}_t) \tag{5.1.1}$$

2. The second assumption is that the observations of rainfall or not $\mathbf{R}_t$ at a specific day t are conditionally independent given the state at this day. This is expressed as:

$$P(\mathbf{R}_t|S_t, \mathbf{R}_{1:t-1}, \mathbf{X}_{1:T}) = P(\mathbf{R}_t|S_t) \tag{5.1.2}$$

As we know, the transition probabilities in a NHMM are not constant as they vary over time and in our case specifically, they are influenced by some covariates X_t. In our example X_t is a vector of dimensions 2160x1 and represents the GCM simulated seasonal average precipitation anomaly, which is driven by historical sea surface temperature. As we know, the temperature of the surface of the ocean play a significant role in the climate and in specific weather patterns, especially in a country like Brazil, which is surrounded by the Atlantic ocean. A GCM is a well-known climate mathematical model, which can simulate various atmospheric quantities like rainfall amounts. GCM models simulate this kind of values by incorporating the interactions between the atmosphere, the sea and the land and as anyone can imagine they play an extremely significant role in recognizing and understanding past, present and future climate events and phenomena. The GCM that been used here is the ECHAM 4.5 model (Roeckner 1996 [24]), a well-known and established GCM that has been widely used for climate research.

The seasonal average rainfall anomaly value has been centered and standardized by its standard deviation. Also, we have to declare that for each season-year, that means for each 90 days, we have the same daily value repeated. By putting the same value for each season in our covariate vector, we simplify the model but at the same time taking into account the seasonal-scale climate information and we are in the same line with the scheme of IRI. To better understand the distibution of our covariates and the values that it takes we can observe the table below:

Statistic	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
X_t	-2.593	-0.672	0.153	-0.069	0.732	1.591

Table 5.1: Summary statistics of the covariate X_t, representing the GCM simulated seasonal precipitation anomaly.

Also, now regarding the transition probability matrix, for each day of the 2160 in our dataset we will compute a transition probability matrix Γ which will be kxk where k is the number of hidden states of our NHMM. More specifically, the transition between states is being modeled by a multinomial logistic regression, where the transition probability from state j to state i at day t is given by:

$$P(S_t = i \mid S_{t-1} = j, \mathbf{X_t} = \mathbf{x}) = \frac{\exp(\sigma_{ji} + \boldsymbol{\rho}_i' \mathbf{x})}{\sum_{k=1}^{K} \exp(\sigma_{jk} + \boldsymbol{\rho}_k' \mathbf{x})}, \tag{5.1.3}$$

where x is the value of the covariate at day t, σ_{ji} are the baseline transition coefficients (σ is a matrix of kxk), and $\boldsymbol{\rho}_i$ are the coefficinets of the covariates and basically is a vector of kxD where D is the number of covariates X_t, so in our example ρ is a vector of kx1. The ' in ρ indicates the transpose of the vector. Also, for the case of the first state, that is the initial state, the probability of the process to start in each state i given the covariate $\mathbf{X_1}$ is:

$$P(S_1 = i \mid \mathbf{X_1} = \mathbf{x}) = \frac{\exp(\lambda_i + \boldsymbol{\rho}_i' \mathbf{x})}{\sum_{k=1}^{K} \exp(\lambda_k + \boldsymbol{\rho}_k' \mathbf{x})}, \tag{5.1.4}$$

To ensure identifiability, we set up the λ_1, ρ_1 and σ_{j1} equal to 0. Also, same modeling of transition probabilities can be found in the papers of [10] and the one of [19]. In addition, we would like to assess the uncertainty around the estimate of the parameter ρ, which quantifies the effect of the GCM' simulated seasonal average rainfall anomaly on the transition probabilities. For this reason, we calculated the standard errors of the estimates of ρ and its the asymptotic normal Confidence Intervals. For the standard error calculation we used the Hessian matrix, though the most correct way would be calculating them through the profile likelihood method. We present those metrics in the table below:

	Estimate	**Se**	**95% CI**
ρ_2	-0.152	0.088	(-0.324, 0.020)
ρ_3	-0.051	0.074	(-0.197, 0.094)
ρ_4	-0.377	0.078	(-0.529, -0.224)

Table 5.2: Standard errors and 95% confidence intervals for the estimated ρ coefficients.

Now, as regards the observation process and the Bernoulli probabilities, the logic is exactly the same as described in the Section 4.3. The only thing that changes in the NHMM is the way the hidden states transit over time as in this case the transition probabilities are modeled as a function of the covariates $\mathbf{X_t}$.

5.2 NHMM fit and parameter estimation

First of all, we have to say that we chose to fit a 4-state NHMM based on the fact that in the previous sections we have fitted Homogeneous Hidden Markov Models with various number of states and based on metrics like BIC and log-likelihood, we saw that the model with the 4 states was the one with the best fit and cocluded that the hidden states that govern the rainfall patterns in the area of Northeast Brazil are 4. So, driven from this fact, we decided to extend this model to a more complicated NHMM framework. The NHMM will not influence the interpretation of the 4 states but simply modify the way that the transitions occur between these states. In practice, to fit the NHMM, we used the R programming language and we implemented several custom functions-methods like the ones described in the previous sections, each one with a specific purpose in the process.

So, the very first thing that we did was to construct a function, where we compute the tpm Γ_t for each day and the initial state probabilities as described in equations (5.1.3) and (5.1.4) correspondingly. We also used the same logic to transform our parameters to working parameters, but only for the Bernoulli probabilities, because ϱ and σ are already free parameters and do not require transformation. Next, we computed the log-likelihood using the forward algorithm but this time the initial state probabilities are computed as we described and the forward probabilities are calculated as:

$$\alpha_t(s) = \sum_{s'} \alpha_{t-1}(s')\Gamma_t(s' \to s)P(X_t \mid S_t = s) \tag{5.2.1}$$

since Γ_t varies in time. More specifically, because Γ_t depends on X_t, and X_t takes the same value for every 90-day period, Γ_t remains fixed within each 90-day window, but is updated at the beginning of each new period. To estimate the optimal parameters Θ of this log-likelihood we used again the nlm method and took random initial values for a uniform distribution in the range of (0.1,0.9) for all our parameters. In Table 5.3 we can see more information about the minimum log-likelihood value, AIC and BIC metrics and the comparison with the same metrics of the homogeneous 4-state HMM:

Model	**Number of Parameters**	**Log-Likelihood**	**AIC**	**BIC**
Homogeneous	52	-12165.15	24436.91	24752.58
Non-Homogeneous	58	-12151.51	24431.02	24794.40

Table 5.3: Comparison of Homogeneous vs Non-Homogeneous 4-state HMM

So, first of all, we observe the increase in the parameters from 52 to 58, which is basically coming from the ϱ parameter and also the initial state distribution in the NHMM. Due to the increased complexity in the modeling of the transition probabilities and the initial distribution, the log-likelihood of the NHMM improves from -12165.15 to -12151.51, which indicates a better fit, so our motivation to fit the NHMM is justified. Also we see an improvement on AIC but the same is not true for BIC, which is a bit higher than the one of the HMM model. BIC is higher in the NHMM because this metric penalizes model complexity more heavily than AIC.

