Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Modelagem de dados pluviométricos no DF por meio de Cadeias de Ordem Variável Estocasticamente Perturbadas

Matheus Ferreira Marques Cavalcante

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção do título de Bacharel em Estatística.

Brasília
2018

MATHEUS FERREIRA MARQUES CAVALCANTE

Aplicações de Cadeias de Ordem Variável Estocasticamente Perturbadas

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos para a obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. **Lucas Moreira**

Brasília
2018

Agradecimentos

Agradeço primeiramente Universidade de Brasília, pelas oportunidades de estudo e aos professores. Agradeço ao professor Dr. Lucas Moreira pelo apoio e a orientação neste trabalho. Finalmente agradeço a minha família e amigos pelo suporte e conselhos em todos esses anos.

Resumo

Neste trabalho modelamos dados pluviométricos no Distrito Federal por meio de Cadeias de Ordem Variável Estocasticamente perturbadas. Inicialmente, estudamos o estimador utilizado neste trabalho, por meio de simulações onde constatamos a eficiência do estimador. Consideramos um alfabeto de tamanho três para os modelos de ordem variável e , para verificar o comportamento do estimador em amostras contaminadas, utilizamos três árvores conhecidas. Para a contaminação da amostra, consideramos dois modelos de pertubação. No primeiro modelo consideramos uma cadeia com alfabeto de tamanho três em que, a cada instante de tempo, um dos símbolos pode ser modicado com uma probabilidade pequena e ?xada. No segundo modelo o processo perturbado assume aleatoriamente o valor da cadeia original ou uma função que depende deste valor, com probabilidade pequena e fixada. Os modelos de contaminação foram comparados através das simulações de amostras perturbadas de processos conhecidos. Pela simplicidade do primeiro modelo de contaminação foi possível recuperar a árvore de contextos do processo original mesmo com alta probabilidade de contaminação. Enquanto o outro modelo, recuperamos a árvore de contextos do processo original apenas quando a probabilidade de perturbação era suficientemente pequena. Em seguida, estudamos o comportamento de chuvas do Distrito Federal, propomos modelos pluviométricos para prever a possibilidade do próximo dia ser sem chuva, chuva moderada ou chuva forte, dado as informações de pluviosidade dos dias anteriores.

Palavras-chave: Cadeias de Ordem Variável, Árvores de Contextos, Modelos de Contaminação estocástica, Modelos Meteorológicos.

Abstract

In this work we model the pluviosity data in Distrito Federal through Stochastically Disturbed Variable Order Chains. Initially, we studied the estimator used in this work, by simulations where we verified the efficiency of the estimator. An alphabet of size three was considered for this work and to verify the behavior of the estimator in contaminated samples, we used three previously known trees. For sample contamination, we considered two perturbation models. In the first model, the value at any instant, can be modified with a small and fixed probability. In the second case, the disturbed process randomly assumes the value of the original chain or a function that depends on this value, with a small and fixed probability. The contamination models were compared by simulations of disturbed samples from known processes. Due to the simplicity of the first contamination model, it was possible to recover a tree of contexts from the original process even with high probability of contamination. While the other model, it retrieved the tree of contexts correctly when it had small values of disturbance. Next, we study the behavior of the rainfalls of the Distrito Federak, and propose pluviometric models to predict the possibility of a next day will not rain or will have moderate rain or heavy rain, given the information of pluviosity from the previous days.

Key Words: Chains of Variable Memory, Context Trees, Contamination Stochastic Models, Meteorological Models.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

Seguindo Quintino e Moreira (2015), a motivação deste trabalho foi estudar Cadeias de Ordem Variável Estocasticamente Perturbadas. Consideramos os modelos de pertubação estocásticas apresentados em Collet, Galves e Leonardi (2008) e por Garcia e Moreira (2015) e propomos, utilizando Cadeias de Ordem Variável, modelos meteorológicos para predizer se o próximo dia choverá ou não dado as informações de pluviosidade dos dias anteriores.

Os modelos onde a ordem de dependência do passado é variável, a porção do passado necessária para predizer o próximo símbolo não é fixa. Essa porção do passado necessária para predizer o próximo símbolo é uma função da sequência dos símbolos passados e estes modelos foram introduzidos por Rissanen (1983) e chamados fontes de memória finita, ou máquinas de árvores. Na literatura estatística recente, estes modelos são chamados *Cadeias de Ordem Variável*.

Rissanen (1983) chamou de *contexto* a porção do passado necessária para predizer o próximo simbolo. O conjunto de todos os contextos pode ser representado por uma árvore probabilística com raiz e rótulos chamada de *árvore de contextos* do processo. Em seu trabalho, Rissanem estudou as cadeias de ordem finita. No entanto, a extensão de um modelo com ordem variável para uma situação não Markoviana, em que os contextos são ainda finitos, porém com comprimento ilimitado, ocorre naturalmente. Com a leitura de trabalhos recentes como Quintino e Moreira (2015), Alex e Moreira (2014) e Denise Duarte e Wecsley O. Prates (2016) é possível fazer um levantamento recente acerca do tema.

Um aspecto vantajoso dos modelos de ordem variável, em relação as Cadeias de Markov de ordem fixa, é a redução do número de parâmetros a serem estimados. Isto ocorre, pois, os modelos de ordem variável levam em conta as dependências estruturais presentes nos dados. Outra característica interessante é que em muitas aplicações, a forma da árvore de contextos tem uma interpretação natural e informativa.

Além de introduzir as Cadeias de Ordem Variável, Rissanen (1983) também propôs um algoritmo para estimar a árvore de contextos, chamado *Algoritmo Contexto*.

Diversos estudos posteriores abordaram a questão da estimação da árvore de contextos para Cadeias de Ordem Variável bem como o correspondente conjunto associado de probabilidades de transição, utilizando variantes do Algoritmo Contexto de Rissanen (1983). Dentre eles destacam-se Bühlmann e Wyner (1999) para o caso de cadeia com ordem limitada, Ferrari e Wyner (2003) para ordem não limitada, o BIC de Csiszar e Talata (2005) e também Duarte et al. (2006) que deram uma majoração para a velocidade de convergência do Algoritmo Contexto para Cadeias de Ordem Variável não limitadas.

Collet, Galves e Leonardi (2008) propuseram um Modelo de Contaminação Estocástica considerando uma Cadeia de Ordem Variável com alfabeto finito em que, a cada instante de tempo, o processo perturbado assume aleatoriamente o valor da cadeia original ou uma função que depende deste valor, com probabilidade pequena e fixada. Além do modelo proposto, provaram que é possível recuperar a árvore de contextos do processo original através de uma amostra contaminada segundo este modelo.

Garcia e Moreira (2015) apresentam um Modelos de Contaminação Estocástica em que consideraram uma Cadeia de Ordem Variável com alfabeto finito que, a cada instante de tempo, um dos símbolos pode ser modificado com uma probabilidade pequena e fixada. Também provaram que utilizando uma amostra contaminada segundo este modelo de contaminação é possível recuperar a árvore de contextos do processo original.

Denise Duarte e Wecsley O. Prates (2016) posteriormente mostraram que é possível recuperar a verdadeira árvore de contextos de uma Cadeia de Ordem Variável Estocasticamente Perturbada e saber o grau de tal perturbação, a depender do grau de perturbação e do regime de perturbação associado.

Através do estudo de simulações verificamos o bom desempenho da versão do Algoritmo Contexto proposta em Galves e Leonardi (2008) quando utilizamos amostras contaminadas, desta forma viabilizando a aplicação em modelos pluviométricos considerados neste trabalho. Para a aplicação de Cadeias de Ordem Variável, modelamos dados pluviométricos do Distrito Federal considerando os regimes de chuva da região e propomos predições do próximo dia ser sem chuva, chuva moderada ou chuva forte. Todas as simulações e estimativas foram realizadas através do ambiente R de computação estatística (R Core Team, 2018).

O presente estudo está organizado da seguinte forma: no Capítulo 1 apresentamos as notações e definições básicas. No Capítulo 2 definimos a versão do Algoritmo Contexto utilizada para estimação da árvore de contextos, assim como os modelos de contaminação considerados nesse trabalho. No Capítulo 3 apresentamos e discutidos os resultados obtidos através das simulações de processos contaminados. No Capítulo 4 modelamos dados pluviométricos do Distrito Federal para aplicação de Cadeias de Ordem Variável baseado no estudo feito do regime chuvas no DF. O Capítulo 5 traz as

conclusões do trabalho.

No Apêndice A apresentamos os códigos desenvolvidos em ambiente R do estimador de árvore de contextos.

Capítulo 1

Revisão Bibliográfica

Neste capítulo definimos formalmente uma Cadeia de Ordem Variável e, para conveniência do leitor, fizemos uma breve revisão das notações e definições que assumimos durante o trabalho.

1.1 Notações e Definições

Considere o alfabeto $\mathcal{A} = \{0, 1, \cdots, N-1\}$ com tamanho $|\mathcal{A}| = N$. Dados dois inteiros $m \leq n$ denotamos a_m^n a sequência de símbolos $a_m, a_{m+1}, \cdots, a_n$ de $\mathcal{A}$ e $\mathcal{A}_m^n$ o conjunto de tais sequências. O comprimento da sequência será $l(a_m^n) = n - m + 1$. Caso $n < m$, $a_m^n = \varnothing$ e $l(a_m^n) = 0$.

O conjunto de todas as sequências semi-infinitas e o conjunto de todas as sequências de símbolos de tamanho finito são denotados, respectivamente, por

$$\mathcal{A}_{-\infty}^{-1} = \mathcal{A}^{\{\cdots,-2,-1\}} \quad e \quad \mathcal{A}^* = \bigcup_{j=0}^{\infty} \mathcal{A}_{-j}^{-1},$$

em que para $j = 0$ corresponde ao conjunto das sequências vazias $\varnothing$.

Dadas duas sequências ω e v, com $l(\omega) < +\infty$, denotamos por $v\omega$ a sequência de comprimento $l(v) + l(\omega)$ obtida pela concatenação das duas sequências. Por exemplo, para $v = \cdots, v_{-n-2}, v_{-n-1}$ e $\omega = \omega_{-n}, \cdots, \omega_{-2}, \omega_{-1}$, a sequência obtida pela concatenação de v e ω será $v\omega = \cdots, v_{-n-2}, v_{-n-1}, \omega_{-n}, \cdots, \omega_{-2}, \omega_{-1}$. Note que, para o caso em que $v = \varnothing$ obtêm-se $v\omega = \varnothing\omega = \omega$. Analogamente ocorre para $\omega = \varnothing$.

Uma sequência u é dita ser um *sufixo* de ω se existir s, com $l(s) \geq 1$, tal que $\omega = su$ e será denotada por $u \prec \omega$. Caso $u \prec \omega$ ou $u = \omega$, será denotado por $u \preceq \omega$. Dada uma sequência finita ω denotamos por suf(ω) o maior sufixo de ω.

