UNIVERSIDADE FEDERAL DO PAMPA

JEAN SAMARONE ALMEIDA FERREIRA

PREDIÇÃO DA VARIABILIDADE ESPACIAL DA PRODUTIVIDADE AGRÍCOLA COM MODELOS OCULTOS DE MARKOV

Bagé
2019

JEAN SAMARONE ALMEIDA FERREIRA

PREDIÇÃO DA VARIABILIDADE ESPACIAL DA PRODUTIVIDADE AGRÍCOLA COM MODELOS OCULTOS DE MARKOV

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada como requisito parcial para a obtenção do título de Mestre em Computação Aplicada.

Orientadora: Ana Paula Lüdtke Ferreira
Coorientador: Naylor Bastiani Perez

Bagé
2019

Ficha catalográfica elaborada automaticamente com os dados fornecidos pelo(a) autor(a) através do Módulo de Biblioteca do Sistema GURI (Gestão Unificada de Recursos Institucionais).

F383P Samarone Almeida Ferreira, Jean

Predição da variabilidade espacial da produtividade agrícola com modelos ocultos de Markov / Jean Samarone Almeida Ferreira.

90 p.

Dissertação (Mestrado) – Universidade Federal do Pampa, PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA, 2019.

"Orientação: Ana Paula Lüdtke Ferreira; Coorientação: Naylor Bastiani Perez".

1. Modelo oculto de Markov. 2. Produtividade agrícola. 3. Variabilidade espacial. 4. Inferência probabilística. I. Título.

JEAN SAMARONE ALMEIDA FERREIRA

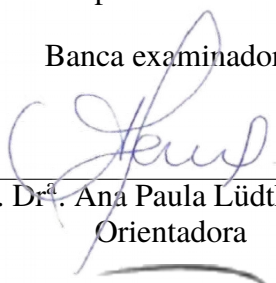
PREDIÇÃO DA VARIABILIDADE ESPACIAL DA PRODUTIVIDADE AGRÍCOLA COM MODELOS OCULTOS DE MARKOV

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada como requisito parcial para a obtenção do título de Mestre em Computação Aplicada.

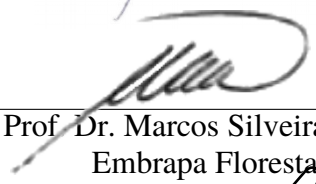
Área de concentração: Tecnologias para a Produção Agropecuária

Dissertação defendida e aprovada em: 05 de dezembro de 2019.

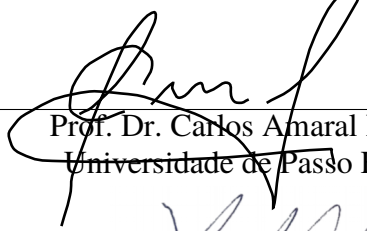
Banca examinadora:



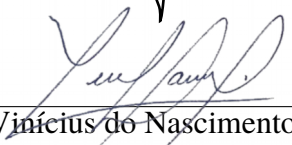
Profª. Drª. Ana Paula Lüdtke Ferreira
Orientadora



Prof. Dr. Marcos Silveira Wrege
Embrapa Florestas



Prof. Dr. Carlos Amaral Hölbig
Universidade de Passo Fundo



Prof. Dr. Vinícius do Nascimento Lampert
Embrapa Pecuária Sul

AGRADECIMENTO

À professora Ana Paula Lüdtke Ferreira pela orientação, ensinamentos e todo o tempo dedicado a este trabalho.

Ao professor Naylor Bastiani Perez pela coorientação e disponibilização dos dados necessários para o desenvolvimento do trabalho.

À banca examinadora pela predisposição e o tempo dedicado na avaliação deste trabalho.

À minha família que entendeu minha ausência e me apoiou durante o desenvolvimento do trabalho.

RESUMO

O trabalho desenvolvido nesta dissertação de mestrado caracteriza-se como uma pesquisa exploratória, que utiliza um estudo de caso com base em dados coletados em uma das áreas de produção da Embrapa Pecuária Sul e revisão de literatura relacionada ao problema. O trabalho é justificado pela necessidade de tentar entender e prever a produtividade de uma determinada área ao longo do tempo. O objetivo é prever o que pode acontecer em uma colheita, usando um modelo de Markov oculto para inferências probabilísticas em dados históricos. Os dados foram organizados em sequências de estados, onde cada estado representa um resultado de produtividade (a parte oculta do modelo) e dados referentes às condições coletadas de dados meteorológicos, do solo, do balanço hídrico e de outros dados (a parte visível do modelo). A implementação do modelo foi feita com a linguagem R . Foi feita uma comparação entre os modelos com dados reais e simulados. Os resultados apontam a necessidade de um conjunto maior de dados de produtividade para que o modelo seja confiável. O modelo mostrou-se adequado para predizer a produtividade ao longo das safras, mas a estimativa da variabilidade dentro de uma determinada área é mais sensível à disponibilidade e discretização dos dados de entrada.

Palavras-chave: Modelo oculto de Markov. Produtividade agrícola. Variabilidade espacial. Inferência probabilística.

ABSTRACT

The work developed in this Master's Thesis is characterized as exploratory research using a case study based on data collected from one of Embrapa Pecuária Sul production areas, and problem-related literature review. The work is justified by the need to try to understand and predict land productivity over different times and seasons. The goal is to predict what might happen in a crop, using a hidden Markov model for probabilistic inference on historical data. The data were organized in state sequences, where each state represents a productivity result (the model hidden part) or data regarding conditions gathered from meteorological, soil, water balance, and other data (the model visible part). Model implementation was done using R software libraries. A comparison was made between models with real and simulated data. The results point to the need for a larger set of productivity data so that the model results are reliable. The model was adequate to predict yield throughout the crop, but the estimation of variability within a given area is more sensitive to input data availability and discretization.

Keywords: Hidden Markov model, Agricultural productivity, Spatial variability, Probabilistic inference.

LISTA DE FIGURAS

Figura 1 Brasil, estado do Rio Grande do Sul, município de Hulha Negra e localização do talhão objeto do estudo 14
Figura 2 Datas de plantio e colheita das safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018 15
Figura 3 Etapas da metodologia 35
Figura 4 Grade de 50 pontos amostrais distribuídos no talhão de estudo 36
Figura 5 Produtividade da soja em t/ha nas safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018 37
Figura 6 Modelo Oculto de Markov 42
Figura 7 Sequências de dados originais, com 50 registros por variável, para as safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018 44
Figura 8 Sequência de dados simulados, com 50 registros por variável, para as safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018 45
Figura 9 Gráfico dirigido da matriz de transição 49
Figura 10 Gráfico dirigido do modelo oculto de Markov com probabilidades de emissão combinadas, probabilidades de transição e probabilidades iniciais 51
Figura 11 Dados simulados, nove variáveis 54
Figura 12 Dados simulados, cinco variáveis, dados de solo 55
Figura 13 Caminho mais provável com dados simulados, seis variáveis e dados meteorológicos 55
Figura 14 Cinco variáveis, dados de solo 56
Figura 15 Nove variáveis e seis variáveis, dados meteorológicos 56
Figura 16 Mapa de Pontos da Produtividade Simulada da Soja 57
Figura 17 Mapa Interpolado da Produtividade Simulada da Soja 58