The improved fit of the NHMM assure us that the transitions between the hidden states are influenced by large-scale climate variables. Thus, this new model seems more realistic and can help us better understand the rainfall patterns in Northeast Brazil.

Finally, fitting the NHMM with the seasonal mean rainfall anomaly did not led to any significant difference in the results of the Bernoulli probabilities and the state composites. However, we see changes in the transition probabilities for the period of the 24 seasons and for instance we will illustrate an example in

Figure 5.1 of how the transition probabilities change through time and based on the seasonal average rainfall anomaly X_t value at each season.

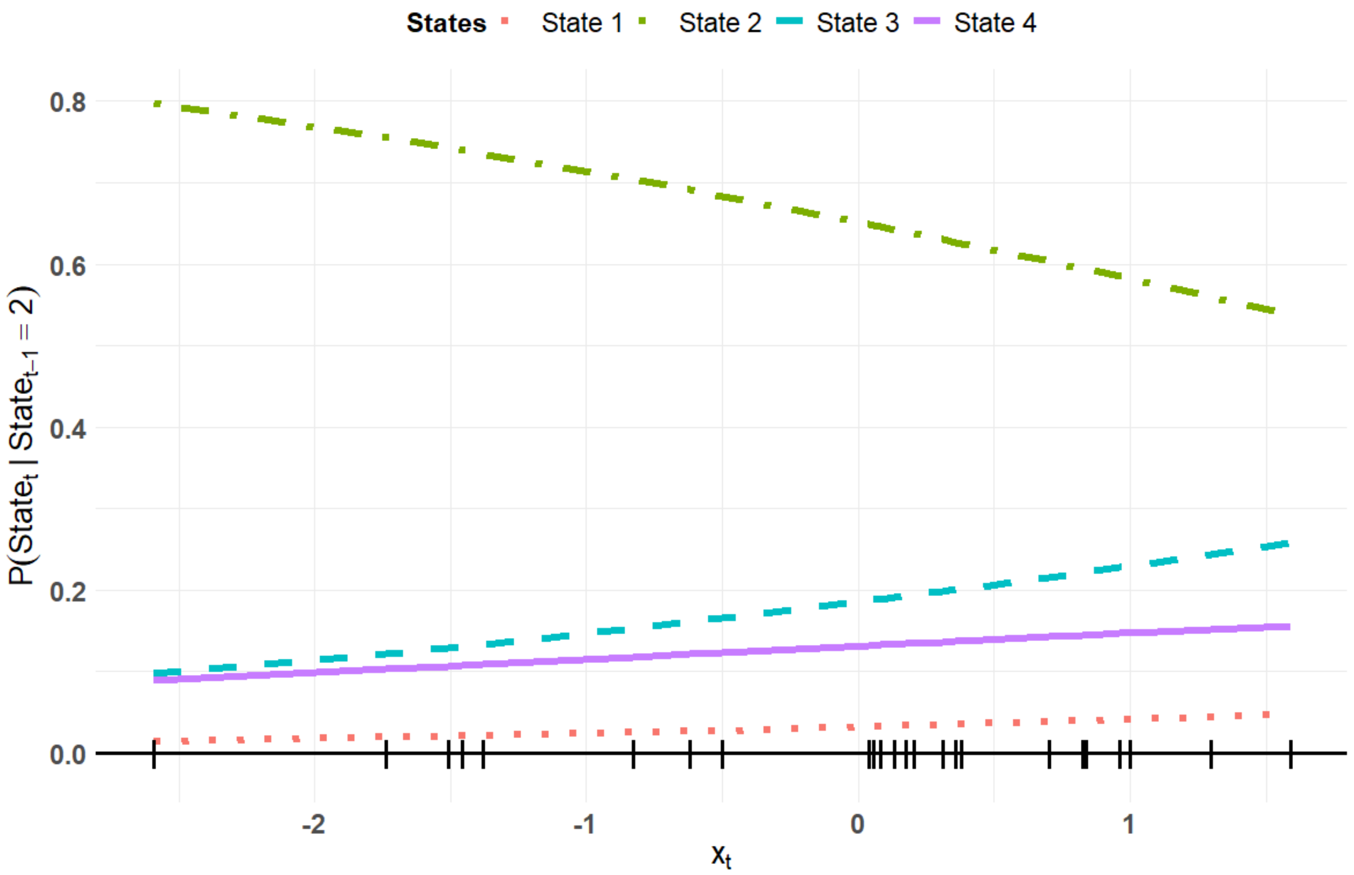


Figure 5.1: Transition probability from state 2 (the ”dry” one)

Thus, from the plot, we observe that the probability of remaining in state 2 decreases as X_t increases, while the probability of transitioning to state 4 and 1 are very low but seem to increase. Furthermore, the probability of transition to state 3 from state 2 increases also as the covariate increases its value. Yet, the most probable as we see is to remain at state 2. We have to refer for better understanding that the short vertical black bars on the x-axis indicate the different values of X_t for each season. Those values were used for estimating the transition probabilities in the NHMM. We can see also the corresponding plots of the remaining states 1,3 and 4 in the Appendix A to understand better the transitions between the states for the Non-Homogeneous model. Overall, we see an influence on the transitions from state to state because of the covariates.

Finally, we calculated the initial state distribution δ based on the equation (5.1.4) and we found out that the process starts from the "dry" state, the state 2 as $\delta = (0, 1, 0, 0)$.

5.3 Incorporating GCM' simulated seasonal mean rainfall anomaly into the observation process

5.3.1 Model fitting process and the benchmark model Mo

After fitting the NHMM and incorporating the covariate in the tpm we realized that this did not affect in a significant way the probability of rain at each station. Thus, we decided to try a different approach and incorporate them this time in the observation process, that is in the Bernoulli probabilities of rainfall occurrence. More specifically, in the new case, the Bernoulli probabilities will depend on the covariate X_t , meaning that the probability of observing the rainfall occurrence vector across the 10 weather stations R_t at day t, given the state $S_t = i$, will now be a function of X_t. We implemented this approach, so to check if large scale climate variables such as the one that we examine here, have an influence in rain occurrences in the area of Northeast Brazil. If we prove something like this, we will enhance the ability of the model to simulate rain patterns. We can see a similar approach at the paper of N.I. Ramesh and C. Onof [25].

Now, as first try and before fitting the model with different rain probabilities across stations given the state, we decided to fit the same "Mo" model which is described in Section 4.4 but this time we modeled the Bernoulli probabilities with a logistic regression framework. Starting with this simpler model, though we have proved that this is not the real case, allowed us to better understand the fitting process while also helped us find reasonable initial values which assisted in the convergence of the station-specific model.