Ao longo desse trabalho consideramos o processo $\mathbf{X} = \{X_t,\ t \in \mathbb{Z}\}$ estacionário e ergódico sobre o alfabeto $\mathcal{A} = \{0, 1, \cdots, N-1\}$. Assumimos que o processo

$\mathbf{X}$ é compatível com a probabilidade de transição $p_X(.|.)$, ou seja,

$$p_X(a|\omega) = \mathbb{P}(X_0 = a|X_{-1} = \omega_{-1}, X_{-2} = \omega_{-2}, \cdots), \tag{1.1}$$

para todo $\omega \in \mathcal{A}_{-\infty}^{-1}$ e para todo $a \in \mathcal{A}$. Para $\omega \in \mathcal{A}_{-j}^{-1}$ a probabilidade estacionária do cilindro definida por essa sequência será denotada por

$$\mu_X(\omega) = \mathbb{P}(X_{-j}^{-1} = \omega). \tag{1.2}$$

Com intuito de estimarmos a árvore de contextos de um processo $\mathbf{X}$, dada uma amostra contaminada desse processo, consideramos que $\mathbf{X}$ satisfaz as seguintes definições.

Definição 1.1 *Dizemos que um processo* $\mathbf{X}$ *é não-nulo se satisfaz*

$$\alpha_X = \inf\{p_X(a|\omega) : \ a \in \mathcal{A}, \ \omega \in \mathcal{A}_{-\infty}^{-1}\} > 0. \tag{1.3}$$

Definição 1.2 *Dizemos que um processo* $\mathbf{X}$ *possui taxa de continuidade somável se*

$$\beta_X = \sum_{k \in \mathbb{N}} \beta_{k,X} < +\infty, \tag{1.4}$$

em que a sequência $\{\beta_{k,X}\}_{k\in\mathbb{N}}$ *é definida por*

$$\beta_{k,X} := sup\left\{\left|1 - \frac{p_X(a|\omega)}{p_X(a|v)}\right| : \ a \in \mathcal{A}, \ v, \omega \in \mathcal{A}_{-\infty}^{-1} \ \ com \ \ \omega_{-k}^{-1} = v_{-k}^{-1}\right\}. \tag{1.5}$$

A sequência $\{\beta_{k,X}\}_{k\in\mathbb{N}}$ é chamada *taxa de continuidade do processo* $\mathbf{X}$. Note que, a condição de não-nulidade do processo $\mathbf{X}$ é necessária para que possamos definir a taxa de continuidade do processo por (1.5). A taxa de continuidade é uma propriedade esperada para o processo $\mathbf{X}$, pois, desejamos que dois passados coincidindo nos últimos k símbolos tenham a mesma influência na predição do próximo símbolo da sequência, a medida que k cresce.

Rissanen (1983) chamou de *contexto* a porção do passado necessária para predizer o próximo símbolo do processo, sendo o tamanho desta sequência é função do próprio passado. Um contexto infinito é uma sequência semi infinita tal que nenhum dos seus sufixos é um contexto. O conjunto de todos os contextos satisfaz a propriedade do sufixo, isto é, nenhum contexto é sufixo de outro contexto. Esta propriedade permite representar o conjunto de todos os contextos (finito ou infinito enumerável) como uma árvore probabilística com raiz e rótulos. Esta árvore é chamada *árvore de contextos* do processo $\mathbf{X}$. A seguir definiremos de maneira mais formal um contexto.

Definição 1.3 *Dizemos que uma sequência $\omega \in \mathcal{A}_{-j}^{-1}$ é um contexto do processo* $\mathbf{X}$ *se para toda sequência semi-infinita $x_{-\infty}^{-1} \in \mathcal{A}_{-\infty}^{-1}$ tendo ω como sufixo satisfazer*

$$\mathbb{P}(X_0 = a | X_{-\infty}^{-1} = x_{-\infty}^{-1}) = p_X(a|\omega), \quad \forall a \in \mathcal{A}, \tag{1.6}$$

e nenhum sufixo de ω satisfaz a equação 1.6 .

Denotamos por $d(\mathcal{T})$ a profundidade da árvore $\mathcal{T}$, ou seja,

$$d(\mathcal{T}) := \max\{l(\omega) : \omega \in \mathcal{T}\}.$$

Uma árvore $\mathcal{T}$ é dita *completa* se qualquer sequência em $\mathcal{A}_{-\infty}^{-1}$ pertence a $\mathcal{T}$ ou tem sufixo que pertence a $\mathcal{T}$. Dizemos que a árvore de contextos é *limitada* se o comprimento do maior contexto é finito. Caso contrário, $\mathcal{T}$ é dita *ilimitada*.

Dizemos que uma árvore é *irredutível* se nenhuma sequência pode ser substituída por um sufixo sem violar a propriedade sufixo. Essa noção foi introduzida em Csiszár e Talata (2006) e generaliza o conceito de árvore completa.

A seguir definiremos de maneira mais formal uma *árvore probabilística de contextos* e uma Cadeia de Ordem Variável.

Definição 1.4 *Uma árvore probabilística de contextos em $\mathcal{A}$ é um par ordenado $(\mathcal{T}, \bar{p})$ que satisfaz*

(1) $\mathcal{T}$ é uma árvore irredutível.

(2) $\bar{p} = \{\bar{p}(.|\omega), \omega \in \mathcal{T}\}$ é uma família de probabilidades de transição sobre $\mathcal{A}$.

Definição 1.5 *Dizemos que o processo* $\mathbf{X}$ *é compatível com a árvore probabilística de contextos $(\mathcal{T}, \bar{p})$ se satisfaz*

(1) $\mathcal{T}$ é a árvore de contextos do processo $\mathbf{X}$*.*

(2) Para qualquer $w \in \mathcal{T}$ e $a \in \mathcal{A}$, $p_X(a|\omega) = \bar{p}(a|\omega)$.

Se $\mathbf{X}$ é compatível com a árvore probabilística de contextos $(\mathcal{T}, \bar{p})$, dizemos que $\mathbf{X}$ é uma *Cadeia de Ordem Variável* e denotamos a árvore de contextos de $\mathbf{X}$ por $\mathcal{T}_{\mathbf{X}}$. Note que em (1.4) da Definição (1.2), se $d(\mathcal{T}_{\mathbf{X}}) < +\infty$, então $\beta_{k,X} = 0$ para $k \geq d(\mathcal{T}_{\mathbf{X}})$, ou seja,

$$\beta_X = \sum_{k=0}^{d(\mathcal{T}_{\mathbf{X}})-1} \beta_{k,X} < +\infty.$$

Em alguns casos podemos estar interessados não na árvore de contextos do processo $\mathbf{X}$ mas na utilização desta árvore com uma restrição no tamanho da maior sequência. Seja K esta restrição. Neste caso, chamaremos de árvore truncada no nível $K \geq 1, K \in \mathbb{N}$. Dessa forma, se definirmos $K \geq d(\mathcal{T}_{\mathbf{X}})$, estaremos considerando a própria árvore de contextos do processo $\mathbf{X}$.

Definição 1.6 *Dado um inteiro K, defina a árvore de contextos truncada no nível K por*

$$\mathcal{T}_{\mathbf{X}}|_K = \{\omega \in \mathcal{T}_{\mathbf{X}} : l(\omega) \leq K\} \cup \{\omega : l(\omega) = K \;\; e \;\; \omega \prec u, \; para \; algum \;\; u \in \mathcal{T}_{\mathbf{X}}\}.$$

Considere $\mathbf{Z} = \{Z_t, \; t \in \mathbb{Z}\}$ um processo tomando valores num alfabeto finito $\mathcal{A} = \{0, 1, \cdots, N-1\}$. Seja $Z_1, \cdots, Z_n$ uma amostra aleatória do processo $\mathbf{Z}$. Para toda sequência finita ω, com $l(\omega) \leq n$, denotamos por $N_n(\omega)$ o número de vezes que observou-se a sequência ω na amostra, ou seja,

$$N_n(\omega) = \sum_{t=0}^{n-l(w)} \mathbf{1}_{\{Z_{t+1}^{t+l(\omega)} = \omega\}}. \tag{1.7}$$

Para todo elemento $a \in \mathcal{A}$ e para toda sequência finita ω, a probabilidade de transição empírica é dada por

$$\hat{p}_Z(a|\omega)_n = \frac{N_n(\omega a) + 1}{N_n(\omega .) + |\mathcal{A}|}. \tag{1.8}$$

Observamos que a definição 1.8 implica que $\hat{p}_Z(a|\omega)_n$ é assintoticamente equivalente ao Estimador de Máxima Verossimilhança de $p_Z(a|\omega)_n$.

Antes de apresentar o estimador da árvore de contextos, definido no Capítulo 2, é necessário definirmos o seguinte operador

$$\Delta_n(\omega) := \max_{a \in \mathcal{A}} |\hat{p}_Z(a|\omega)_n - \hat{p}_Z(a|suf(\omega))_n|, \tag{1.9}$$

para qualquer sequência finita $\omega \in \mathcal{A}^*$.O operador $\Delta_n(\omega)$ calcula a distância entre as probabilidades empíricas dado uma sequência ω e a sequência passada $suf(\omega)$.

Exemplo 1.1 (Representação da Árvore de Contextos) *Considere* $\mathbf{X}$ *uma Cadeia de Ordem Variável tomando valores em um alfabeto binário* $\mathcal{A} = \{0, 1\}$ *e com árvore de contextos* $\mathcal{T}_{\mathbf{X}} = \{0, 01, 11\}$ *representada pela Figura 1.1. Podemos representar* $\mathcal{T}_{\mathbf{X}}$ *pela Figura 1.1.*

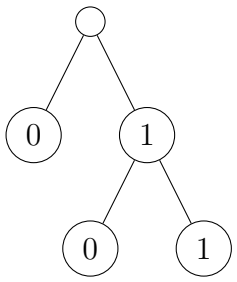

Figura 1.1: Representação da Árvore de Contextos.