Figura 18 Deficit e excedente hídrico na safra 2012-2013 89
Figura 19 Deficit e excedente hídrico na safra 2014-2015 89
Figura 20 Deficit e excedente deficit hídrico na safra 2016-2017 90
Figura 21 Deficit e excedente hídrico na safra 2017-2018 90

LISTA DE TABELAS

Tabela 1 Dados relacionados à produtividade nas safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018 15
Tabela 2 Variáveis relacionadas à produtividade em cada safra 15
Tabela 3 Níveis críticos de compactação em função da classificação dos solos 32
Tabela 4 Variáveis diárias relacionadas ao clima 38
Tabela 5 Datas de plantio, emergência, colheita e número de dias das safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018 39
Tabela 6 Variáveis relacionadas ao solo 40
Tabela 7 Frequências de classes das variáveis 41

LISTA DE ABREVIATURAS E SIGLAS

CONAB	Companhia Nacional de Abastecimento
UNIPAMPA	Universidade Federal do Pampa
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
SIG	Sistemas de Informações Geográficas
BDMEP	Banco de Dados Meteorológicos para Ensino e Pesquisa
OMM	Organização Meteorológica Mundial
INMET	Instituto Nacional de Meteorologia
HMM	Hidden Markov Model
IDW	Inverse Distance Weighting
SGDB	Sistemas de Gerenciamento de banco de Dados
OSGEO	Open Source Geospatial Foundation
GPS	Global Positioning System
ha	hectares
mm	milímetros
ton	toneladas

SUMÁRIO

1 INTRODUÇÃO 11
1.1 Contexto e justificativa 11
1.2 A área de estudo 13
1.3 Objetivos 16
1.4 Estrutura do trabalho 16
2 REVISÃO BIBLIOGRÁFICA 18
2.1 Variabilidade espacial da produtividade 18
2.2 Modelos e métodos para análise da variabilidade espacial 20
2.3 Modelos de Markov 23
2.3.1 Definições e aplicações 23
2.3.2 Algoritmo Viterbi 27
2.3.3 Ferramentas para construção de modelos de Markov 28
2.4 Clima 30
2.5 Compactação do solo 31
2.6 Sistemas de informações geográficas 33
3 MATERIAL E MÉTODOS 35
3.1 Caracterização e etapas do método 35
3.2 Organização e pré-processamento dos dados 35
3.2.1 Dados de produtividade 36
3.2.2 Dados meteorológicos 38
3.2.3 Dados de solo 39
3.3 Definição de classes 40
3.4 Modelo oculto de Markov para análise da variabilidade espacial da produtividade 41
3.5 Simulação e comparação de modelos 43
3.6 Espacialização dos resultados do modelo 46
4 RESULTADOS 48
4.1 O modelo oculto de Markov 48
4.2 O caminho mais provável 52
4.3 Comparação de modelos 54
4.4 Espacialização dos dados 57
5 CONCLUSÃO 59
5.1 Discussão dos resultados obtidos 59
5.2 Trabalhos Futuros 60
REFERÊNCIAS 62
APÊNDICE A – CÓDIGO R UTILIZADO PARA O MODELO OCULTO DE MARKOV 68
APÊNDICE B – CÓDIGO R UTILIZADO NA SIMULAÇÃO 78
APÊNDICE C – CÓDIGO R UTILIZADO PARA A GERAÇÃO DO MAPAS 83
APÊNDICE D – FUNÇÕES AUXILIARES 86
APÊNDICE E – GRÁFICOS DE DEFICIT E EXCEDENTE HíDRICO 89

1 INTRODUÇÃO

1.1 Contexto e justificativa

A agricultura é uma atividade humana que tem como objetivo prover o sustento alimentar da população mundial e consiste na principal atividade econômica de muitos países. Ao longo do tempo os processos de produção agrícola sofrem transformações, principalmente em relação à busca de técnicas de produção mais eficazes, que aumentem a produtividade ao mesmo tempo que façam o melhor uso possível dos recursos necessários para o processo de produção. A busca da melhoria da produtividade agrícola transforma as formas de produção, reduz desperdício de recursos como combustível e insumos, melhora a qualidade de sementes, implementos e formas de manejo (solo, pragas e hídrico). A produção de grãos aparece com destaque nesse cenário, pois é utilizada tanto para a alimentação humana quanto para a fabricação de ração animal, produtos não alimentícios e ainda uma alternativa para utilização na fabricação de biocombustíveis (COMPANHIA NACIONAL DE ABASTECIMENTO, 2017, p.7), (SIQUEIRA, 2004, p.131).

Segundo a Companhia Nacional de Abastecimento (COMPANHIA NACIONAL DE ABASTECIMENTO, 2018), na safra 2017-2018 a produção de grãos no Brasil foi estimada em 228,6 milhões de toneladas, uma redução de 3,8% em relação à safra anterior. Nessa safra, um dos fatores para a queda na produtividade foi o estresse hídrico ocorrido em função da menor taxa de precipitação, que impactou a produção do milho de segunda safra. A produção do milho de primeira safra (verão) teve boa produtividade, mas de forma geral, a produção desse grão diminuiu em função das condições de mercado pouco favoráveis. A produção de soja, por sua vez, teve um recorde de 119 milhões de toneladas, 4,3% maior do que na safra passada.

A cultura da soja constitui uma *commodity* no mercado internacional de grãos e serve de matéria-prima para vários produtos. É hoje a cultura mais importante do agronegócio mundial e em 2018 movimentou cerca de 31,7 bilhões de dólares, com os Estados Unidos como maior produtor, seguido por Brasil, Argentina, China e Índia (MINISTÉRIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO, 2019). Conforme o Departamento de Agricultura dos Estados Unidos (UNITED STATES DEPARTAMENT OF AGRICULTURE, 2019), as mudanças nas práticas de manejo e expansão da área colhida permitiram que o Brasil se tornasse um dos principais exportadores de soja, impulsionado pelo aumento da demanda global, altos preços e

avanços tecnológicos que ocorreram nas duas últimas décadas.

Ainda na safra 2017-2018, conforme o relatório da EMATER/RS (EMPRESA DE ASSISTÊNCIA TÉCNICA E EXTENSÃO RURAL, 2019), no estado do Rio Grande do Sul a área plantada de soja foi de 5.758.133 ha, a produção 17.546.405 toneladas e a produtividade média 3.047 kg/ha. Na região de Bagé a produtividade média foi de 2.001 kg/ha, a menor produtividade entre todas as regiões avaliadas. Conforme a CONAB (COMPANHIA NACIONAL DE ABASTECIMENTO, 2017), o estado apresenta um dos menores rendimentos nacionais, principalmente em função de períodos de estiagem nas fases vitais de desenvolvimento da planta, o que evidencia a importância dos fatores climáticos.