So, now, the likelihood computation will be calculated as:

$$L_T = \delta P(R_1|S_1, x_1)\Gamma_2 P(R_2|S_2, x_2) \cdots \Gamma_T P(R_T|S_T, x_T)1', \tag{5.3.1}$$

and the only difference is the impact of the covariates x_t in the state-dependent probabilities $P(R_t|S_t, x_t)$

and in the corresponding matrices:

$$P(R_t \mid S_t, x_t) = \text{diag}(P(R_t \mid S_t = 1, x_t), \ldots, P(R_t \mid S_t = m, x_t)) \tag{5.3.2}$$

Now, the simplest case but not the realistic is the one of the model with shared Bernoulli probabilities across state where they are modeled as:

$$\text{logit}\, P(R_t \mid S_t = i, x_t) = \beta_{0,i} + \beta_{1,i} x_t \tag{5.3.3}$$

, where $\beta_{0,i}$ is the intercept for each hidden state,$\beta_{1,i}$ is effect of the covariate (slope) for each hidden state and x_t is the value of the seasonal mean rainfall anomaly at this day. The slope indicates how sensitive is the rainfall occurrence to changes in the covariate. Now, as we can understand from the new form of the Bernoulli probabilities, the parameters in the new model that refer to the emission probabilities part are increased to 8 for a 4-state model because now we have the extra terms of the slopes. Of course, if $\beta_{1,i} = 0$ for every state, that means there is no effect of the covariate on the emission probabilities, the model reduces to the Homogeneous stationary ”Mo” model that we fitted in section 4.4 and in which the rainfall occurrence probabilities were constant for each day. In our new model, these probabilities will vary depending of the different values of x_t and because as we have said the values of x_t remain constant for each season (90 days), we will have as a result 24 different Bernoulli probabilities for each state, because of the 24 seasons. These probabilities will be calculated using the optimal estimated parameters $\beta_{0,i}$ and $\beta_{1,i}$ in the logistic regression framework that we described.

So, first of all, we fitted the same stationary Homogeneous model ”Mo” with 4 hidden states, with the only difference that we just described. For the fitting process we used the same methodology as the simple ”Mo” model. After fitting the new model we estimated the optimal parameter vector Θ by maximizing the log-likelihood that we see in the Equation (5.3.1). Below, we present the values the optimal parameters $\beta_{0,i}$ and $\beta_{1,i}$ for each state:

	State 1	**State 2**	**State 3**	**State 4**
$\beta_{0,i}$	-0.686	-2.676	1.706	0.453
$\beta_{1,i}$	0.393	0.459	0.026	0.134

Table 5.4: Estimated optimal parameters $\beta_{0,i}$ and $\beta_{1,i}$ for each hidden state for the Mo model

For instance, based on the Table 5.4 the Bernoulli probability for each hidden state i at time t is given by the logistic regression formula:

$$P(R_1 \mid S_1 = i, x_1) = \frac{1}{1 + \exp(-(\beta_{0,i} + \beta_{1,i}x_1))} \tag{5.3.4}$$

So, as we said this model is not a realistic one, but it is a benchmark for passing to the most complex model with the station-specific Bernoulli probabilities for each state.

5.3.2 Incorporating Station-Specific Intercepts with a Shared Covariate Effect - Case 1

Fitting the baseline Mo model helped us understand the influence of our covariate to the rainfall occurrences in a simplified framework. Now is the time that we will introduce the station-specific model and we would like to try more than one ways to incorporate the GCM' simulated seasonal average rainfall anomaly in the observation process through a logistic regression. Before, in the simplest Mo model $\beta_{0,i}$ and $\beta_{1,i}$ were the same across stations. Now, we can have two different cases and in this section we will describe the first one. So, firstly we will suppose that we have different intercepts (β_0) across stations, but common slopes (β_1) In this formulation, the intercepts of the logistic regression vary for each station, capturing station-specific rainfall occurrence patterns, but the effect of the covariate (the slope β_1) is the same across all stations. This modeling framework will assist us in assessing wether large-scale climate variable has a similar effect in all stations. So, now the Bernoulli probabilities are modeled as:

$$\text{logit } P(R_t^j = 1 | S_t = s, x_t) = \beta_{0,s}^j + \beta_{1,s}x_t, \tag{5.3.5}$$

where $\beta_{0,s}^j$ is the intercept for state s at station j, $\beta_{1,s}$ is the slope for state s, which is shared across all stations and x_t is the covariate.

As we understand, we have to estimate 44 parameters as regards the Bernoulli probabilities, 40 that refer to the different intercepts per state and station and 4 which refer to the different slope of each hidden state. In the Table 5.5 we see the estimated intercepts $\beta_{0,s,j}$. In this case, those are responsible for the differences in rainfall occurrence probabilities across the 10 weather stations.

State	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7	Station 8	Station 9	Station 10
1	0.581	1.374	1.210	0.466	0.735	0.597	0.780	0.505	2.163	1.933
2	-3.295	-2.431	-2.113	-3.982	-1.415	-2.544	-3.160	-3.301	-2.287	-2.277
3	-1.299	0.011	-1.356	-1.091	0.401	-2.146	-0.419	-0.903	0.615	0.703
4	-0.696	-0.835	1.183	-1.263	-0.421	0.484	-1.593	-1.031	0.076	-0.034

Table 5.5: Estimated optimal intercept parameters $\beta_{0,s,j}$ for each hidden state and station

Furthermore, in Table 5.6 we observe the optimal estimated slopes for the model with the 4 hidden states. We see that while state 1 and state 4 have different slopes, state 2 and state 3 have the same one, indicating that maybe the effect of the covariate in these state is pretty similar. Thus, one could say that because of this, rainfall occurence in these specific states might be influenced by similar factors. Finally, all the slopes are positive, which means that the seasonal mean rainfall anomaly has a positive effect on the rain probability. Higher values of X_t lead a higher likelihood of rainfall occurrence.

	State 1	State 2	State 3	State 4
$\beta_{1,i}$	0.071	0.384	0.385	0.231

Table 5.6: Estimated slopes $\beta_{1,i}$ and their se for each hidden state for the Mt model-case 1

However, since the slopes quantify the influence of the GCM' seasonal average rainfall anomaly to the rainfall occurrence probabilities, we would also like to quantify the uncertainty of those estimates, so in the table 5.7 we added their standard errors (Se) and the asymptotic normal Confidence Intervals (CI) so to assess their reliability. We calculated the standard errors of these estimates using the Hessian matrix, though the most correct way to find them would be through the profile likelihood method.

	Estimate	Se	95% CI
$\beta_{1,1}$	0.071	0.038	(-0.003, 0.145)
$\beta_{1,2}$	0.384	0.087	(0.214, 0.554)
$\beta_{1,3}$	0.385	0.047	(0.293, 0.477)
$\beta_{1,4}$	0.231	0.093	(0.049, 0.414)

Table 5.7: Se and 95% CI for β_1 estimates

Before proceeding to the next section, one thought is that instead of fitting a model where all the intercepts and the slopes will be different per state and station, one possible adjustment could be to assign the same slope to the states 2 and 3 as we observed from the table 5.6.