A profundidade da árvore de contextos é $d(\mathcal{T}_{\mathbf{X}}) = 2$*, pois, os contextos de maior comprimento são* $\omega = 10$ *e* $v = 11$*. A concatenação destes dois contextos é dada por* $\omega v = 1011$*. A sequência* 1 *é o maior sufixo tanto* ω *como* v*, ou seja,* $suf(\omega) = suf(v) = 1$*. A árvore de contextos truncada no nível* $K = 1$ *é dada por* $\mathcal{T}_{\mathbf{X}}|_K = \{0, 1\}$*, que é equivalente a árvore de contextos de uma Cadeia de Markov de ordem* $k = 1$*.*

Exemplo 1.2 (Estimação das Probabilidades de Transição) *Considere* $\mathbf{X}$ *uma Cadeia de Ordem Variável tomando valores num alfabeto* $\mathcal{A} = \{0, 1\}$ *e com árvore de contextos* $\mathcal{T}_{\mathbf{X}} = \{0, 01, 11\}$*. Seja* 111001011010111 *uma amostra aleatória do processo* $\mathbf{X}$*.*

Note que o número de ocorrências das sequências $\omega = 0$*,* $v = 01$ *e* $u = 11$ *foram dadas, respectivamente, por* $N_{15}(\omega.) = 5$*,* $N_{15}(v.) = 4$ *e* $N_{15}(u.) = 5$*. Com intuito de estimar as probabilidades de transição deste processo é necessário determinar o número de ocorrências da concatenação entre cada contexto com o estado* 0*. Foram obtidas* $N_{15}(00) = 1$*,* $N_{15}(010) = 2$ *,* $N_{15}(110) = 2$*. e* $|\mathcal{A}| = 2$*. As probabilidades de transição estimadas foram* $\hat{p}_X(0|0)_{15} = \frac{N_{15}(00)+1}{N_{15}(0.)+|2|} = \frac{1+1}{5+2} = 0,286$*,* $\hat{p}_X(0|01)_{15} = 0,5$ *e* $\hat{p}_X(0|11)_{15} = 0,428$*. Na Tabela 1.1 trazemos um quadro resumo das probabilidades de transição do processo :*

Tabela 1.1: Tabela das Probabilidades de Transições Estimadas

$\hat{p}_Z(a\|\omega)_{15}$	α	
	0	1
0	0.286	0.714
01	0,5	0.5
11	0,428	0,572

Exemplo 1.3 (Interpretação para Cadeias de Ordem Variável) *Considerando o processo do Exemplo 1.2 representado pela Figura 1.1 com Tabela 1.1.*

Para predição do próximo estado do processo , precisamos olhar para um período anterior, caso o estado observado seja 0 *então não é mais necessário olhar para estados a mais no passado. Especificamente para o processo do Exemplo 1.2, a probabilidade do próximo estado ser* 1 *dado que no estado anterior foi observado* 0 *é de* $\hat{p}_Z(1|0)_{15} = 0.714$*, e* $\hat{p}_Z(0|0)_{15} = 0.286$ *para a probabilidade do próximo estado ser* 0 *. Caso o estado anterior observado seja* 1*, é preciso olhar para mais um período anterior na cadeia para predizermos o próximo estado, seja este estado* 0 *temos* $\hat{p}_Z(1|01)_{15} = 0.5$ *é a probabilidade do próximo estado ser 1 dado que observamos* 1 *e* 0 *no passado,* $\hat{p}_Z(1|11)_{15} = 0.572$ *é a probabilidade do próximo estado ser 1 dado que se observou* 1 *e* 1 *no passado .*

Capítulo 2

Metodologia

Neste capítulo definimos a versão do Algoritmo Contexto proposta por Galves e Leonardi (2008) utilizada nesse trabalho para a estimação da árvores de contextos. Em seguida, apresentamos os Modelos de Contaminação Estocástica descritos por Collet, Galves e Leonardi (2008) e em Garcia e Moreira (2015). Utilizamos o ambiente R de computação estatística (R Core Team, 2018) e a função em R desenvolvida em Alex e Moreira (2015) para programar o estimador de árvore de contextos e os Modelos de Contaminação Estocástica apresentados neste capítulo.

2.1 Uma versão do Algoritmo Contexto

O algoritmo de estimação da árvore de contextos utilizado nesse trabalho foi proposto por Galves e Leonardi (2008) e é uma modificação do Algoritmo Contexto de Rissanem (1983). Primeiramente, considere o operador $\Delta_n(\omega)$ apresentado na Equação (1.9) do Capítulo 1.

Definição 2.1 (Galves e Leonardi, 2008) *Para todo $\delta > 0$ e $d < n$ a árvore de contextos estimada $\hat{\mathcal{T}}_n^{\delta,d}$ é o conjunto contendo todas as sequências $\omega \in \bigcup_{i=1}^{d} \mathcal{A}_{-i}^{-1}$ tais que $\Delta_n(asuf(\omega)) > \delta$ para algum $a \in \mathcal{A}$ e $\Delta_n(u\omega) \leq \delta$ para todo $u \in \bigcup_{i=1}^{d-l(\omega)} \mathcal{A}_{-i}^{-1}$.*

Note na Definição 2.1 que as constantes $\delta > 0$ e $d < n$ são fundamentais para o estimador, pois, inicialmente é considerada a árvore de contextos maximal. Assim, cada sequência ω candidata a contexto possui comprimento $l(\omega) = d$, ou seja, $\omega \in \mathcal{A}_{-d}^{-1}$. Em seguida, o estimador reduz o comprimento das sequências ω que não satisfazem o critério de poda, apresentado na Definição 2.1, tomando $suf(\omega)$ como novo candidato a contexto. Este procedimento é repetido até que a condição de parada seja satisfeita para todas as sequências $\omega \in \hat{\mathcal{T}}_n^{\delta,d}$.

2.2 Modelos de Contaminação Estocástica

Nesta seção apresentamos dois Modelos de Contaminação Estocástica, sendo que um dos modelos foi definido em Collet, Galves e Leonardi (2008) e o outro modelo definido por Garcia e Moreira (2015). Em cada modelo, os autores mostraram que é possível recuperar a árvore de contextos de um processo através de uma amostra contaminada da cadeia segundo um dos modelos de contaminação especificado nos respectivos trabalhos.

Consideramos os processos $\mathbf{X}$ e $\mathbf{Y}$ sendo independentes, não-nulos e com taxa de continuidade somável. Denotamos por $\mathbf{Z} = \{Z_t,\ t \in \mathbb{Z}\}$ os dois processos estocasticamente perturbados. A seguir definimos e comentamos os modelos.

Definição 2.2 (Garcia e Moreira, 2015) *Considere* $\mathbf{X} = \{X_t,\ t \in \mathbb{Z}\}$ *um processo estacionário e ergódico tomando valores num alfabeto* $\mathcal{A} = \{0, 1, \cdots, N-1\}$ *com tamanho* $|\mathcal{A}| = N$. *Seja* $\boldsymbol{\xi} = \{\xi_t, t \in \mathbb{Z}\}$ *uma sequência de variáveis aleatórias Bernoulli i.i.d. tomando valores em* $\{0, 1\}$, *independentes do processo* $\mathbf{X}$, *com*

$$\mathbb{P}(\xi_t = 1) = 1 - \varepsilon,$$

em que ε *é um parâmetro de pertubação fixado em (0,1). Definimos o Modelo de Contaminação Zero Inflado por*

$$Z_t = X_t.\xi_t, \quad t \in \mathbb{Z}. \tag{2.1}$$

Através da Definição 2.2 podemos observar que, no modelo Zero Inflado, a perturbação pode ocorrer apenas quando $X_t \geq 1$ e $\xi_t = 0$. No entanto, no modelo apresentado a seguir, em qualquer instante de tempo, a perturbação pode ocorrer para todos dos estados do processo $\mathbf{X}$.

Definição 2.3 (Collet, Galves e Leonardi, 2008) *Considere* $\mathbf{X} = \{X_t,\ t \in \mathbb{Z}\}$ *um processo estacionário e ergódico tomando valores num alfabeto* $\mathcal{A} = \{0, 1, \cdots, N-1\}$ *com tamanho* $|\mathcal{A}| = N$. *Seja* $\boldsymbol{\xi} = \{\xi_t,\ t \in \mathbb{Z}\}$ *uma sequência de variáveis aleatórias i.i.d. tomando valores em* $\{0, 1\}$, *independentes de* $\mathbf{X}$, *com*

$$\mathbb{P}(\xi_t = 0) = 1 - \varepsilon,$$

em que ε *é um parâmetro de perturbação fixado em* $(0, 1)$. *Definimos o Modelo de Contaminação por Congruência por*

$$Z_t \equiv X_t + \xi_t (mod|\mathcal{A}|), \tag{2.2}$$

em que (2.2) denota a função de congruência módulo $|\mathcal{A}|$.

Garcia e Moreira (2015) mostraram que o estimador apresentado na Definição 2.1 deste capítulo é robusto, ou seja, se considerarmos os Modelo de Conta-

minação Zero Inflado apresentado na Definição 2.2 , mesmo se na estimação utilizarmos uma amostra perturbada do processo, o estimador consegue recuperar a árvore de contextos do processo original. Collet, Galves e Leonardi (2008) mostraram a robustez deste estimador quando consideramos uma amostra perturbada do processo segundo o Modelo de Contaminação por Congruência.

Capítulo 3

Estudos Simulados de Processos de Ordem Variável

Neste capítulo avaliamos o desempenho do estimador de árvore de contextos, apresentado na Definição 2.1, em amostras contaminadas segundo os modelos de Contaminação Zero Inflado e por Congruência descritos, respectivamente, nas Definições 2.2 e 2.3. Antes de partimos para o estudo simulado de Processos Contaminados, uma análise da constante de poda δ foi feita para verificar o comportamento da mesma quando alteramos os processos e para observamos onde obtemos mais precisão no estimador. Em seguida, utilizando o valor mais adequado de δ, para verificar a eficácia do estimador, utilizamos amostras de tamanhos diferentes e valores de pertubação ϵ diferentes. Durante o estudo dos modelos de Contaminação verificamos a robustez do estimador utilizado neste trabalho.

3.1 Cadeias de Ordem Variável não Contaminadas

Nesta seção verificamos o comportamento da constante de poda do estimador de árvore de contextos δ, apresentado na Definição 2.1, utilizando simulações de amostras não contaminadas de Cadeias de Ordem Variável. A seguir consideramos três Cadeias de Ordem Variável, estacionárias e ergódicas tomando valores no alfabeto $\mathcal{A} = \{0, 1, 2\}$. As Figuras 3.1,3.2 e 3.3 apresentam as árvores de contextos desses processos.

Com o intuito de estudarmos o comportamento da constante de poda δ, na Definição 2.1. Realizamos 100 repetições para as Árvores 1, 2 e 3 onde consideramos que $\delta \in \{0.001, 0.011, 0.021, ..., 0.501\}$. Verificamos o percentual de retorno correto para cada um dos cenários, ou seja, quantas vezes a estimação da correspondente árvore foi correta, para um dado valor de δ. Para a Árvore 1, as Figuras 3.4, 3.5 e 3.6, trazem o percentual de estimações corretas para cada um dos possíveis valores de δ, para os

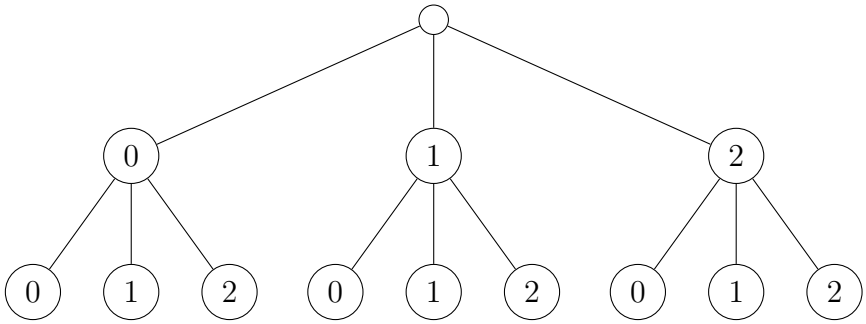

Figura 3.1: Árvore de Contextos 1

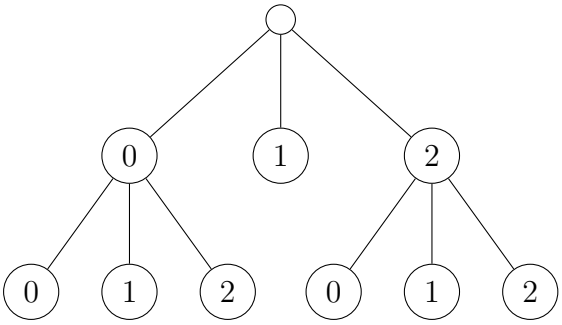

Figura 3.2: Árvore de Contextos 2

tamanhos de $n = 500$, $n = 1000$ e $n = 10000$, respectivamente. Um procedimento análogo foi feito para as Árvores 2 e 3. As Figuras 3.7, 3.8 e 3.9 trazem os resultados correspondentes à Árvore 2, enquanto as Figuras 3.10, 3.11 e 3.12 correspondem à Árvore 3.