A CONAB (COMPANHIA NACIONAL DE ABASTECIMENTO, 2017), define a *produtividade média* ou *rendimento médio* como "o quociente obtido pela divisão da produção agrícola pela área plantada, ou seja, a produtividade média é a quantidade de produto auferido em razão do mais fundamental insumo da produção agrícola, a terra". Os autores destacam que o aumento da produtividade é importante pois, além do aumento na renda do produtor, propicia uma menor demanda por novas frentes de cultivo e diminui o impacto ambiental. Segundo Silva (2019), "a produtividade na cultura é resultado da interação entre o potencial genético da cultivar[1] e as condições ambientais durante o período de cultivo". O autor destaca que o manejo adequado deve levar em conta a escolha de cultivares recomendadas para a região, sementes de boa qualidade, a recomendação técnica de adubação, plantio de qualidade e na época recomendada e o manejo adequado de plantas daninhas, pragas e doenças. A produtividade representa o desempenho econômico de determinada cultura, é um importante indicador agrícola e o estudo de sua variação ao longo do espaço e do tempo pode ajudar a resolver problemas em áreas que apresentem alta variabilidade em sua produção.

O presente trabalho encaixa-se no contexto atual de crescimento e popularização da utilização de técnicas de agricultura de precisão. O detalhamento da informação sobre a produção e sobre as características de uma área permite que novas relações possam ser descobertas e estabelecidas. Esse fato possibilita que novas formas de manejo resultem em um melhor aproveitamento dos recursos (naturais, financeiros e humanos) e minimizam problemas ou erros que antes eram imperceptíveis. Além de preços mais acessíveis, a coleta e a pré-análise de dados estão cada vez mais rápidas, automatizadas e com maior precisão. Os dados de colheita, que já ofereceram uma informação primária

[1] Lei Nº 9.456, de 25 de abril de 1997.

e direta para o produtor, podem também servir como entrada de dados para sistemas de predição mais complexos, como o modelo proposto neste trabalho.

O problema abordado nesse trabalho é enfrentado safra após safra, por agricultores, agrônomos e agências governamentais: como garantir uma produtividade com menor variabilidade possível em uma determinada área, ao longo do tempo, com os recursos agronômicos (correções de solo, adubação e tratos culturais) disponíveis e lidando com uma alta parcela de incerteza (principalmente o clima). A agricultura é uma atividade complexa, pois depende de uma série da fatores para que se atinja o objetivo final que é uma boa safra e o retorno financeiro ao produtor. Existe uma quantidade significativa de variáveis e um alto grau de incerteza que pode influenciar a variabilidade da produtividade, mas cabe destacar a importância do subsistema ecológico, assim nomeado por Diniz (1984, p.110), que compreende os solos, as condições climáticas e o relevo. Conforme o autor, o solo sustenta fisicamente as plantas e oferece as condições de alimentação necessárias ao seu crescimento. O clima exerce grande influência sobre a agricultura, pois as plantas exigem certa quantidade de calor, água e insolação. O relevo é importante pois consiste em um moderador das condições climáticas e de uso da terra.

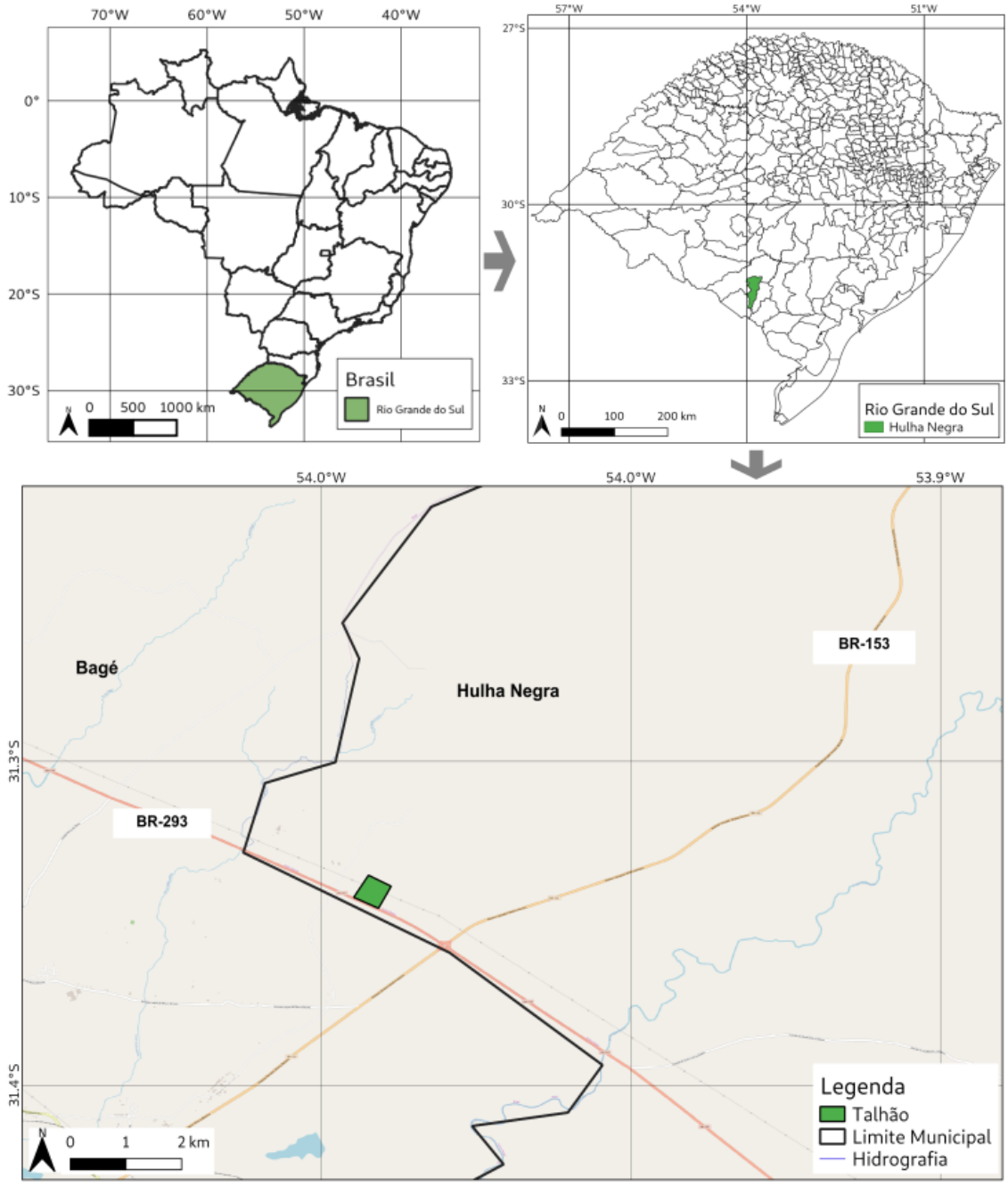
Esse trabalho justifica-se pela necessidade de tentar entender o comportamento da produtividade dentro de uma determinada área e ao longo do tempo, prevendo o que pode acontecer em uma safra, por meio de um processo de inferência probabilística sobre os dados históricos de produção.