5.3.3 Independent Effects of Covariates Across Stations - Case 2

Now we are ready to introduce a much more flexible model where each station will have its own distinct relationship with the covariate. In this case, there are different slopes and different intercepts across all 10 stations and the 4 states. In this case, we have the maximum variability and what we would like to check for spatial hetoregenity and see how the different weather cites respond to this simulated large climate variable. So, now the Bernoulli probabilities are modeled as:

$$\text{logit } P(R_t^j = 1|S_t = s, x_t) = \beta_{0,s}^j + \beta_{1,s}^j x_t, \tag{5.3.6}$$

where $\beta_{0,s}^j$ is the intercept of each state s and station j and $\beta_{1,s}^j$ is the slope for each station j and sate s. This formulation allows each station to be influenced by the covariate in a unique way, but of course increases a lot the parameters that need to be estimated and the computational burden of our fitting process. So, now the parameters that refer to the rainfall occurrence probabilities increase to 80 as we have 40 different intercepts $\beta_{0,s,j}$ and 40 different slopes $\beta_{1,s,j}$ per state s and station j. The specific values of these 80 parameters but also the standard errors of the sloves and their CI can be found in the Appendix B. So, as we understand the rainfall occurrence probability for each station at each state will differ from season to season as also in the previous section, because of the incorporation of the covariate in the observation process through a logistic regression model, this time with different slopes and intercepts. To illustrate this more clearly, we found out the three largest changes in rainfall occurrence probabilities across seasons for different stations and states. The stations 2,7 and 8 show the biggest variability across the 24 years and more specifically we observe for the state 2 station a change of approximately 0.28 from the first year until the 24rth one, for the state 4,station 7 a change of 0.27 and for state 4, station 2 a change of around 0.19.

Yet, with the increase in the parameters comes also the risk of convergence issues. Furthermore, with this formulation because we have a much more compelx optimization, we need much longer run times and the convergence process becomes extremely slow. To fit this new model, we used as initial values for $\beta_{0,s,j}$ and

$\beta_{1,s,j}$ the values that we see in the tables 5.5 and 5.6.

In addition, this approach might be not so realistic and appropriate for our data, because as we observed in the previous model, some states share the same slope. Finally, as we will see later, we observe an improvement in the log-likelihood of this model compared to the previous models, but AIC is about the same and BIC worsen compared to the model described in the previous section.

5.3.4 Model comparisons and results

Finally, we implemented this analysis to check if which model would give us a better statistical fit overall. To evaluate this, we would compare metrics like the log-likelihood and BIC and check differences in these for the Homogeneous model, the Non-Homogeneous one and the two models which incorporate the Bernoulli probabilities in the observation process. Below, the Table 5.8 summarizes the results for all these four models.

Model	Log-Likelihood	AIC	BIC
Homogeneous Mt	-12165.15	24436.91	24752.58
Non-Homogeneous Mt	-12151.51	24431.02	24794.4
Mt with covariates in the observation process - case 1	-12135.3	24382.59	24700.55
Mt with covariates in the observation process- case 2	- 12099.14	24382.28	24904.64

Table 5.8: Comparison of log-likelihood and BIC for the Mt model without covariates and with covariates in the observation process.

So, in the Table 5.8, we observe that both the log-likelihood and AIC decrease as we go from the simplest to the most complex model. Also, as we expected BIC increases as we add complexity with an exception in the model where we have the same slopes but different intercepts in the logistic regression that incorporates the covariates in the observation process. Finally, in order to avoid adding too many parameters and overfit our model we would go with the Mt model of the first case. That is the most preferable model as it seems to has the best statistical fit depending on the combination of these 3 metrics. By this model, we understand that the impact of the GCM'simulated seasonal average rainfall anomaly is similar across stations, that is that large climate variables change the rain probabilities in a similar uniform manner. Finally, the insights that gained from the model can be used by experts in order to make seasonal forecasting and predict or simulate rainfall patterns and in this way assist local communities of Northeast Brazil.

Additionally, to have a clearer picture of the comparison of these four models we made a graphical representation of AIC and BIC, where in the x axis we see the number of each model as they are in the table 5.8 and in y axis the values of the two metrics. This visualization reinforces our belief of which model is the optimal choice.

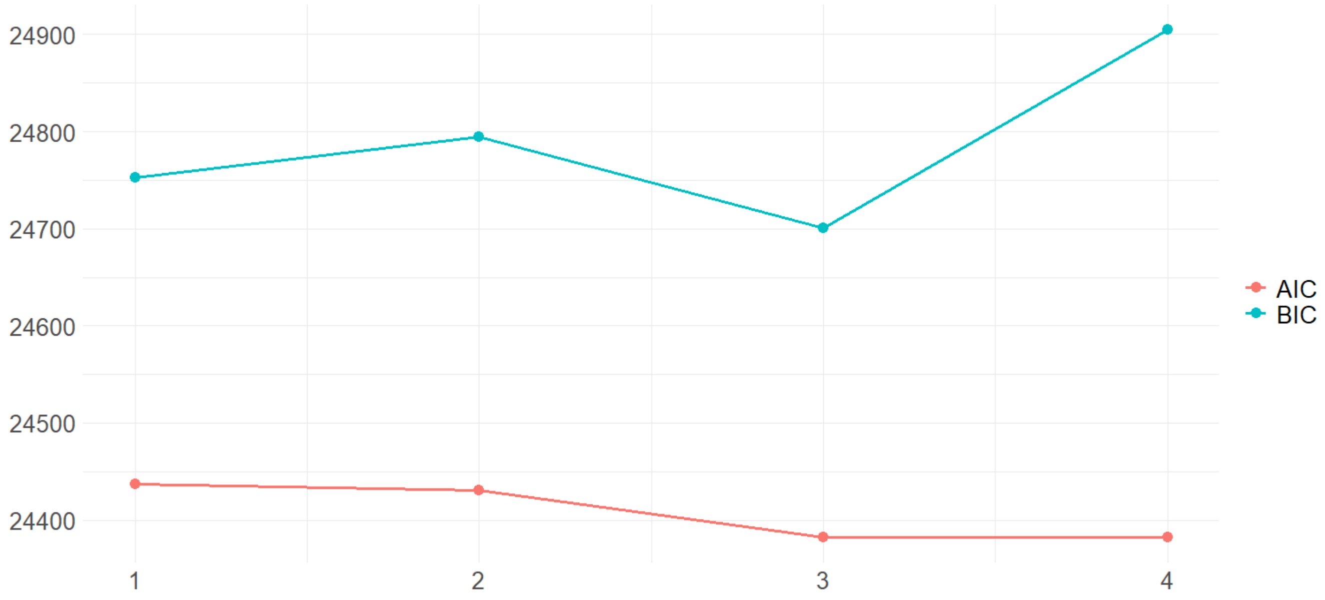


Figure 5.2: AIC and BIC comparison between the 4 models

Chapter 6

HMM for independent seasonal sequences of rainfall occurrence over Northeast Brazil

6.1 Modeling HMM with independent seasonal sequences

Until now, we have modeled our data as a continuous sequence of 2160 days. However, based on the fact that our data consist of 24 seasons-years, each one with 90 days (February - April), we thought that we could modify our approach by partitioning the dataset into 24 independent sequences. By modeling each year as an independent sequence, we allow for yearly variations in rainfall occurrence patterns and we basically treat each year separately. In this manner, we account for the seasonal and interannual variability. This may be a more realistic approach than the previous one as rain patterns can fluctuate a lot. Also, the approach of the continuous sequence of 2160 days assumes that a rain at a specific day at a specific year is influenced by past rainfall occurrences, even if those rains happened in years far before the given day. However, this assumption might not be fully justified, as each year can start with different initial conditions.