Para a Árvore 1, observamos que o estimador é mais eficiente para pequenos valores de δ, enquanto para a Árvore 2, as simulações revelam que o critério de poda só é sensível para valores de δ próximos de 0.18. Tais comportamentos podem ser explicados por meio do operador $\Delta_n(\omega)$, que calcula a diferença entres as probabilidades de transição empíricas dado uma sequência e o seu maior sufixo. Baseado neste operador, o estimador $\hat{\mathcal{T}}_n^{\delta,d}$ inicialmente considera uma árvore completa com profundidade d, podando-a de acordo com o limiar δ. Como a Árvore 1 é completa e de profundidade 3, esperamos que o estimador seja mais eficaz para pequenos valores de δ, uma vez que ao aumentamos o valor de δ o estimador tende a podar mais galhos da Árvore inicial. No caso da Árvore 2 que não é completa, esperamos que o estimador seja mais eficaz a medida que o limiar δ aproxima-se de uma constante, que neste caso é 0.18. O comportamento do estimador no caso da Árvore 3, em termos dos valores de δ, é semelhante ao comportamento no caso da Árvore 2.

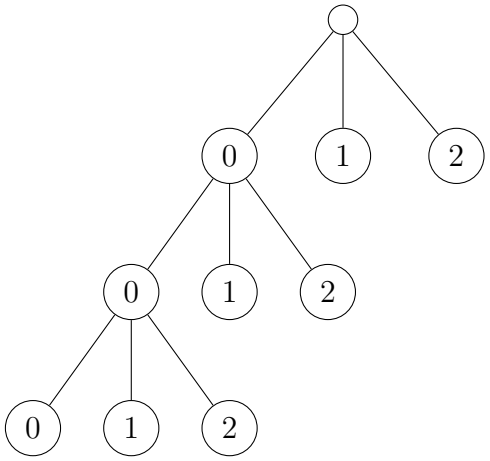

Figura 3.3: Árvore de Contextos 3

3.2 Cadeias de Ordem Variável Contaminadas

Nesta seção contaminamos os processos correspondes às Árvores 1, 2 e 3 da Seção 3.1 segundo os modelos de contaminação apresentados na Seção 2.2. Para a estimação de tais processos por meio de amostras contaminadas utilizamos os valores mais adequados de δ, conforme analisamos na Seção 3.1. Realizamos 100 repetições das simulações para cada uma das árvores de contextos e contaminamos as amostras considerando diferentes valores do parâmetro de perturbação ε, para amostras de tamanhos 10.000 e 100.000. Após as replicações, avaliamos a proporção de retornos corretos do estimador $\hat{\mathcal{T}}_n^{\delta,d}$ das árvores de contextos em cada um dos cenários. Os resultados obtidos considerando o Modelo de Contaminação Zero Inflado, são apresentados nas Tabelas 3.1 e 3.2. Enquanto os resultados referentes ao Modelo de Contaminação por Congruência, seguindo a mesma metodologia, são apresentados nas Tabelas 3.3 e 3.4.

Tabela 3.1: Proporção de retornos, modelo de contaminação Zero Inflado, ε fixado, $n = 10.000$ e 100 repetições.

Árvore de Contextos	Parâmetro de Perturbação (ε)				
	0,001	0,01	0,05	0,1	0,20
Árvore 1	1,00	1,00	1,00	0,70	0,02
Árvore 2	1,00	1,00	1,00	1,00	0,96
Árvore 3	0,86	0,77	0,59	0,44	0,12

A Tabela 3.1 apresenta a proporção de retornos corretos da árvore de contextos do processo original, segundo o modelo Zero Inflado, para diferentes valores do parâmetro de perturbação, com tamanho de amostra $n = 10.000$ e 100 repetições. Para as simulações de amostras contaminadas da Árvore 2, considerando as probabilidades