1.2 A área de estudo

A área de estudo está localizada no interior de uma área experimental da EMBRAPA Pecuária Sul, no município de Hulha Negra, próximo ao limite com o município de Bagé, região da Campanha do Rio Grande do Sul, mesorregião Sudoeste Riograndense, conforme mostra a Figura 1.

A área plantada tem cerca de 13,4 ha, com drenagem moderada e altitudes que variam entre 230 e 250 metros. É acompanhada por equipe técnica especializada, recebe um manejo adequado e é monitorada por meio de coletas de solo e dados de resistência à penetração. Mesmo assim, algumas áreas não apresentam resultados satisfatórios, com diferenças na produtividade dentro da mesma área e também ao longo das safras. Esse comportamento, além de prejudicar financeiramente os agricultores, não permite um planejamento adequado ao longo dos anos.

Figura 1 – Brasil, estado do Rio Grande do Sul, município de Hulha Negra e localização do talhão objeto do estudo

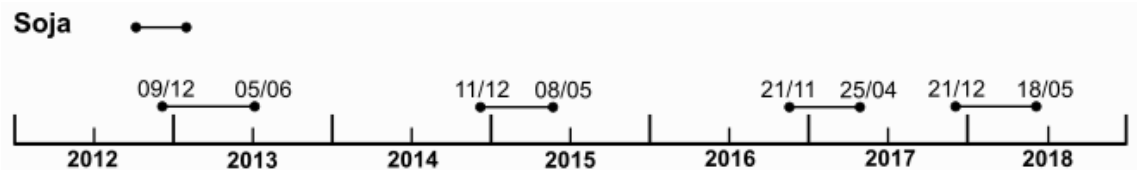


Fonte: Autor (2019)

Nos últimos sete anos é utilizada em regime de rotação com plantio de soja e pastagem (sorgo e azevém), conforme apresentado na linha do tempo representada na Figura 2.

Ao longo das safras a área tem apresentado alta variabilidade na produtividade,

Figura 2 – Datas de plantio e colheita das safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018



Fonte: Autor (2019)

conforme pode ser verificado na Tabela 1. As colunas denominadas AreaAp[2], AreaFeita, PesoSeco, Peso Umido, TxMin e TxMax têm sua descrição apresentada na Tabela 2.

Tabela 1 Dados relacionados à produtividade nas safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018

Safra	AreaAp	AreaFeita	PesoSeco	PesoUmido	TxMin	TxMed	TxMax
2012/2013	12,48	10,78	25362,62	27401,47	0,81	2,03	2,76
2014/2015	14,10	13,02	30098,38	30111,2	0,36	2,37	4,22
2016/2017	13,73	12,59	47870,21	48705,13	0,99	3,74	5,68
2017/2018	13,54	12,53	12997,77	13842,24	0,20	1,04	1,86

Fonte: Autor (2019)

Tabela 2 Variáveis relacionadas à produtividade em cada safra

Nome	Unidade	Descrição
AreaAp	ha	área aplicada com transpasse
AreaFeita	ha	área aplicada sem transpasse
PesoSeco	Kg	peso colhido descontado a umidade
PesoUmido	Kg	peso colhido sem descontar a umidade
TxMin	t/ha	taxa mínima colhida por hectare
TxMed	t/ha	taxa média colhida por hectare
TxMax	t/ha	taxa máxima colhida por hectare
Altitude	m	altitude em metros

Fonte: Autor (2019)

[2]O transpasse refere-se à passagem da colheitadeira mais de uma vez pela mesma área.

1.3 Objetivos

Este trabalho tem como objetivo desenvolver um modelo oculto de Markov que permita o estudo da variabilidade da produtividade agrícola, dentro de uma determinada área e também ao longo do tempo.

São objetivos específicos deste trabalho:

- Criação de série temporal com dados meteorológicos e balanço hídrico;
- Organização do conjunto de dados em sequência de estados;
- Modelagem do problema como um modelo oculto de Markov;
- Implementação do modelo através de bibliotecas de software;
- Aplicação do modelo aos dados reais e simulados;
- Comparação dos modelos;
- Interpretação dos resultados.

1.4 Estrutura do trabalho

O trabalho está organizado em cinco capítulos, divididos conforme explicado a seguir. O Capítulo 1 traz a introdução, e destaca o papel da produção agrícola no Brasil e no mundo. Trata do cenário da produção de grãos, especialmente da soja, no estado do Rio Grande do Sul, na região do município de Bagé, e do conceito de produtividade. Por fim, destaca a importância, o contexto do trabalho, a sua justificativa e seus objetivos.

No Capítulo 2 é discutida a variabilidade espacial da produtividade, os modelos e métodos mais utilizados para tratar a questão. É introduzido o conceito de modelo de Markov e destacadas as ferramentas de software disponíveis para sua implementação. São abordados, ainda, o clima e os conceitos básicos de resistência à penetração e compactação do solo. Por fim, é feita um revisão sobre Sistemas de Informações Geográficas (SIG) e como eles servem de base para a exploração de dados georreferenciados.

No Capítulo 3 é feita a caracterização da área de estudo, apresentado o histórico das safras e o resumo dos dados de produtividade, meteorológicos e de resistência à penetração. É feita uma explicação sobre a preparação de cada uma das variáveis para inserção no modelo construído. Em seguida trata do modelo oculto de Markov proposto. O capítulo trata ainda da definição das classes de variáveis da espacialização

dos resultados do modelo.

No Capítulo 4 é mostrada a aplicação do modelo aos dados simulados. É feita uma comparação entre modelos com número de parâmetros diferentes. Por fim, é apresentado o resultado da simulação em forma espacializada.

O Capítulo 5 apresenta a conclusão do trabalho, resultados, objetivos atingidos e indica trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

2.1 Variabilidade espacial da produtividade

Conforme Franchini *et al.* (2016, p. 18), “a produtividade de grãos de soja é dependente das características genéticas das plantas, do ambiente de produção e da interação entre esses fatores". Destacam que o solo e o clima influenciam de forma relevante a produtividade da cultura em cada região. Os autores realizaram um estudo no estado do Paraná, com o intuito de avaliar a variabilidade espacial e temporal do desempenho agronômico da soja em relação aos fatores ambientais. Fizeram uso de semivariogramas para a estimação de dados faltantes e análise de agrupamento para identificar grupos de municípios que apresentavam características semelhantes. Os resultados apontaram que existe uma alta variabilidade na produtividade de grãos e que o fator ambiental mais fortemente associado à produtividade da soja é a temperatura do ar. Houve aumento tanto da área cultivada como da produtividade em regiões com temperaturas mais amenas e estabilização da produtividade em áreas com clima mais quente, solo arenoso e com baixa capacidade de armazenamento de água.