Now, based on our new approach we will compute the likelihood separately for each season and the total likelihood of our model will be the product of these 24 likelihoods. So, instead of calculating and evaluating the likelihood for the entire sequence of 2160 at once, we iterate over each season-year and compute the likelihood independently. In this way, we do not force any strict continuity between the years. The total likelihood is given by:

$$L_{\text{total}} = L(\Theta, R_1) \cdot L(\Theta, R_2) \cdots L(\Theta, R_{24}) = \prod_{s=1}^{24} L(\Theta, R_s), \tag{6.1.1}$$

where R_s indicates the rainfall occurrence data for each season s. For instance the R_1 is a 90x10 matrix with the first 90 days of our dataset and the rainfall occurrence across the 10 weather stations. Of course, we take the logarithms in order to avoid common problems like numerical underflow and so now we have that:

$$\log L_{\text{total}} = \sum_{s=1}^{24} \log L(\Theta, R_s), \tag{6.1.2}$$

Θ is the parameter vector that we have to estimate by optimizing the log-likelihood and is common across all the likelihoods.

The basic difference in this modeling approach is that now because each season is treated as an independent sequence, it will have its own initial state distribution δ_s.

6.2 Same initial state distributions across seasons

Our first model based on the new approach will be one in which for each season the δ_s will be equal to the stationary distribution and this is expressed mathematically as:

$$\delta_1 = \delta_2 = \cdots = \delta_{24}$$

By this model, we do not increase the number of parameters need to be estimated as Θ remains the same. Also, both the rainfall occurrence probabilities and the transition probability matrix remain the same.

Of course, we need to assess if this approach improves the statistical fit of our data and this is why we will compare key model selection metrics like AIC and BIC between this model and the one which treated the data as a continuous sequence of 2160 days. We can see clearer this comparison of those 4-state HMM models in Table 6.1.

6.3 Different initial state distributions for each season

In the previous section, we assumed that the initial state distribution for each season is the same and equal to the stationary distribution. Now we would like to explore an alternative approach where instead of having one δ same across seasons, we will have 24 different initial state distributions $\delta_1, \delta_2, \ldots, \delta_{24}$ By having different δ_s we allow seasonal variations in the hidden states. That means that the probability of the process being in each state at the first day t=1 varies across the 24 years. Surely, as we understand just by having all these δ_s, the number of parameters need to estimate increases a lot. With the previous approach where we had just one δ, we only needed to estimate $m-1$ parameters as regards the initial state distribution part but now we need to estimate $24 * (m-1)$ parameters. We also already know that the sum of the elements of δ in each season should sum up to 1. This is expressed mathematically as:

$$\sum_{i=1}^{m} \delta_{s,i} = 1, \quad \forall s \in \{1, 2, \ldots, 24\}$$

and that all of them should be positive, that is $\delta_{s,i} > 0$ for all hidden states i and seasons s. We have fitted this model and calculated all these different deltas and we found out that:

- **Seasons starting in State 1 (”Wet” State):** The seasons of 1977, 1984, 1990, 1995 and 2001 begin in the wet state, suggesting that these years start with a wet regime.
- **Seasons starting in State 2 (”Dry” State):** We observed that in the majority of seasons and that is the years 1975, 1979, 1980, 1981, 1983, 1985, 1987, 1988, 1991, 1993, 1994, 1998 and 1999, the probability of starting at state 2 is extremely high if not exactly 1, so they begin in a dry period.
- **Seasons starting in State 3:** Only in the sixth season that is the year of 1981, we see that it is certain (probability equal to 1) that we start in this state.
- **Seasons starting in State 4:** A small number of seasons and more specifically the seasons 1992, 1997 and 2002 begin in state 4.
- **Seasons with mixed probabilities:** Only at the year of 1989 we observe a more balanced initial distribution between state 1 (24.5%) and state 2 (75.5%), indicating a potential transitional phase at the beginning of that season.

By knowing the exact initial state distribution at each of the 24 years we learn about the weather conditions at the start of each season but also we gain useful insights of the variability of the climate in Northeast Brazil throughout these years.

6.4 Model comparison and results

After fitting these two models, we would like to find out which approach is better for our data and of course if treating our data as 24 independent sequences makes any sense. So below we present the AIC and BIC of these modeling approaches so we can evaluate better their performance.

Model	AIC	BIC
Single-sequence HMM (2160 days)	24436.91	24752.58
Seasonal HMM (24x90 days) - Common δ	24424.21	24719.46
Seasonal HMM (24x90 days) - Variable δ	24508.86	25212.91

Table 6.1: Comparison of AIC and BIC for the two different modeling approaches

As we observe from the Table 6.1, AIC and BIC are significantly lower for the Seasonal HMM with a common initial state distribution δ which is also the stationary distribution which indicates a better statistical fit to our data. Furthermore, while we have tried the approach with different δ_s, we observe that maybe because of the big increase in the model parameters and the increased complexity of the model, the fit is not improved.

The above finding is reasonable and something that we expected as rainfall occurrence patterns in Northeast Brazil are influenced by specific weather phenomena like the Atlantic Intertropical Convergence Zone (ITCZ) and the South American Monsoon System (SAMS) which may vary from year to year. The ITCZ is a group of clouds and rain which during the rain period moves north and unites with the SAMS near the Amazon. This combination-union leads to rainfall in Northeast Brazil. This seasonal pattern though, changes from year to year and for example if the ITCZ moves south instead of north this may lead to dryness in the area. However, the fact that the model with common δ outperforms the one with variable δ suggests that although there are some vairation in rainfall occurrences between years, the underlying hidden states and their initialization remain relatively stable over time. This could indicate that it is not necessary that seasonal factors lead to a unique initialization of the states.

So the model with the 24 seasons approach with the stationary distribution as the initial state distribution for each season, may be a better one because the position and the strength of the ITCZ fluctuate annually due to climate phenomena like the El Niño and the changes in SST in the Atlantic ocean.

Chapter 7

Conclusions and Discussion

The purpose of this thesis was to explore the application of HMMs in modeling the occurrence of daily rainfall at ten weather stations in Ceará area in northeast Brazil during the February - April season from 1975 to 2002. Firstly, the study aimed to identify the hidden states that govern the rainfall dynamics in the area and analyze their meteorological significance. Next, through this thesis, we tried to understand and evaluate the impact of GCM's simulated seasonal rainfall anomaly on the transition between the states but also on the rainfall occurrence probabilities at each station.