Figura 3.4: Proproções de acerto para a Àrvore 1 N=500

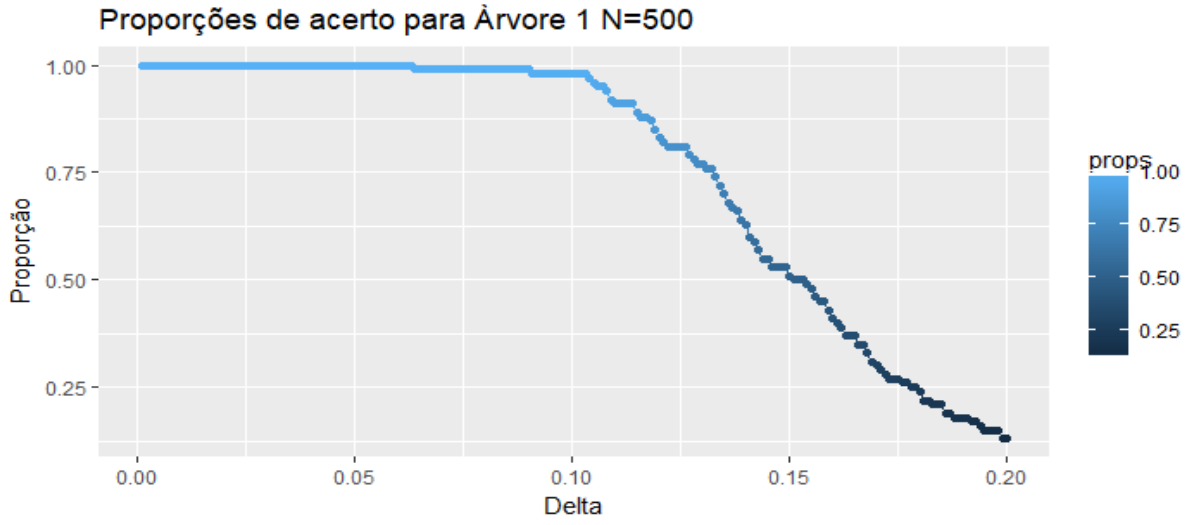

Figura 3.5: Proproções de acerto para a Àrvore 1 N=1000

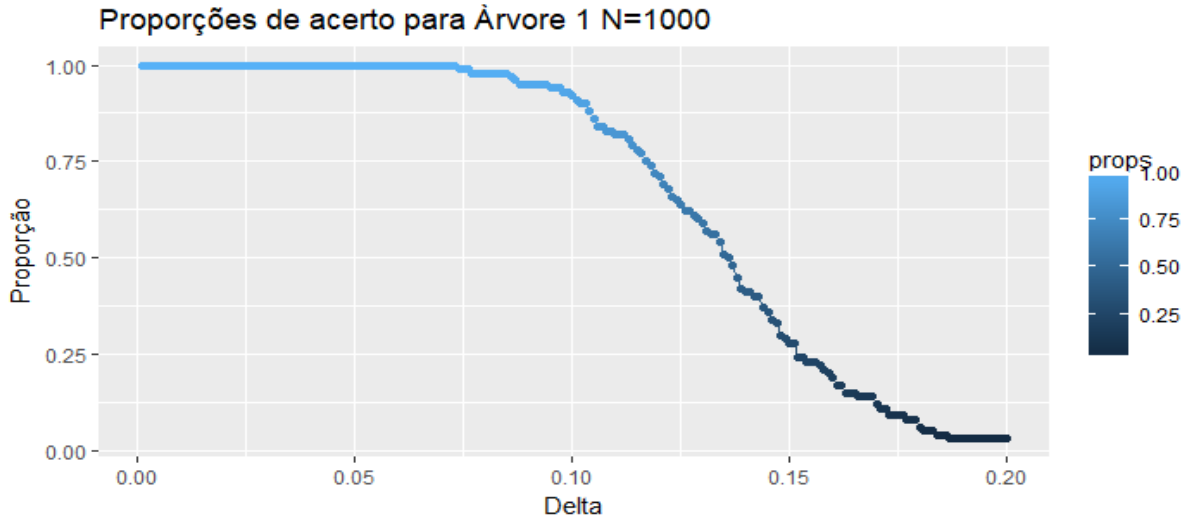

Figura 3.6: Proproções de acerto para a Àrvore 1 N=10000

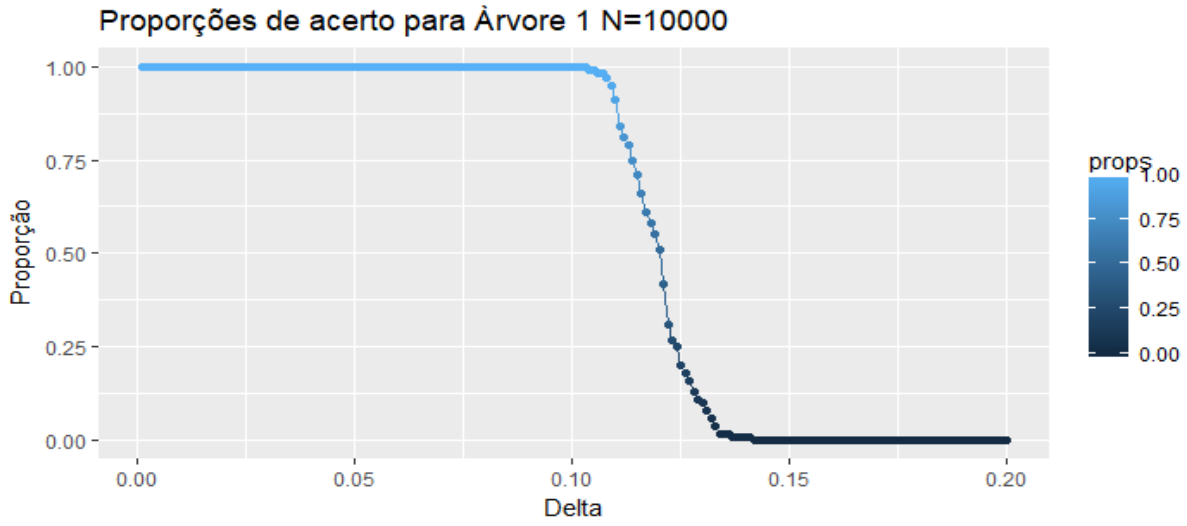

Figura 3.7: Proproções de acerto para a Àrvore 2 N=500

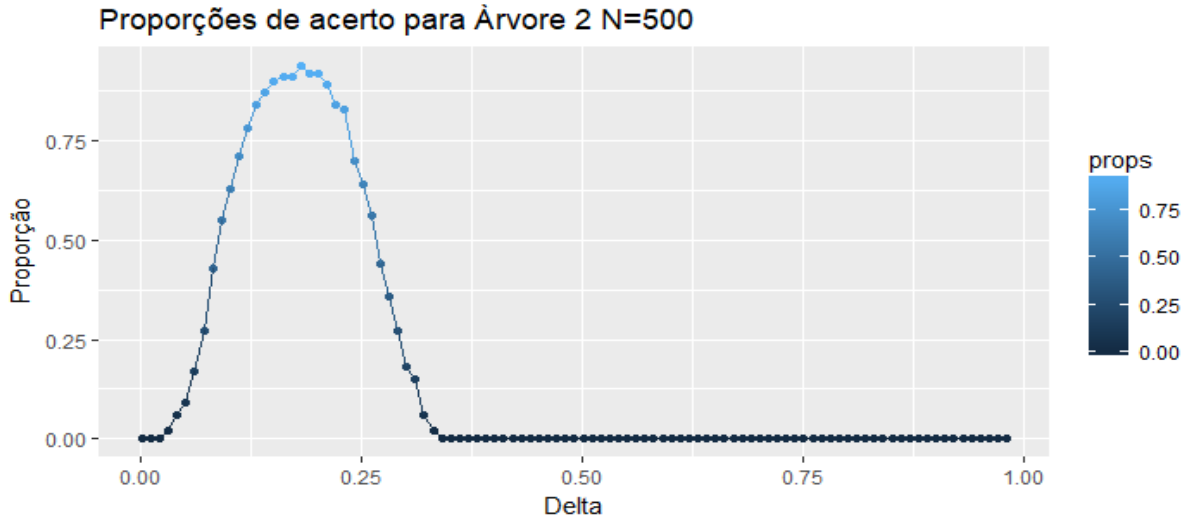

Figura 3.8: Proproções de acerto para a Àrvore 2 N=1000

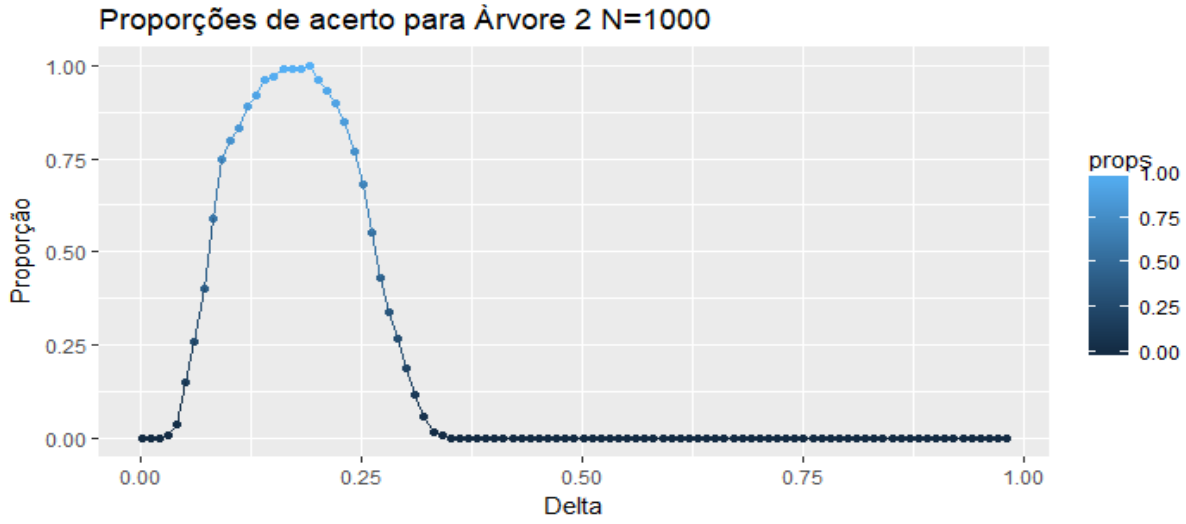

Figura 3.9: Proproções de acerto para a Àrvore 2 N=10000

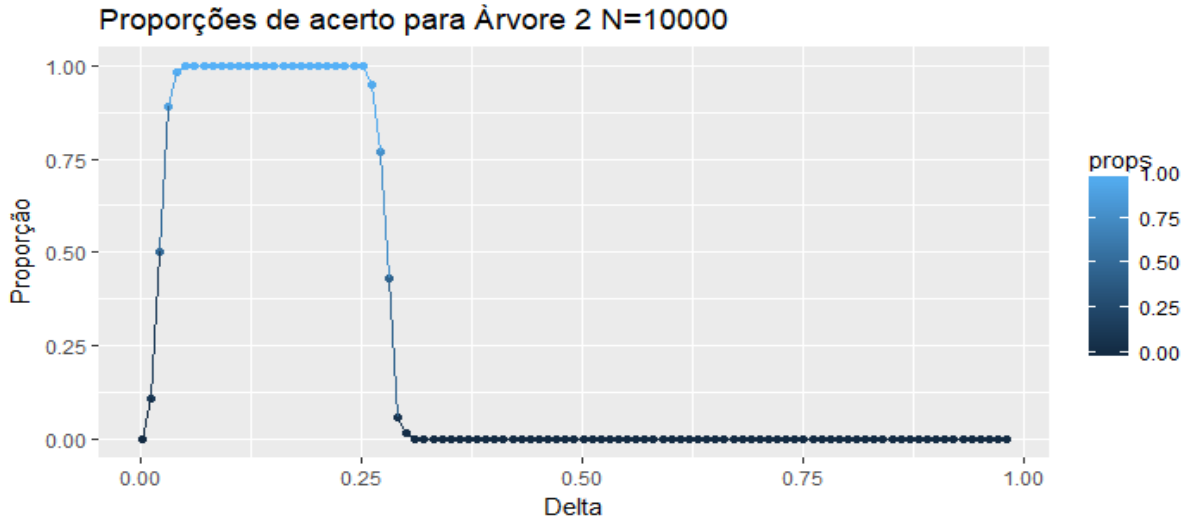

Figura 3.10: Proproções de acerto para a Àrvore 3 N=500

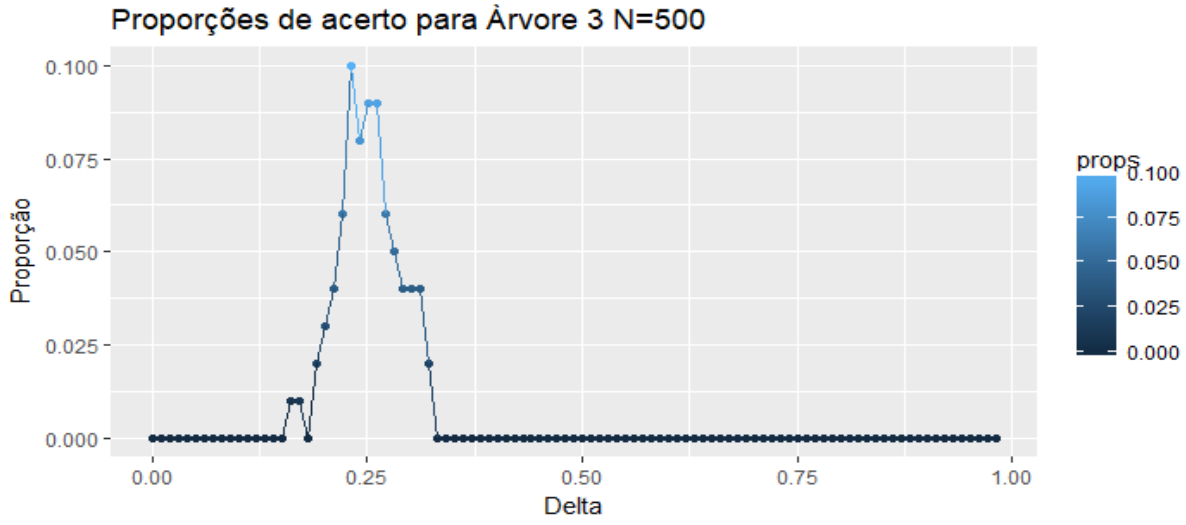

Figura 3.11: Proproções de acerto para a Àrvore 3 N=1000

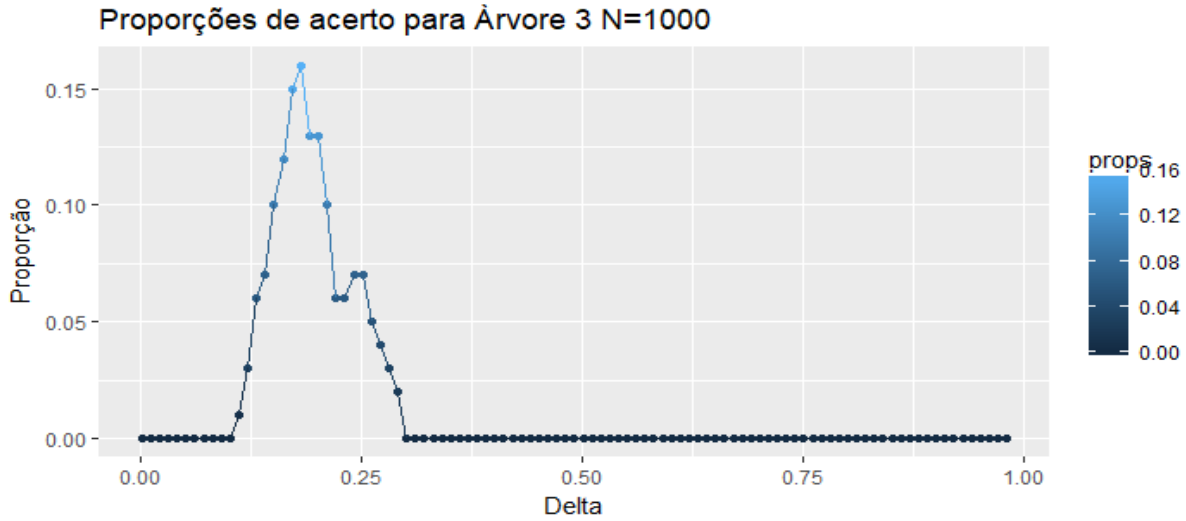

Figura 3.12: Proproções de acerto para a Àrvore 3 N=10000

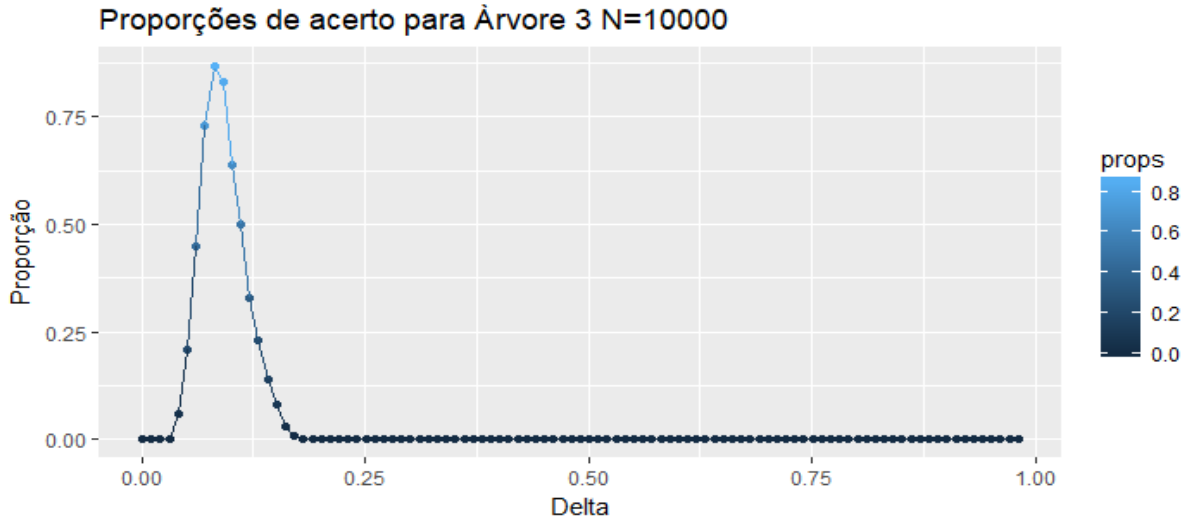

Tabela 3.2: Proporção de retornos, modelo de contaminação Zero Inflado, ε fixado, $n = 100.000$ e 100 repetições.