Segundo Guedes-Filho (2009, p. 1), “para melhor entender os fatores que afetam a produtividade das culturas, um novo componente passou a ser considerado no manejo da produção agrícola: a variabilidade espacial”. O autor destaca a importância de entender a variabilidade espacial dos atributos do solo a fim de minimizar os efeitos dessa variabilidade na produtividade das culturas, servindo como parâmetro de tomada de decisão e aplicação de um manejo mais adequado. Em seu estudo avaliou as produtividades de soja, milho, aveia, centeio, triticale, labelabe e mamona, analisando atributos químicos e físico-hídricos do solo com ferramentas de geoestatística, estudo de semivariogramas e interpolação por *kriging*. Um maior detalhamento sobre o uso dessas técnicas pode ser encontrado em (REIMANN *et al.*, 2008, p.103) e (NETELER; MITASOVA, 2008, p.348). As produtividades apresentaram alta variabilidade, sendo a mais alta, em anos com deficit hídrico ou distribuição irregular de água. Houve relação espacial entre as produtividades das culturas e os atributos físicos e químicos do solo, indicando três zonas de manejo sobre a área de estudo.

Kravchenko e Bullock (2000) destacam que a análise da variabilidade na produção é uma questão importante na pesquisa agrícola, estando as características topográficas entre os fatores mais importantes que afetam a produtividade. Em seu estudo avaliaram

as relações produtividade-topografia-solo utilizando dados de produtividade de milho e soja entre os anos de 1994 e 1997, matéria orgânica do solo, capacidade de troca de cátions e concentrações de fósforo (P) e potássio (K) em oito áreas. Os dados topográficos utilizados foram elevação, declividade, curvatura e acumulação do fluxo de água. Os resultados comprovaram que os dados topográficos em combinação com informações do solo são úteis para explicar a variabilidade de rendimento.

Vieira e Gonzalez (2003) avaliaram a variabilidade espacial de propriedades do solo e do rendimento de culturas sob plantio direto, em função do tempo, em duas condições de solo e clima no Estado de São Paulo. Em duas áreas diferentes, de aproximadamente 1 (um) hectare cada, foram cultivados milho, soja, algodão, aveia, aveia preta, trigo, arroz, centeio e adubos verdes, sendo avaliados fertilidade, propriedades físicas do solo e rendimento. A amostragem foi feita a cada dois anos e os dados foram analisados usando semivariogramas e geração de mapas interpolados por *kriging*. Os resultados apontaram que os fatores que afetam a variabilidade se alteram de uma cultura para outra e as variações no rendimento, de um ano para outro, sugerem que as próprias causas também mudam com o tempo. Os resultados dos semivariogramas cruzados entre o teor de fósforo nas folhas e o rendimento da soja reforçam a conclusão do trabalho.

Usowicz e Lipiec (2017) investigaram a variação das propriedades químicas e físicas do solo e a produtividade de grãos de aveia, centeio, aveia e triticale nos anos de 2001, 2002, 2003 e 2015. Os dados analisados incluíram os componentes do solo: areia, silte, argila, carbono orgânico, capacidade de troca de cátions, pH na camada superficial e subsolo, quantidade de água e densidade aparente. A produção de aveia foi avaliada em 2001 e a de centeio e triticale em 2002 e 2015. Os resultados foram analisados utilizando geoestatística por meio de semivariogramas e mapas gerados com o método inverso da potência das distâncias (*inverse distance weighting* ou IDW). Os resultados do estudo apontaram que a produtividade de grãos foi correlacionada significativamente e positivamente com a capacidade de troca de cátions na camada superficial e subsolo, quantidade de água no solo, argila e carbono orgânico. Os autores concluem que a geoestatística é uma ferramenta útil para determinar as inter-relações espaciais do rendimento das culturas e das propriedades do solo.

2.2 Modelos e métodos para análise da variabilidade espacial

Diferentes técnicas e modelos têm sido utilizados para estudar os fatores que afetam a produtividade buscando entender a variabilidade espacial na agricultura. A estatística descritiva, que faz uso de medidas como média, mediana, variância, coeficiente de variação, assimetria e curtose, é utilizada com o objetivo de verificar a existência de tendência central e dispersão dos dados. Tem função primordial de descrever e ajudar compreender os dados de uma determinada distribuição. Diante de um grande conjunto de dados costuma-se agrupá-los, muitas vezes, reduzindo seu volume de forma a poder relacioná-los e compará-los, facilitando a sua interpretação. Esse processo de redução, geralmente, leva à perda de informação, que pode impactar no resultado do modelo ou experimento, vide King (1998, p.12), Reimann *et al.* (2008, p.53) e Gerardi e Silva (1981, p.43).

Conforme Cai *et al.* (2013), a relação entre o clima e o rendimento das culturas tem sido estudada a partir de dois tipos de modelos: modelos de crescimento de culturas (*crop growth models*) e modelos de regressão (*regression models*). Os modelos de crescimento de culturas são simulações feitas por meio do uso de métodos computacionais baseados na integração entre os processos biológicos, físicos e químicos envolvidos no crescimento de plantas. Esses modelos incorporam informações sobre clima, datas de plantio e colheita, aplicação de insumos, irrigação e propriedades do solo, para simular a produtividade das culturas, o que implica em um conjunto mais completo e detalhado dos dados. Modelos de crescimento são comumente usados para simulação quando os dados podem ser gerados dentro de distribuições estatísticas conhecidas ou pertinentes. Os modelos de regressão demandam menor quantidade de dados, mas necessitam que um conjunto apropriado de fatores seja previamente determinado, visto que a quantidade excessiva de variáveis pode resultar em estimativas não adequadas, principalmente se o tamanho da amostra for pequeno.

Análises de correlação e regressão são comumente utilizadas para estabelecer a existência de dependências ou associações lineares entre as variáveis estudadas e podem ser utilizadas para estudar a variabilidade na produtividade. Conforme Berquo, Souza e Gotlieb (1981, p.98), em pesquisas que envolvem duas ou mais variáveis deve-se estudá-las simultaneamente, procurando uma possível correlação ou outro tipo de relacionamento entre elas. Resumindo, busca-se determinar como as alterações sofridas em uma variável afetam as demais variáveis do problema.

A correlação fornece o grau de relacionamento linear entre duas variáveis: no caso das variáveis X e Y apresentarem variações lineares no mesmo sentido a correlação é positiva, quando as variações são em sentidos contrários a correlação é negativa e quando a variação de X não está linearmente relacionada a Y, não existe correlação. Segundo Gerardi e Silva (1981), na correlação examina-se, particularmente, até que grau duas variáveis são interdependentes ou covariantes, isto é, variam juntas. Kravchenko e Bullock (2000) utilizaram o coeficiente de correlação de Pearson para correlacionar o rendimento de grãos de milho e soja com a topografia e propriedades do solo.