Summary of the methodology approach and results

Our research initially started by applying the HMM methodology to the Old Faithful Geyser data so as to better understand how this methodology works and also how the identified hidden states can capture patterns in binary time series data. In this case, we found out that there are 2 states which indicate short and long eruptions and also that the non-stationarity assumption did not led to any significant better fit, which means that the process influencing the eruptions remains stable in time. This analysis assisted us in proceeding to the most complex case of the rainfall occurrence dataset of Northeast Brazil.

The first application on these data was to fit several homogeneous HMMs so to identify the optimal number of states. The best model was the one with 4 hidden states, where each of them represent different rainfall patterns across the stations. What we found out is that there is a dry state where the probabilities of rain across all stations is pretty low, a wet state where the same probabilities are high enough and two states that show contrasting behavior in rainfall occurrences across stations. More specifically, at one state there are high probabilities at northern stations and low at the southern while in the other state there is the same pattern but inverted. As we understand from what we describe above, there is a unique probability of rain for each state and station, as we also tried an approach of shared probability across stations for each state but we saw that this assumption is not the reality as someone could expected.

Next, we thought that a NHMM may be a more realistic model for our data because it incorporates time-dependency in the process. So we included the GCM ' s simulated seasonal mean rainfall anomaly as an input in our model and in this way we accounted for a dynamic transition probability matrix. By comparing the homogeneous with the non-homogeneous approach we observed that our input variable derived by the GCM had a significant impact in transition dynamics which means that large scale climate variables influence the transition between wet and dry states in the area of Northeast Brazil.

In addition, we also explored the impact that our input variable may have on the rain probabilities across stations, by incorporating it in the observation process through a logistic regression framework. We achieved this by using two alternative formulations, one where the input variable assumed to have the same influence on rain probabilities and one where there is an independent slope-covariate effect for each station. As we understood by evaluating and comparing both models, the first one fits better to our data and suggested that GCM's simulated seasonal rainfall anomaly modulate the likelihood of rainfall within each state in a uniform way.

Finally, we tried an alternative approach by modeling the rain data as independent seasonal sequences so to better capture yearly variations in rainfall occurrences. First we fitted a model with same initial distribution across seasons, which is also the stationary distribution and then a more complicated model where each season has its unique initial distribution. The model with the common initial distribution outperformed the one with the different initial distribution per season but also the homogeneous model so what we found out from our analysis that treating the data as independent sequences may make more sense because of the weather phenomena like ITCZ and SAMS which change drastically from year to year.

Limitation and further research

Throughout this thesis, we faced several specific challenges and limitations. Firstly, as we have already explained, our models rely on the conditional independence assumption between the stations given the state, but this does not imply that the rainfall processes across stations are spatially independent in general. In fact, spatial dependence is modeled through the hidden state. However, in this approach we assume that the spatial correlations are fully explained only by the shares state, which may be not fully true as there might be some dependencies also within each state, especially with stations that are near each other or share common features. Furthermore, another limitation of our thesis could be the incorporation of only one external variable as an input in the NHMM framework but also in the observation process. While the GCM' s seasonal mean rainfall anomaly is a large scale climate variable which provided us valuable insights, we could use more useful factors like the wind average anomaly, humidity metrics and many more. By limiting the covariate to one variable, we may not fully capture the influence of climate variables especially during years with complex climate conditions.

Something that was not included in this study and seems as a valuable direction for future research, is the simulation of rainfall occurrences through the NHMM framework that we applied. Those simulations can be used for plenty of forecasting applications and also be a key factor on climate risk assessments, water resource planning, and agricultural applications.In addition, while we identified and interpreted the 4 hidden states that are present in our data, there is plenty of room for further refining. It seems that there is a need for deeper meteorological research by combining also satelited-derived climate variables or other metrics like the temperature, solar radiation, atmospheric pressure or the wind average, so to enhance the understanding of these states. Also, one more meaningful direction for further research could be the relaxation of the conditional independence assumption by incorporating dependencies between the observations. For instance, we could fit higher-order HMMs to discover these dependencies and better capture the temporal structure of our data. Finally, an quite promising alternative approach that can be followed, for future work is the use of Copula based methods to better capture the dependencies in the rainfall data. Copula is basically a function that links a suitable multivariate distribution to its marginal distributions and capture the dependencies between our variables. By copulas, we can model the components of the rainfall occurrence vector of the stations using different marginals and in this way allow for a more realistic approach and representation of the precipitation patterns. For further details of a copula approach, check the paper of Cappé et al. (2009) [26], the report by Martino [27], which applies a Copula-HMM framework to disease progression modeling.

Appendix A

Transition probabilities of the Non-Homogeneous model

Here we observe the transition probabilities from the states 1,3 and 4 of the Non-Homogeneous HMM with 4 hidden states. The probabilities are represented using different line types and colors, one for each state.

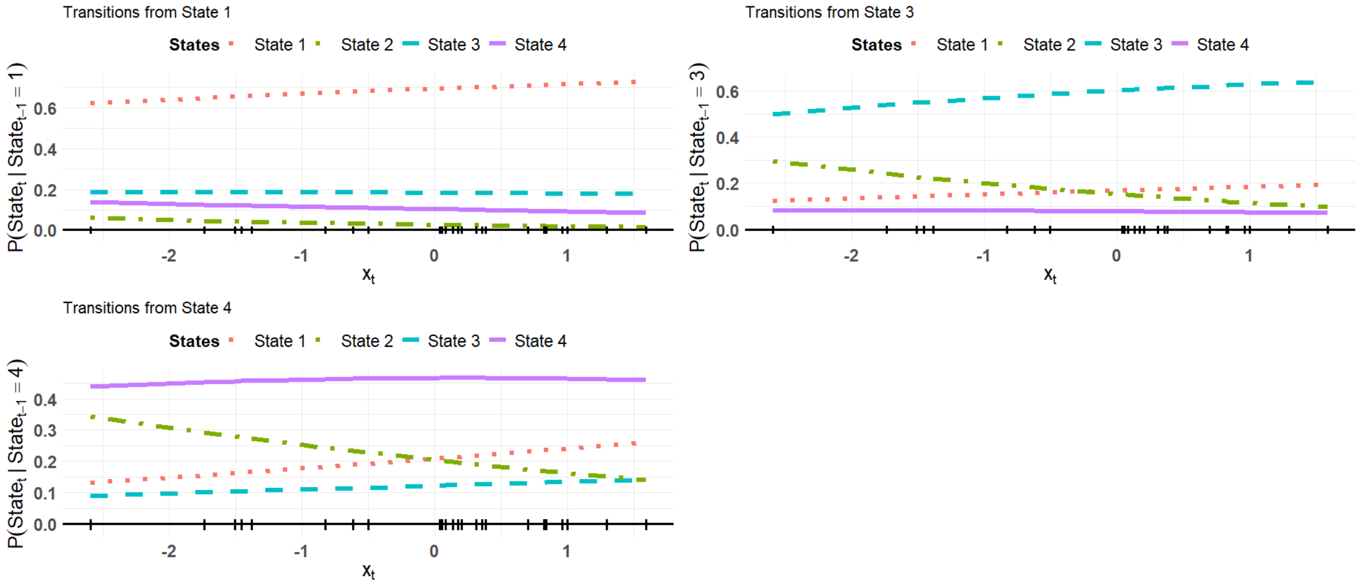


Figure A.1: Transition probabilities from states 1,3 and 4)

Appendix B

Bernoulli Probabilities and Logistic Regression Parameters

In this appendix, we present the Bernoulli probabilities of the model with the covariate in the observation process of the case 1 and the logistic regression parameters used in the model of case 2 which is the one with different slopes and different intercepts per state and station.