Árvore de Contextos	Parâmetro de Perturbação (ε)				
	0,001	0,01	0,05	0,1	0,20
Árvore 1	1,00	1,00	1,00	0,92	0,00
Árvore 2	1,00	1,00	1,00	1,00	1,00
Árvore 3	1,00	1,00	0,92	0,46	0,01

Tabela 3.3: Proporção de retornos, modelo de contaminação Congruência, ε fixado, $n = 10.000$ e 100 repetições.

Árvore de Contextos	Parâmetro de Perturbação (ε)				
	0,001	0,01	0,05	0,1	0,20
Árvore 1	1,00	1,00	0,93	0,19	0,00
Árvore 2	1,00	1,00	1,00	0,44	0,00
Árvore 3	0,85	0,79	0,45	0,10	0,00

de contaminação fixadas com $\varepsilon \leq 0,1$ a proporção de retorno correto foi igual ou maior do que o para as simulações da Árvore 1, sendo que ambas as árvores possuem a mesma profundidade. Note que as simulações para a Árvore 3 não atingem 100% de acerto para tamanhos de amostra 10.000. Isso se deve ao fato da grande diferença entre o número de parâmetros da árvore real e o número de parâmetros a serem podados pelo estimador da árvore completa inicial. É grande a quantidade de parâmetros a serem podados relativa ao tamanho da amostra, por isso não há informação suficiente em 10.000 observações para o estimador distinguir se um parâmetro deve ser podado ou não.

Aumentando-se o tamanho de amostra, em cada repetição de 10.000 para 100.000, considerando ainda o modelo Zero Inflado, podemos observar através da Tabela 3.2 que as simulações da Árvore 3, o estimador retornou corretamente a árvore de contexto em todas as simulações em que $\varepsilon < 0,1$ enquanto as simulações para a Árvore 2 retornaram corretamente o árvore de contextos para altos níveis de contaminação, em específico $\varepsilon = 0.2$.

Para as simulações de amostras contaminadas das Árvores 1, 2 e 3 e considerando o modelo Zero Inflado, através das Tabelas 3.1 e 3.2 observamos que o estimador da árvore de contextos aumentou a proporção de retorno correto a medida que o tamanho de amostra aumentou. Porém, mesmo com $n = 100.000$ o estimador apresentou baixas proporções de retornos corretos para as simulações da Árvore 3 quando fixamos $\varepsilon \geq 0,10$.

Apesar das Árvores 1 e 2 possuírem a mesma profundidade $d = 2$, as simulações das amostras contaminadas da Árvore 2 foram aquelas que apresentaram o

Tabela 3.4: Proporção de retornos, modelo de contaminação Congruência, ε fixado, $n = 100.000$ e 100 repetições.

Árvore de Contextos	Parâmetro de Perturbação (ε)				
	0,001	0,01	0,05	0,10	0,20
Árvore 1	1,00	1,00	1,00	0,00	0,00
Árvore 2	1,00	1,00	1,00	0,46	0,00
Árvore 3	1,00	1,00	0,14	0,00	0,00

nível de retorno correto maior, quando o parâmetro de perturbação fixado foi $\varepsilon \geq 0,10$.

As Tabelas 3.3 e 3.4 apresentam a proporção de retornos corretos considerando o modelo de Contaminação por Congruência assumindo níveis do parâmetro de perturbação ε fixados, realizadas 100 repetições, para $n = 10.000$ e $n = 100.000$, respectivamente.

As simulações de amostras contaminadas das Árvores 1 e 2 com $n = 10.000$, considerando o modelo de contaminação por Congruência, apresentaram alta proporção de retornos corretos com o parâmetro de perturbação fixado em até $\varepsilon \leq 0,05$. As simulações de amostras contaminadas das Árvores 1 e 2 para $\varepsilon > 0,1$ apresentaram baixa proporção de retornos corretos mesmo quando $n = 100.000$. A Árvore 3 só apresenta proporção alta de retornos para baixos níveis de contaminação, mesmo aumentando o tamanho da amostra.

Podemos resumir os resultados apresentados nesta seção do seguinte modo: o modelo de Contaminação Zero Inflado apresentou maiores proporções de retornos corretos que os Modelos de Contaminação por Congruência, mesmo na presença de alta probabilidade de contaminação. Este fato ocorreu devido a simplicidade do modelo Zero Inflado uma vez que, a cada instante de tempo, o processo original pode ser contaminado, com probabilidade pequena e fixa, apenas se o símbolo do processo nesse instante de tempo for igual a 1 ou 2. Por outro lado, o modelo de Contaminação por Congruência , a cada instante de tempo, pode contaminar a amostra para todos os símbolos do alfabeto, com probabilidade pequena e fixa.

Podemos ver que para contaminações pequenas o estimador é capaz de identificar a árvore de contextos corretamente, entretanto, quando se tem uma grande discrepância entre a árvore real e a árvore truncada cheia, o estimador pode não ser tão efetivo para amostras pequenas como é o caso da árvore 3.3. Destacamos a robustez e o bom comportamento do estimador de árvores de contextos $\hat{\mathcal{T}}_n^{\delta,d}$ dada uma amostra contaminada do processo. Dessa forma, viabilizamos a aplicação destes modelos de contaminação na modelagem de dados meteorológicos apresentados no próximo capítulo.

Capítulo 4

Aplicação de Cadeias de Ordem Variável

Neste capítulo fizemos um estudo geral do Regime de Chuvas no Distrito Federal, e propomos modelos pluviométricos para a probabilidade do próximo dia não chover, chover moderadamente ou chover forte.

4.1 Regime de Chuvas no Distrito Federal

O clima de Brasília é tropical com estação seca, com temperaturas médias mensais superiores a 18ºC e índice pluviométrico de 1480 milímetros (mm) anuais concentrados entre os meses de outubro a abril. Durante a estão seca , compreendida entre os meses de maio a setembro é comum os niveis de umidade relativa dor ar ficarem abaixo de 30%. A Figura 4.1 apresentada a seguir traz um quadro resumo do clima de Brasília.

Figura 4.1: Quadro Resumo dos Dados climatológicos para Brasília.

Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Ano
Temperatura máxima recorde (°C)	32,6	31,4	32,1	31,6	30,2	31,6	30,8	33	35,8	36,4	34,5	33,7	36,4
Temperatura máxima média (°C)	26,5	27	26,7	26,6	25,9	25	25,3	26,9	28,4	28,2	26,7	26,3	26,6
Temperatura média (°C)	21,6	21,7	21,6	21,3	20,2	19	19	20,6	22,2	22,4	21,5	21,4	21,4
Temperatura mínima média (°C)	18,1	18	18,1	17,5	15,6	13,9	13,7	15,2	17,2	18,1	18	18,1	16,8
Temperatura mínima recorde (°C)	12,2	11	14,5	10,7	3,2	3,3	1,6	5	9	10,2	11,4	13,5	1,6
Precipitação (mm)	209,4	183	211,8	133,4	29,7	4,9	6,3	24,1	46,6	159,8	226,6	241,5	1 477,4
Dias com precipitação (≥ 1 mm)	17	14	14	8	3	1	1	2	5	11	17	19	112
Umidade relativa (%)	76,2	74,7	76,8	72,2	66,2	58,7	52,7	46,8	50,3	62,8	74,5	78	65,8
Horas de sol	150,9	158,9	166,5	204,6	239,5	254,3	268,9	264,4	210,5	183,1	139,9	126,8	2 368,3

Observamos que 92% das chuvas ocorridas no Distrito Federal se encontram entre os meses de Outubro e Abril, o que caracteriza a estação chuvosa em Brasília.

Com base nessas informações propomos modelos de previsão para a pluviosidade do Distrito Federal.

4.2 Modelos de Cadeias de Ordem Variável

Para a aplicação de Cadeias de Ordem Variável utilizaremos dados pluviométricos diários nos quais constam a medição da pluviosidade em milímetros(mm) do Distrito Federal para cada dia entre as datas de 22/08/1961 à 07/05/2018 . Os dados podem ser acessados através do portal eletrônico do INMET[1] para a estação BRASILIA - DF (OMM: 83377).

O critério que adotamos para distinguir entre dias sem chuva, chuva moderada ou chuva forte foi: se a precipitação daquele dia for menor que 1mm de chuva então será considerado como sem chuva, se a precipitação for entre 1mm e menor que 25mm de chuva o dia será considerado chuva moderada, e acima de 25mm de chuva o dia é considerado como chuva forte. Assim, a cada instante de tempo o processo $\mathbf{X}$ assume 1 se choveu entre 1mm e 25mm, e 2 se choveu mais que 25mm. Para a aplicação consideramos o alfabeto $\mathcal{A} = \{0, 1, 2\}$ para o processo $\mathbf{X}$. O tamanho da amostra foi de $n = 20.415$.

Para a estimação das árvores de contextos $\mathcal{T}_{\mathbf{X}}$ do processo $\mathbf{X}$ utilizando a versão do Algoritmo Contexto apresentada na Definição 2.1 do Capítulo 2, fixamos os parâmetros δ,d e n necessários para o estimador $\hat{\mathcal{T}}_n^{\delta,d}$, ou seja, a profundidade das árvores d e o parâmetro $\delta > 0$.

Considerando o período histórico de 1961 à 2018,as Tabelas 4.1 e 4.2 trazem as probabilidades de transição estimadas para ás Árvores estimadas representadas pelas Figuras 4.2 e 4.3, onde consideramos a profundidade $d = 2$ e $d = 3$.

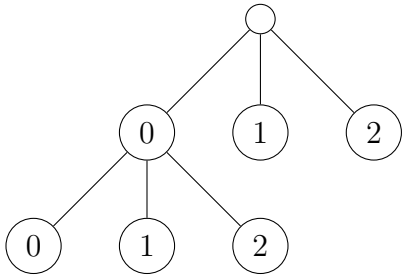

Figura 4.2: Árvore de contextos estimada com profundidade $d = 2$.