Conforme Diniz (1984, p.168), "os coeficientes de correlação medem apenas a relação entre duas variáveis". Além disso, se essa relação existe, ela só pode ser medida por um coeficiente de correlação se a relação for linear. Murphy (2012, p.45) apresenta diversos exemplos de variáveis fortemente relacionadas que não apresentam correlação, pois o relacionamento existente não é linear. No caso de estimar o valor de uma variável, quando a outra assumir uma determinada posição, caracterizando um problema de predição, este pode ser resolvido pelas técnicas de regressão. A análise de regressão descreve, por meio de uma equação, o comportamento de uma das variáveis em função do comportamento de uma (linear simples) ou mais variáveis (regressão múltipla). Conforme Gerardi e Silva (1981), em uma análise de regressão, pretende-se saber se é possível, partindo de uma variável, predizer a outra, ou seja, predizer que o valor de uma variável Y corresponde ao valor dado em uma variável X. Geralmente X é a variável independente e Y a dependente. As técnicas de análise de regressão e semivariogramas foram utilizadas por (MILLER; SINGER; NIELSEN, 1998) para determinar a relação entre a variabilidade espacial do rendimento de trigo e propriedades do solo. Konopatzki *et al.* (2012), analisaram a variabilidade espacial da produtividade de peras associada a outros parâmetros, utilizando correlação de Pearson e Spearman, semivariogramas e regressão múltipla.

Um semivariograma ou variograma, em geoestatística, é utilizado para expressar a variabilidade espacial numa direção pré-definida. Conforme Grego, Oliveira e Vieira (2014), a hipótese básica sobre a qual a geoestatística se baseia é que os dados vizinhos são mais parecidos que os dados distantes. Nesse sentido, o semivariograma mede o grau de semelhança entre os vizinhos. Semivariogramas foram utilizados por (MILLER; SINGER; NIELSEN, 1998), (USOWICZ; LIPIEC, 2017), (VIEIRA; GONZALEZ, 2003), (GUEDES-FILHO, 2009), (KONOPATZKI *et al.*, 2012), (FRANCHINI *et al.*, 2016), (AL-OMRAN *et al.*, 2013), (MATTIONI; SCHUCH; VILLELA, 2011) e

(ACOSTA *et al.*, 2019).

Devido à quantidade de variáveis, alguns modelos utilizam técnicas estatísticas para seleção e redução da dimensionalidade. A análise de componentes principais foi utilizada por (SANTI *et al.*, 2012), para obter a redução na dimensionalidade de atributos químicos e físicos do solo, com o objetivo de compreender a variabilidade espacial e temporal da produtividade de culturas de grãos. A análise hierárquica de agrupamento foi utilizada por (FRANCHINI *et al.*, 2016), para avaliar a variabilidade espacial e temporal do desempenho agronômico da soja em relação aos fatores ambientais. Segundo Cai *et al.* (2013) esses métodos apresentam a desvantagem de somente levar em consideração os dados e, inadvertidamente, descartar algum aspecto agronomicamente importante.

O estudo feito por (MATIS *et al.*, 1985) propôs uma metodologia para a previsão de produtividade de culturas em intervalos de tempo durante a fase de crescimento, utilizando um modelo de Markov. Nesta abordagem, uma cadeia de Markov é construída com dados históricos a fim de fornecer distribuições de previsão para o rendimento das culturas em várias classes de condição de umidade do solo. Segundo os autores, a metodologia é comparada a uma análise de regressão em que as variáveis independentes são as várias culturas e condições de umidade do solo. Destacam que os modelos de Markov exigem suposições menos rigorosas e fornecem mais informações do que a abordagem de regressão. Uma base de dados que simula o crescimento e desenvolvimento do cultura do milho foi utilizada para demonstrar o desenvolvimento das previsões de rendimento.

A utilização de modelos de Markov para prever o rendimento do algodão, a partir de dados de pré-colheita, oriundo de uma pesquisa de rendimento de larga escala é apresentada por (MATIS; BIRKETT; BOUDREAUX, 1989). No trabalho, as matrizes de transição foram estimadas a partir de três anos (1981-1983). Os erros de previsão foram relativamente pequenos para o ano de 1984, bem como para anos anteriores, realizado através de um estudo de reamostragem. Os autores concluem que o procedimento é de fácil adaptação para aplicações similares e, portanto, a abordagem é recomendada para a previsão de culturas a partir de dados operacionais. A aplicação é diferente em muitos aspectos da aplicação anterior apresentada em (MATIS *et al.*, 1985) e demonstra a generalidade e praticidade da metodologia básica dos modelos de Markov.

Conforme Thirunavukkarasu (2015), análises de taxas de crescimento são amplamente empregadas para estudar o comportamento de diferentes produções agrícolas. As taxas de crescimento das culturas são estimadas principalmente através de

modelos paramétricos e séries temporais. O autor destaca três desvantagens relacionadas ao uso desses modelos: a) os dados podem não seguir modelos lineares ou não lineares, exigindo modelos polinomiais de grau mais elevado; b) dependem de pressupostos como normalidade e aleatoriedade de resíduos; e c) são baseados em suposições hipotéticas. Em sua tese é feita uma avaliação de tendência (aumento ou diminuição) da produção de cevada na Índia, fazendo uso de um modelo de Markov, que só depende das características do conjunto de dados passado, ao contrário do pressupostos citados acima.

2.3 Modelos de Markov

Os estudos comumente realizados focam somente na relação entre as variáveis em um determinado instante e deixam de lado a evolução do processo ao longo do tempo. Os modelos ocultos de Markov estabelecem um relacionamento probabilístico entre os estados ocultos e as observações feitas ao longo do processo, que evolui no tempo por meio das transições entre seus estados. A escolha desta abordagem para a busca da solução do problema se deu em função de saber como o ambiente afeta a produção, o que uma estruturação em estados não é capaz de responder. A pergunta "qual é o rendimento de colheita mais provável desta plantação, tendo como base os dados atuais?" determina um problema de inferência de um conjunto de condições para um resultado final provável. Os modelos ocultos de Markov separam as variáveis observáveis (fatores do ambiente de produção) das variáveis ocultas (resultados da produção) que não podem ser observadas diretamente.

2.3.1 Definições e aplicações

Apesar da pouca literatura encontrada sobre o uso de modelos de Markov em problemas relacionados à produtividade agrícola, Paegelow e Olmedo (2008) afirmam que as cadeias de Markov é uma das abordagens mais utilizadas na modelagem ecológica e ambiental. Diniz (1984, p.250), destacou a potencialidade de aplicação dos princípios markovianos e dos recursos de tecnologia da informação aos estudos de transição e transformação da agricultura. As cadeias de Markov são nomeadas em homenagem ao matemático russo Andrei Andreyevich Markov (1856-1922), que iniciou a teoria dos processos estocásticos.