B.1 Estimated rainfall occurrence probabilities

The corresponding probabilities of rainfall occurrence for each state and station across the 24 seasons can be seen in the plots below. In the figure B.1 we observe these probabilities for the states 1 and 2 while in the figure B.2 there are the graphs for the states 3 and four. So, our very first observation is that the seasonal average rainfall anomaly influences the rainfall occurrence probabilities in the same way across all stations as there is the same slope for each one of them and also across states as the slopes have very small differences. That is why we see same ups and downs in the plots.

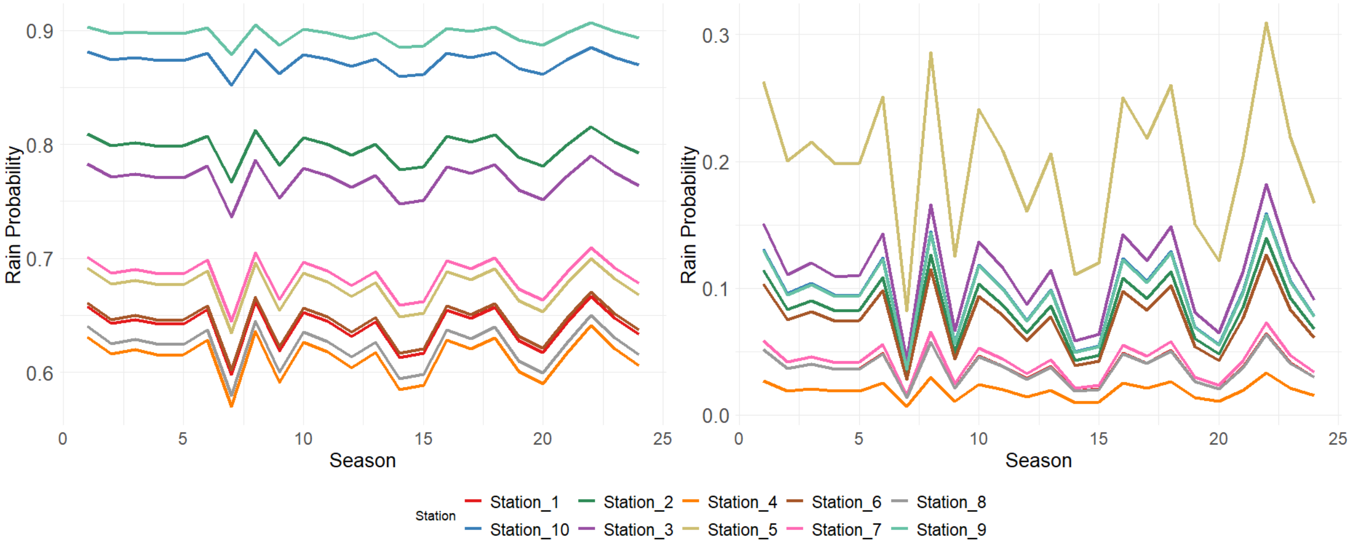


Figure B.1: Rainfall occurrence probabilities per station for states 1 (left) and 2 (right) across the 24 seasons

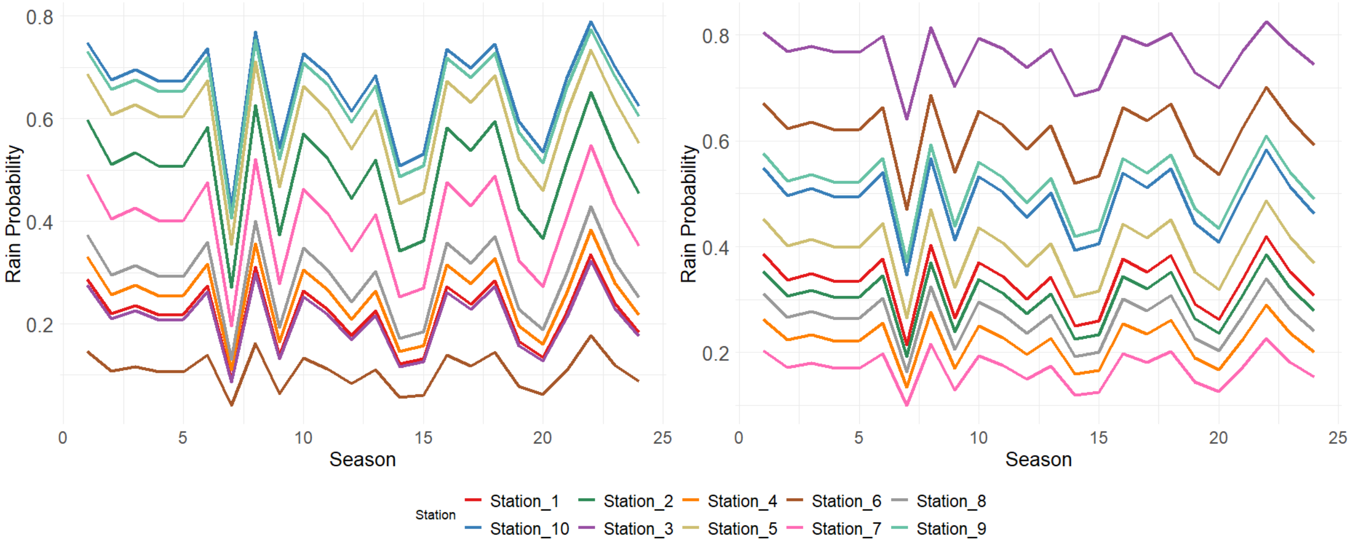


Figure B.2: Rainfall occurrence probabilities per station for states 3 (left) and 4 (right) across the 24 seasons

B.2 Logistic Regression Parameters

The logistic regression model for rainfall occurrence is given by:

$$P(R_t^i = 1 | S_t = s, x_t) = \beta_{0,s,j} + \beta_{1,s} x_t$$

where: - $\beta_{0,s,j}$ represents the intercept for each hidden state s and station j. - $\beta_{1,s}$ is the slope coefficient for each state s and station j. Below we see two tables: the first one, the B.1 with the estimates of intercepts of the model $\beta_{0,s,j}$ and the second one B.2 with the estimated values of the slopes $\beta_{1,s}$, their Se and also their CI. We also calculated these metric so to quantify the uncertainty regarding those estimates.