Uma característica dos modelos de ordem variável é que em muitas aplicações a forma da árvore de contextos tem uma interpretação natural e informativa. Pela árvore de contextos estimada e apresentada na Figura 4.2, podemos predizer se o

[1]http://www.inmet.gov.br/portal/

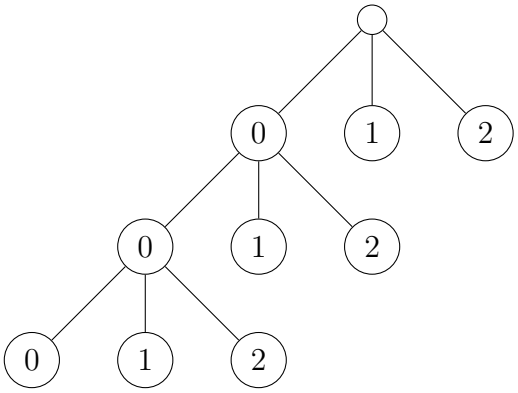

Figura 4.3: Árvore de contextos estimada com profundidade $d = 3$.

próximo dia irá chover ou não, apenas considerando no máximo as informações dos dois dias anteriores. Por exemplo, considerando a árvore da Figura 4.2 e a Tabela 4.1, a probabilidade de que hoje chova moderadamente dado que ontem choveu moderadamente é de $0,47$. Por outro lado, a probabilidade de que hoje chova forte dado que ontem choveu foi de $0,096$. Se considerarmos que ontem não choveu, para a probabilidade estimada de que o dia presente chova é necessário olhar mais um dia antes não sendo necessário olhar para mais dias anteriores.

Tabela 4.1: Tabela das Probabilidades de Transições Estimadas da Figura 4.2.

$\hat{p}_Z(a\|\omega)$	α		
	0	1	2
00	0.877	0.104	0.017
10	0.587	0.338	0.074
20	0.532	0.393	0.074
1	0.433	0.470	0.096
2	0.366	0.504	0.130

Interpretações semelhantes podem ser feitas para a árvore da Figura 4.3 com a Tabela 4.2.

Outra característica dos modelos de ordem variável é a representação gráfica do processo em forma de uma árvore, nos dá a informação de como o processo se comporta, ou se sofreu alteraçõe no decorrer do regime. Para ilustrar essa característica analisaremos se houve mudança no processo de chuva no DF ao analisarmos o dados de chuva até 2000, e comparar-los com os dados posteriores à 2000. Considerando $d = 2$, as árvores estimadas para os anos até 2000 e pós 2000 é a mesma para ambos os períodos e está representada pela Figura 4.4. As Tabelas 4.3 e 4.4 representam as probabilidades estimadas dos contexto para os períodos antes de 2000 de após 2000, respectivamente.

Considerando $d = 3$ para o mesmo cenário, as árvores são representadas

Tabela 4.2: Tabela das Probabilidades de Transições Estimadas da Figura 4.3.

$\hat{p}_Z(a\|\omega)$	α		
	0	1	2
000	0.9049437	0.08236305	0.01269326
100	0.6930380	0.25712025	0.04984177
200	0.6153846	0.32307692	0.06153846
10	0.5876337	0.3384472	0.07391911
20	0.5327869	0.3934426	0.07377049
1	0.4338443	0.4701493	0.09600645
2	0.3660000	0.5040000	0.13000000

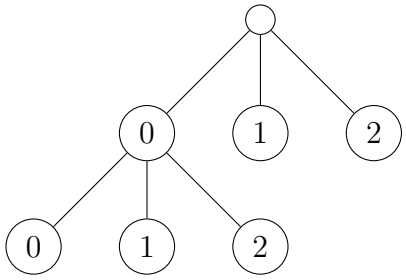

Figura 4.4: Árvore de contextos estimada com profundidade $d = 2$.

pelas figuras 4.5, 4.6 e as probabilidades estimada dos contextos pelas tabelas 4.5 e 4.6.

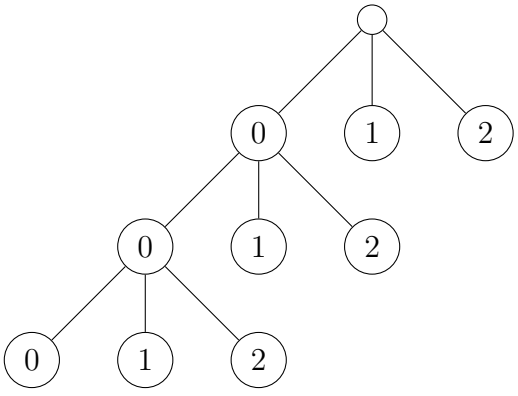

Figura 4.5: Árvore de contextos estimada com profundidade $d = 3$ Período até 2000.

Pelas Tabelas 4.2, 4.3 e pela Figura 4.4 podemos ver que o processo não sofre alterações nos períodos considerados, além dos contextos terem a mesma probabilidade quando consideramos $d = 2$ também. Além disso, o processo se é idêntico ao processo de quando consideramos toda a série histórica. Porém, note que, quando consideramos $d = 3$, a árvore é diferente para os períodos antes de 2000 e após 2000. Para determinar se um dia irá chover forte é necessário olhar para mais dias anteriores, o que não acontece com o período até 2000. Mais especificamente, precisamos ver se no dia anterior choveu ou não, caso não, precisamos ir mais um dia antes.

Tabela 4.3: Tabela das Probabilidades de Transições Estimadas Período até 2000

$\hat{p}_Z(a\|\omega)$	α		
	0	1	2
00	0.8743913	0.1075041	0.01810463
10	0.5976536	0.3326432	0.06970324
20	0.5408560	0.3852140	0.07392996
1	0.4360518	0.4679506	0.09599759
2	0.3813056	0.4851632	0.13353116

Tabela 4.4: Tabela das Probabilidades de Transições Estimadas Período pós 2000.

$\hat{p}_Z(a\|\omega)$	α		
	0	1	2
00	0.8847035	0.0992619	0.01603461
10	0.5669516	0.3504274	0.08262108
20	0.5137615	0.4128440	0.07339450
1	0.4298836	0.4739743	0.09614207
2	0.3343558	0.5429448	0.12269939

De acordo com o INMET[2], para o Distrito Federal, e como vistos na Tabela 4.1 da sessão anterior os dias compreendidos entre os meses de Outubro e Abril é onde se apresenta o maior volume de chuva, concentrando, em média, 92% da precipitação acumulada para o ano, referida como estação chuvosa, e os outros meses como estação de estiagem. Com base nessas informações modelamos a série história para estimarmos o processo de chuva no período de estiagem e de chuva. As figuras 4.7 e 4.8 representam as árvores estimadas para os respectivos períodos considerando $d = 2$. Para $d = 3$ as árvores estimadas são as Figuras 4.9 e 4.10.

Através das Figuras 4.7 e 4.8, podemos ver que durante os períodos de chuvas, para se predizer se o próximo dia será chuvoso, basta apenas olharmos para o dia anterior, tanto na estimação usando $d = 2$ e $d = 3$. Contudo, para os períodos estiagem, ao olhar o passado da cadeia, para predizermos o próximo precisamos verificar se nos três dias anteriores choveram ou não, e caso não tenha chovido, não é preciso mais olhar para o passado conforme as árvores das figuras 4.9 e 4.10 indicam.

[2]http://www.inmet.gov.br/portal/

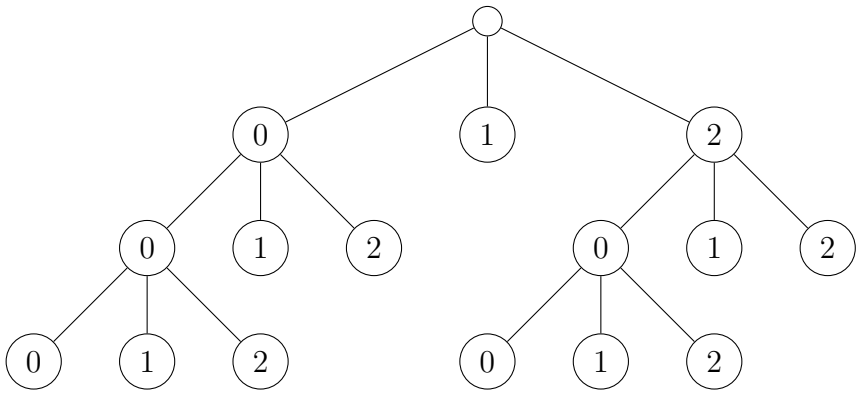

Figura 4.6: Árvore de contextos estimada com profundidade $d = 3$ Período pós 2000.

Tabela 4.5: Tabela das Probabilidades de Transições Estimadas da Figura 4.5.

$\hat{p}_Z(a\|\omega)$	α		
	0	1	2
000	0.9030416	0.08236305	0.01269326
100	0.6882217	0.25712025	0.04984177
200	0.5899281	0.32307692	0.06153846
10	0.5976536	0.3384472	0.07391911
20	0.5408560	0.3934426	0.07377049
1	0.4360518	0.4701493	0.09600645
2	0.3813056	0.5040000	0.13000000

Tabela 4.6: Tabela das Probabilidades de Transições Estimadas da Figura 4.6

$\hat{p}_Z(a\|\omega)$	α		
	0	1	2
000	0.9087770	0.07942446	0.01179856
100	0.7035176	0.24623116	0.05025126
200	0.6785714	0.28571429	0.03571429
10	0.5669516	0.3504274	0.08262108
20	0.5137615	0.4128440	0.07339450
1	0.4301471	0.4742647	0.09558824
002	0.3650794	0.42857143	0.20634921
102	0.4655172	0.43103448	0.10344828
202	0.1250000	0.75000000	0.12500000
12	0.3057325	0.5859873	0.10828025
22	0.2500000	0.6750000	0.07500000

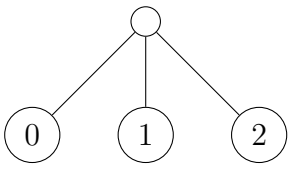

Figura 4.7: Árvore de contextos estimada com profundidade $d = 2$ Período de Chuvas

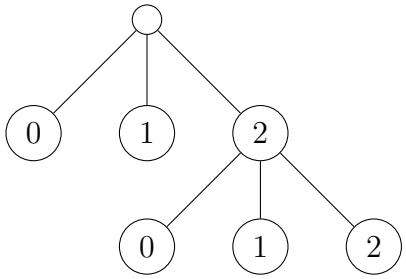

Figura 4.8: Árvore de contextos estimada com profundidade $d = 2$ Período de Estiagem

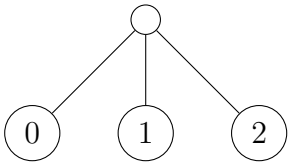

Figura 4.9: Árvore de contextos estimada com profundidade $d = 3$ Período de Chuvas