Um processo estocástico é um modelo matemático que representa processos que evoluem ao longo do tempo de maneira probabilística (TAHA, 2007) e onde os estados futuros dependem dos estados passados. Segundo Christofoletti (1999, p. 10), "os modelos probabilísticos ou estocásticos são expressões que envolvem variáveis, parâmetros e constantes matemáticas, juntamente com um ou mais componentes aleatórios resultantes de flutuações imprevisíveis dos dados de observação ou da experimentação". Paegelow e Olmedo (2008) afirmam que a ideia básica de um modelo estocástico é que uma observação reflete um entre os possíveis estados do sistema, sendo a lista de possíveis estados conhecida. As variáveis aleatórias relacionadas são ditas estocásticas e as transições de estado sucessivas formam um processo discreto. Uma variável do sistema não pode mudar continuamente no tempo, mas apenas em uma série crescente de instantes.

Uma *cadeia de Markov* (*Markov chain*) é um tipo especial de processo estocástico, no qual o estado futuro depende somente do estado presente e não de todos os estados passados. Conforme Diniz (1984, p.249), "é fundamental em uma cadeia de Markov que as probabilidades de transição não dependam de estados anteriores", ou seja, se a memória do sistema se estender por mais de um estado para trás não se tem uma cadeia de Markov, mas outro tipo de processo estocástico. A propriedade de Markov é dada pela equação 1, que evidencia o fato de que o valor da probabilidade do sistema estar no estado S_j no tempo t, dados os estados em que o sistema estava nos tempos $t-1, t-2, \ldots$ é igual à probabilidade do sistema estar no estado S_j no tempo t, dado somente o estado em que estava no tempo $t-1$.

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j | q_{t-1} = S_i) \tag{1}$$

No que se segue neste texto, as definições usadas têm como referência e são adaptadas de (JURAFSKY; MARTIN, 2018) e (TAHA, 2007). Formalmente, uma cadeia de Markov pode ser especificada como um sistema de transição de estados, apresentado na Definição 1.

Definição 1 (Cadeia de Markov) *Uma* cadeia de Markov *é uma tupla* $M = (S, P, \pi)$*, onde* S *é um conjunto de estados,* $P : S \to S$ *é a função de probabilidade de transição de estados e* $\pi : S \to [0,1]$ *é uma função de probabilidade, ou seja,* $\sum_{s \in S} \pi(s) = 1$.

Na definição de cadeia de Markov como uma tripla $M = (S, P, \pi)$, não se faz suposições sobre o conjunto de estados S; usualmente, e nos modelos trabalhados a

partir daqui, esse conjunto será sempre finito e representado como $S = \{S_1, S_2, ..., S_n\}$. A matriz de probabilidades de transição, com domínio finito, também será uma matriz finita. Particularmente, P é uma matriz quadrada, conforme apresentado na equação 2, onde cada elemento p_{ij} em P significa $P(q_t = S_j | q_{t-1} = S_i)$, ou a probabilidade do sistema estar no estado S_j no tempo t dado que estava no estado S_i no tempo $t-1$. π é uma distribuição de probabilidade inicial, onde $\pi(S_k)$, ou a probabilidade do sistema estar no estado S_k no tempo $t = 1$ será representado por π_k e $\pi_k = P(q_1 = S_k)$, para todo $k = 1, \ldots, n$.

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix} \tag{2}$$

A probabilidade condicional de um evento A, dada a ocorrência prévia de um evento B, ou seja, $P(A|B)$ pode ser calculada a partir da probabilidade conjunta $P(A,B)$, como $P(A|B) = \frac{P(A,B)}{P(B)}$, desde que o valor de $P(B)$ não seja nulo.

Segundo Taha (2007), os estados de uma cadeia de Markov podem ser classificados, baseado na probabilidade de transição p_{ij} de P:

1. um estado j é dito *absorvente* se retorna a si mesmo com certeza em uma transição $P_{ij} = 1$;
2. um estado j é *transiente* se puder alcançar outro estado, mas não puder ser alcançado de outro estado;
3. um estado j é *recorrente* se a probabilidade de ser alcançado a partir de outros estados for 1. Isso pode acontecer se, e somente se, o estado não for transiente;
4. um estado j é *periódico* com período $t > 1$, se um retorno for possível apenas em $t, 2t, 3t, \ldots$ passos. Isto significa que $p_{jj}^{(n)} = 0$ sempre que n não for divisível por t.

Uma cadeia de Markov é útil quando precisamos calcular uma probabilidade para uma sequência de eventos observáveis. Em muitos casos, no entanto, os eventos nos quais estamos interessados estão ocultos, ou seja, não os observamos diretamente. Nos problemas relacionados à descoberta de fatores causais entre variáveis, comumente os elementos observáveis são as consequências e não as causas. Na aplicação específica deste trabalho, elementos como compactação do solo, pluviosidades, insolação, etc. são observáveis e mensuráveis. A produção, contudo, somente pode ser observada no final do período. Um *modelo oculto de Markov* (ou *hidden Markov model* – HMM) permite

falar sobre os eventos observados e eventos ocultos, que são considerados como os fatores causais das variáveis observáveis nesse modelo probabilístico.

Um modelo oculto de Markov pode ser especificado conforme a Definição 2, adaptada das definições apresentadas em (FINK, 2008), (JURAFSKY; MARTIN, 2018) e (RABINER; JUANG, 1986).

Definição 2 (Modelo oculto de Markov) *Um* modelo oculto de Markov *é uma tupla* $M = (S, V, A, B, \pi)$*, onde*

- $S = \{S_1, S_2, ..., S_n\}$ *é um conjunto finito de estados,*
- $V = \{V_1, V_2, ..., V_m\}$ *é um conjunto de valores observáveis em cada instante,*
- *A é uma matriz de transição de probabilidades, em que cada elemento* $\{a_{ij}\}$ *da matriz A corresponde ao valor da probabilidade* $P(q_{t+1} = S_j | q_t = S_i)$*, ou à probabilidade do sistema estar no estado* S_j *no instante* $t+1$ *dado que estava no estado* S_i *no instante* t*.*
- *B a matriz de distribuição de probabilidade de observação (ou emissão), em que cada elemento* $\{b_j(k)\}$ *de B corresponde à probabilidade* $P(O_t = V_k | q_t = S_j)$*, ou à probabilidade do símbolo* V_k *ser observado no instante t, dado que o sistema está no estado* S_j *nesse mesmo instante.*
- $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$*, onde cada* $\pi_j = P(q_1 = S_j)$*, ou a probabilidade do sistema estar no estado* S_j *no tempo inicial* $t = 1$*.*
- $\lambda = (A, B, \pi)$ *é a notação compacta dos parâmetros de um HMM, conforme (RABINER; JUANG, 1986, p.8).*

Rabiner e Juang (1986) elencaram três problemas em que modelos ocultos de Markov podem ser aplicados em situações reais, sendo eles: i) o problema da avaliação (do original, em inglês, *likelihood* ou *probability evaluation*); ii) o problema da decodificação (*decoding* ou *optimal state sequence*); e iii) o problema do aprendizado (*learning* ou *parameter estimation*). A seguir, cada um dos problemas é formalmente apresentado:

Avaliação : dado o modelo $\lambda = (A, B, \pi)$ e uma sequência de observações $O = \{o_1, o_2, o_3, \ldots, o_k\}$, qual a probabilidade dessa sequência ter sido gerada pelo modelo? A solução é dada pelo algoritmo *Forward e Backward*.