State	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7	Station 8	Station 9	Station 10
1	-3.183	-2.395	-2.033	-3.802	-1.413	-2.649	-3.062	-3.456	-2.227	-2.318
2	0.641	1.319	1.271	0.494	0.732	0.647	0.737	0.539	2.195	1.934
3	-0.768	-0.928	1.432	-1.418	-0.353	0.693	-1.798	-1.248	0.026	-0.059
4	-1.222	0.027	-1.182	-1.057	0.329	-1.867	-0.453	-0.907	0.623	0.720

Table B.1: Estimated optimal intercept parameters $\beta_{0,s,j}$ for each hidden state and station

State	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6	Station 7	Station 8	Station 9	Station 10
1	0.405	0.472	0.141	0.035	0.257	0.906	0.142	1.010	0.268	0.696
SE	0.099	0.119	0.102	0.109	0.091	0.173	0.109	0.187	0.100	0.102
CI	[0.21, 0.60]	[0.02, 0.48]	[-0.06, 0.34]	[-0.23, 0.19]	[0.34, 0.69]	[0.13, 0.81]	[-0.18, 0.25]	[0.64, 1.38]	[0.20, 0.59]	[-0.06, 0.34]
2	-0.018	0.037	0.186	-0.089	0.067	0.114	0.090	0.184	0.107	0.156
SE	0.109	0.109	0.252	0.237	0.235	0.250	0.267	0.206	0.203	0.222
CI	[-0.23, 0.20]	[-0.18, 0.25]	[-0.31, 0.68]	[-0.55, 0.38]	[-0.39, 0.53]	[-0.35, 0.63]	[-0.43, 0.61]	[0.50, 1.31]	[-0.21, 0.59]	[-0.25, 0.62]
3	0.480	-0.514	0.066	0.612	0.187	-0.238	-0.554	0.997	0.130	0.354
SE	0.131	0.117	0.171	0.138	0.189	0.218	0.218	0.187	0.144	0.131
CI	[0.22, 0.74]	[-0.74, -0.28]	[-0.03, 0.64]	[0.34, 0.88]	[0.39, 1.13]	[-0.98, -0.13]	[-0.98, -0.13]	[0.64, 1.38]	[0.07, 0.64]	[-0.10, 0.41]
4	0.249	0.520	0.396	0.302	0.127	0.186	0.756	0.513	0.364	0.459
SE	0.119	0.091	0.100	0.087	0.093	0.094	0.092	0.091	0.106	0.093
CI	[0.02, 0.48]	[0.34, 0.70]	[0.20, 0.59]	[0.16, 0.57]	[0.09, 0.45]	[-0.08, 0.29]	[0.28, 0.64]	[0.52, 0.87]	[0.16, 0.57]	[0.81, 1.18]

Table B.2: Estimated slopes $\beta_{1,s,j}$, standard errors (SE), and confidence intervals (CI) for each hidden state and station.

Bibliography

[1] Andrew W. Robertson, Sergey Kirshner, and Padhraic Smyth. Daily rainfall occurrence over northeast brazil and its downscalability using a hidden markov model. *International Research Institute for Climate Prediction, The Earth Institute at Columbia University*, 2003. October 21, 2003.

[2] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[3] Timo Koski. *Hidden Markov Models for Bioinformatics*. Kluwer Academic Publishers, Dordrecht, 2001.

[4] M. W. Pedersen, T. A. Patterson, U. H. Thygesen, and H. Madsen. Estimating animal behavior and residency from movement data. *Oikos*, 120(9):1281–1290, 2011.

[5] Roland Langrock, Iain L MacDonald, and Walter Zucchini. Some nonstandard stochastic volatility models and their estimation using structured hidden markov models. *Journal of Empirical Finance*, 19(1):147–161, 2012.

[6] Walter Zucchini, Iain L MacDonald, and Roland Langrock. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC, second edition edition, 2017.

[7] B. H. Juang and Lawrence R. Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[8] Walter Zucchini and Peter Guttorp. A hidden markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923, 1991.

[9] Iain L. MacDonald and Walter Zucchini. *Hidden Markov and Other Models for Discrete-Valued Time Series*. Number 70 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, London, 1997.

[10] J. P. Hughes and P. Guttorp. Incorporating spatial dependence and atmospheric data in a model of precipitation. *Journal of Applied Meteorology*, 33(12):1503–1515, December 1994.

[11] Loukia Meligkotsidou and Petros Dellaportas. Bayesian forecasting using hidden markov models. *Journal of Statistical Computation and Simulation*, 81(1):97–113, 2011.

[12] Paul Taylor. Hidden markov models for grapheme to phoneme conversion. In *INTERSPEECH 2005*, 2005.

[13] J.P. Hughes. *A class of stochastic models for relating synoptic atmospheric patterns to local hydrologic phenomena*. Ph.d. dissertation, Department of Statistics, University of Washington, 1993.

[14] G. Celeux and J.-B. Durand. Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564, 2008. DOI: 10.1007/s00180-007-0106-5.

[15] A. Azzalini and A.W. Bowman. A look at some data on the old faithful geyser. *Applied Statistics*, 39:357–365, 1990.

[16] John A. D. Aston and David Martin. Statistical analysis of the old faithful geyser data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(4):453–460, 2007.

[17] R. Dennis Cook and Sanford Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, 1982.

[18] Roland Langrock. Flexible latent-state modelling of old faithful's eruption intervals. *Australian New Zealand Journal of Statistics*, 54(3):279–296, 2012.

[19] J. P. Hughes, P. Guttorp, and S. P. Charles. A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society Series C Applied Statistics*, 48(1):15–30, 1999.

[20] Walter Zucchini and Peter Guttorp. A hidden markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923, 1991.

[21] Steven C. Amstrup, Trent L. McDonald, and Bryan F.J. Manly. *Handbook of Capture-Recapture Analysis*. Princeton University Press, Princeton, NJ, 2005.

[22] Lauren E Hay, Gregory J McCabe, David M Wolock, and Mark A Ayers. Simulation of precipitation by weather type analysis. *Water Resources Research*, 27(4):493–501, 1991.

[23] András Bárdossy and Erich J Plate. Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, 28(5):1247–1259, 1992.

[24] E. Roeckner and Coauthors. The atmospheric general circulation model echam4: Model description and simulation of present-day climate. Report 23, Max-Planck-Institut für Meteorologie, 1996.

[25] N. I. Ramesh and C. Onof. A class of hidden markov models for regional average rainfall. *Journal of Hydrology*, 247(1-2):105–119, 2001.

[26] O. Cappé, E. Moulines, and T. Rydén. Inference in hidden markov models. *Proceedings of EUSFLAT Conference*, pages 14–16, 2009.

[27] Andrea Martino, Giuseppina Guatteri, and Anna Maria Paganoni. Multivariate hidden markov models for disease progression. Technical Report MOX-Report No. 59/2018, MOX, Dipartimento di Matematica, Politecnico di Milano, 2018.