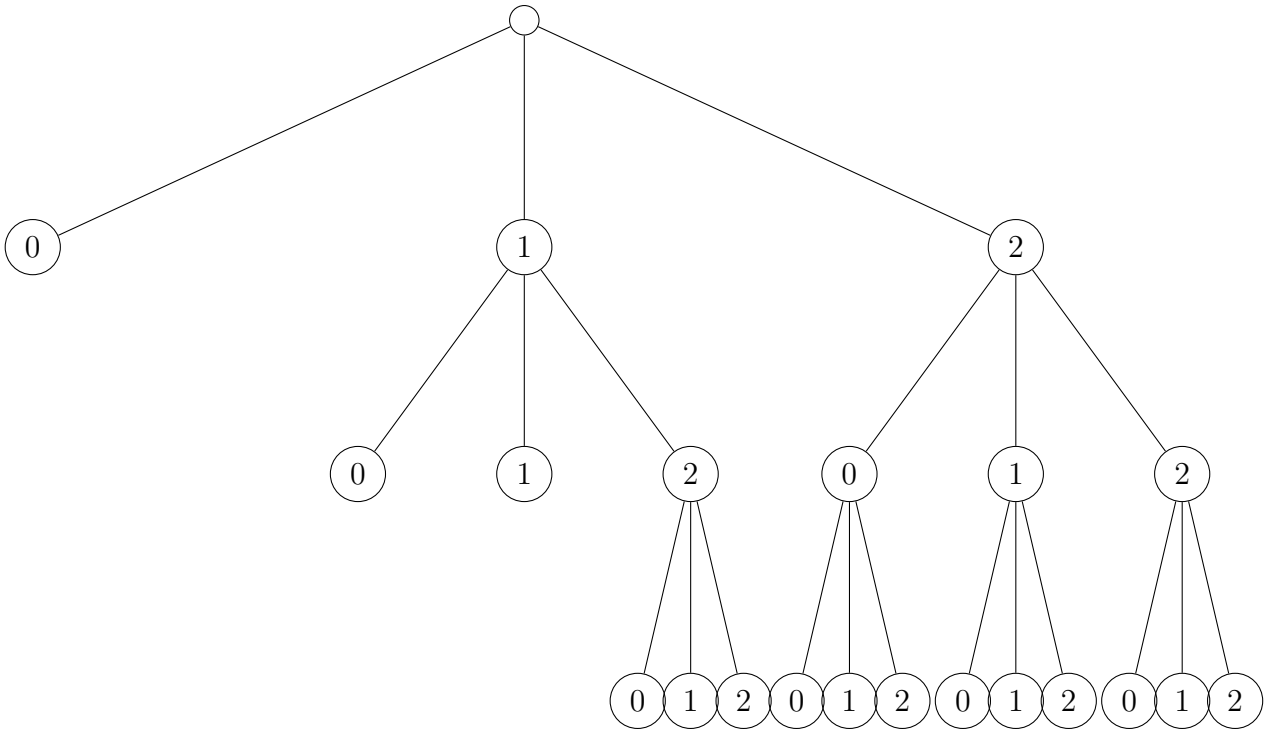

Figura 4.10: Árvore de contextos estimada com profundidade $d = 3$ Período de Estiagem

Capítulo 5

Considerações Finais

Neste trabalho estudamos Cadeias de Ordem Variável Estocasticamente Perturbadas, tomando valores em um alfabeto $\mathcal{A} = 0, 1, 2$.Constatamos a robustez do estimador utilizado neste estudo em recuperar corretamente a árvore de contextos para níveis de contaminação fixados. Propusemos modelos pluviométricos para prever se o próximo dia seria sem chuva, com chuva moderada ou com chuva forte.

Verificamos que regime de chuva para o período que antecede o ano 2000 é o mesmo para o período pós 2000, entretanto, para predizer se o próximo dia irá chover ou não, é necessário observamos mais dias no passado quando observamos os processos para o período posterior ao ano 2000. Esse aumento de "incerteza", talvez possa ser explicado pelo crescimento na emissão, ou acumulo, de monóxido de carbono no DF após o ano de 2000. Observamos que o processo de chuvas no DF para os períodos de estiagem é diferente para os períodos de chuvas, caracterizando as estações típicas do Distrito Federal.

Utilizamos o modelo de Contaminação Zero Inflado, definidos em Garcia e Moreira (2015), e o modelo de Contaminação por Congruência, definido em Collet, Galves e Leonardi (2008). Em cada um destes modelos de Contaminação os respectivos autores mostraram que é possível recuperar a árvore de contextos do processo original utilizando uma amostra contaminada do processo.

Para a comparação entre os modelos de Contaminação utilizamos a versão do Algoritmo Contexto apresentada em Galves e Leonardi (2008). Foram realizadas simulações de amostras não perturbadas e perturbadas segundo cada um destes modelos de contaminação e destacamos o bom desempenho do estimador obtido através das proporções de retornos corretos das árvores de contextos do processo original.

Sugerimos para estudos futuros o estudo do estimador para o parâmetro de contaminação apresentado por Denise Duarte e Wecsley O. Prates (2016). Destacamos a redução do número de parâmetros necessários para predizer o próximo símbolo do processo em comparação com uma possível modelagem dos dados através de uma

Cadeia de Markov de ordem k, e as dificuldades na implementação de algoritmos de Cadeias de Ordem Variável que demandam muito tempo de processamento.

Referências Bibliográficas

[1] Bühlmann, P., Wyner, A. J. (1999). Variable length Markov chains, **Ann. Statist. 27**: 480-513.

[2] Csiszár, I., Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL, **IEEE Trans. Inform. Theory 52**(3): 1007-1016.

[3] Collet, P., Galves, A., Leonardi, F., Random Perturbations of Stochastic Processes with Unbounded Variable Length Memory. **Eletronic Journal of Probability**, Vol. 13, Paper nº. 48, 13451361,2008.

[4] Duarte, D., Galves, A., Garcia, N. (2006). Markov approximation and consistent estimation of unbounded probabilistic suffix trees, **Bull, Braz. Math. Soc.** p. Aceito.

[5] Ferrari, F. e Wyner, A. (2003). Estimation of general stationary processes by variable length Markov chains, **Scand. J. Statist.**, 30(3): 459-480.

[6] Galves, A., Leonardi, F., Exponential inequalities for empirical unbounded context trees. Vol. 60 of **Progress in Probability**, Birkhauser, 257–270,2008.

[7] Galves, A., Locherbach, E., Stochastic chains with memory of variable length. **TICSP Series 38: 117-133**, 2008.

[8] Galves, A., Maume-Deschamps, V., Schmitt, B., Exponential inequalities for VLMC empirical trees. **ESAIM Prob. Stat.**, 2006.

[9] Garcia, Nancy. L., Moreira, Lucas. Stochastically Perturbed Chains of Variable Memory. **Journal of Statistical Physics**, v. 159, p. 1107-1126, 2015.

[10] Matta, D. H., **Algoritmos de estimação para Cadeias de Markov de Alcance Variável - aplicações a detecção do ritmo em textos escritos**.

Dissertação (Mestrado em Estatística) - Instituto de Matemática, Estatística e Computação Científica, UNICAMP. Campinas, 2008.

[11] Moreira, Lucas. **Processos de Ordem Infinita Estocasticamente Perturbados**. 2012. 54 p. Tese (Doutorado em Estatística) - Instituto de Matemática, Estatística e Computação Cientifica, Universidade Estadual de Campinas, Campinas.

[12] R Core Team (2018). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[13] Rissanen, J., A universal data compression system, **IEEE Trans. Inform. Theory** 29(5): 656-664, (1983).

[14] Quintino, S. Felipe. **Aplicações de Cadeias de Ordem Variável Estocasticamente perturbadas**. 2014. Trabalho de Conclusão de Curso. (Graduação em Estatística) - Universidade de Brasília. Orientador: Lucas Moreira.

[15] Bomfim, B.A. Alex. **Estudo de Estimadores de Árvores de Contexto Aplicados à Lingística**. 2014. Trabalho de Conclusão de Curso. (Graduação em Estatística) - Universidade de Brasília. Orientador: Lucas Moreira.

[16] Prates, O. Wecsley, Duarte, Denise. **Inferência em alguns modelos de processos estocasticamente perturbados**. 2016. 56 p. Tese (Doutorado em Estatística) - Departamento de Estatística, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais. Minas Gerais.

Apêndice A

Códigos da versão do Algoritmo Contexto

Neste apêndice apresentamos os códigos desenvolvidos por em linguagem R de computação estatística (R Core Team, 2018) da versão do Algoritmo Contexto apresentada na Definição 2.1 do Capítulo 2.

```
#####################################################
####### Uma versao do  Algoritmo Contexto #########
fgalves <- function(dados,d,delta,perturbacao = 0){
# pertubação
if (perturbacao != 0){
for (i in 1:length(dados)){
dados[i] <- dados[i]*rbinom(1,1,1-perturbacao)
}
}
# valor de |A| (considerando A = {0,1,...,|A| - 1})
alfabeto <- max(dados)+1
# função para converter (i,j) em sequência
fconverte <- function(i,j){
conversao <- character(1)
for (l in (d+1-j):1){
conversao <- paste(floor((i-1)/alfabeto^(l-1)),conversao,sep="")
i <- ((i-1) %% alfabeto^(l-1)) + 1
}
conversao
}
# função para completar matriz
fcompleta <- function(matriz,q) {
d <- (ncol(matriz)-1)
45
for (j in 1:d){
for (i in 1:(q^d)){
```

```
matriz[floor((i-1)/q)+1,j+1] <- matriz[floor((i-1)/q)+1,j+1] + matriz[i,j]
}
}
matriz
}
# contagem de N_n(s,a) e N_n(s)
num <- array(0,c(alfabeto^d,d+1,alfabeto))
for (tempo in (d+1):length(dados)){
i <- 0
ajuste <- 0
for (passado in (tempo-d):(tempo-1)){
i <- i + (alfabeto^ajuste)*(dados[passado])
ajuste <- ajuste + 1
}
num[i+1,1,(dados[tempo])+1] <- num[i+1,1,(dados[tempo])+1] + 1
}
for (i in 1:alfabeto){
num[,,i] <- fcompleta(num[,,i],alfabeto)
}
numt <- apply(num,c(1,2),sum)
# probabilidades de transição estimadas
46
tr <- array(0,c(alfabeto^d,d+1,alfabeto))
for (j in 1:(d+1)){
for (i in 1:alfabeto^(d+1-j)){
for (k in 1:alfabeto){
if (numt[i,j] == 0) tr[i,j,k] <- 1/alfabeto
else tr[i,j,k] <- num[i,j,k]/numt[i,j]
}
}
}
# matriz com os Deltas
adelta <- array(0,c(alfabeto^d,d,alfabeto))
for(a in 1:alfabeto){
for (j in 1:d){
for (i in 1:alfabeto^(d-j+1)){
adelta[i,j,a] <- abs(tr[i,j,a]-tr[floor((i-1)/alfabeto)+1,j+1,a])
}
}
}
mdelta <- apply(adelta,c(1,2),max)
# achando a matriz de zeros e uns
matriz <- matrix(0,alfabeto^d,d+1)
for (j in 1:d){
47
for (i in 1:alfabeto^(d-j+1)){
```

```
if (matriz[i,j] == 1){
matriz[floor((i-1)/alfabeto)+1,j+1] <- 1
}
else if (matriz[floor((i-1)/alfabeto)+1,j+1] == 0){
matriz[floor((i-1)/alfabeto)+1,j+1] <- as.integer(mdelta[i,j] > delta)
}
}
}
# achando a árvore
arvore <- character()
arvore[1] <- "sequencia vazia"
index <- 1
valor <- 0
for (i in 1:alfabeto^d){
for (j in 1:d){
valor <- 0
if (matriz[i,j] == 0 && matriz[floor((i-1)/alfabeto)+1,j+1] == 1){
valor <- 1
}
if (valor == 1){
arvore[index] <- fconverte(i,j)
index <- index + 1
}
}
}
arvore
}
```