Decodificação : dado o modelo $\lambda = (A, B, \pi)$ e uma sequência de observações $O = \{o_1, o_2, o_3, \ldots, o_k\}$, qual é a sequência mais provável de estados a partir da observações? A solução é dada pelo algoritmo *Viterbi*.

Aprendizado : dado o modelo $\lambda = (A, B, \pi)$ e uma sequência de observações $O = \{o_1, o_2, o_3, \ldots, o_k\}$, como ajustar os parâmetros que melhor representam $P(O|\lambda)$. A solução é dada pelo algoritmo *Baum-Welch*. .

Conforme Rabiner e Juang (1986, p.10), a técnica formal para resolver o problema da decodificação e encontrar a melhor sequência de estados, é o algoritmo Viterbi.

2.3.2 Algoritmo Viterbi

O algoritmo Viterbi (VITERBI, 1967) foi proposto em 1967, como um método de decodificação de códigos convolucionais. Conforme (FORNEY, 1973, p.268), esse algoritmo tem sido utilizado para uma variedade de problemas, sendo uma solução ótima recursiva para o problema de estimar a sequência de estados ocultos de um processo de Markov de estado finito em tempo discreto.

Para Jurafsky e Martin (2018, p.8), o algoritmo Viterbi é um tipo de programação dinâmica que pode ser visto como uma treliça[1]. Dado o modelo $\lambda = (A, B, \pi)$ e uma sequência de observações $O = (o_1, o_2, o_3, \ldots, o_T) \subseteq V^T$. Deseja-se encontrar qual a sequência mais provável de estados $Q = (q_1, q_2, \ldots, q_T) \subseteq S^T$.

Ainda conforme o autor, a ideia é processar a sequência de observações da esquerda para a direita, preenchendo a treliça. Cada célula da treliça, $v_t(j)$, representa a probabilidade do HMM estar no estado j depois de observar o primeiro t e passando pela sequência de estado mais provável $q_1, \cdots, q_{t-1}$, dado o modelo λ. O valor de cada célula $v_t(j)$ é calculado tomando-se recursivamente o caminho mais provável que poderia levar a essa célula. Formalmente, cada célula expressa a probabilidade, conforme a Equação 3.

$$v_t(j) = \max_{q_1,\ldots,q_{t-1}} P(q_1 \cdots q_{t-1}, o_1, o_2 \cdots o_T, q_t = j|\lambda) \tag{3}$$

Representa-se o caminho mais provável, assumindo o máximo de todas as possíveis sequências de estados anteriores $\max_{q_1,\ldots,q_{t-1}}$. Para um dado estado q_j no tempo t, o valor $v_t(j)$ é calculado como mostrado na Equação 4

$$v_t(j) = \max_{i=1}^{N} \{v_{t-1}(i) a_{ij} b_j(o_t)\} \tag{4}$$

Os três fatores multiplicados na equação 4 são apresentados abaixo:

[1] ETIM fr. treillis 'tapume vazado formado por entrelaçamento de ripas de madeira'

- $v_{t-1}(i)$ é o valor da probabilidade do caminho mais provável até a observação $t-1$ estar no estado S_i
- a_{ij} valor da probabilidade de transição do estado S_i para o estado S_j.
- $b_j(o_t)$ é a probabilidade de se ter a observação o_t no estado S_j

O algoritmo completo de Viterbi pode ser descrito da seguinte forma:

1. Inicialização

$$v_1(j) \leftarrow \pi_j b_j(o_1) \quad 1 \leq j \leq N$$
$$bt_t(j) \leftarrow 0 \quad 1 \leq j \leq N$$

2. Recursão

$$v_t(j) \leftarrow \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t) \quad 1 \leq j \leq N, 1 < t \leq T$$
$$bt_t(j) = \leftarrow \arg\max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t) \quad 1 \leq j \leq N, 1 < t \leq T$$

3. Finalização

$$P^* \leftarrow \max_{i=1}^{N} v_T(i)$$
$$q_T^* \leftarrow \arg\max_{i=1}^{N} v_T(i)$$

O algoritmos Viterbi fornece uma solução eficiente para o problema da decodificação, que está diretamente ligado à predição de resultados de produção agrícola. A pergunta que deve ser respondida, nesse caso, é: qual é o resultado de produção mais provável, dado que as variáveis observadas nos últimos *n* instantes de tempo (dias, semanas, etc.) são os seguintes? A modelagem dos dados via modelos ocultos de Markov pode fornecer esse tipo de resposta, ao mesmo tempo que as probabilidades associadas às matrizes de emissão (principalmente) possibilitam uma análise das relações causais entre as variáveis modeladas.

2.3.3 Ferramentas para construção de modelos de Markov

Conforme Spedicato *et al.* (2014), as cadeias de Markov representam uma classe de processos estocásticos de grande interesse para um amplo espectro de aplicações práticas. Em particular, as cadeias de Markov em tempo discreto (DTMC – *discrete-time Markov chains*) permitem modelar as probabilidades de transição entre estados discretos com auxílio de matrizes. Vários pacotes em linguagem *R* implementam funcionalidades relacionadas aos modelos de Markov, entre esses pacotes, destacam-se: *HMM, Uhmm, mcmcr, marcovchain e seqHMM*.

O pacote *HMM* (HIMMELMANN, 2015) é uma biblioteca fácil de utilizar, configurar e realizar inferências tanto em cadeias de Markov de tempo discreto quanto para modelos ocultos de Markov. O pacote *Uhmm* (POISSON-CAILLAULT; TERNYNCK, 2019) oferece uma interface para detectar eventos comuns ou extremos em um conjunto de dados e para caracterizar sua dinâmica, construindo um modelo oculto de Markov usando um algoritmo de aprendizado não supervisionado. O pacote *mcmcr* (THORLEY, 2019) fornece funções e classes para armazenar, manipular e resumir modelos de Monte Carlo via cadeias de Markov, modelos também conhecidos como *Monte Carlo Markov Chain* (MCMC).

O pacote *markovchain* fornece um conjunto de funções e métodos S4[2] para criar e gerenciar cadeias de Markov em tempo discreto com mais facilidade. Também oferece funções para realizar análises estatísticas, como ajuste e modelagem de variáveis aleatórias e probabilísticas, como análise de propriedades estruturais (SPEDICATO, 2019). No pacote estão disponíveis alguns exemplos aplicados aos tópicos de economia, finanças e ciências naturais.

Segundo (HELSKE; HELSKE, 2019b), o pacote *seqHMM* foi projetado para treinar vários tipos de modelos ocultos de Markov para dados sociais em sequência e outras séries temporais categóricas. O pacote suporta modelos para um ou vários domínios com uma ou várias sequências paralelas (canais). Fornece ainda funções para avaliar e comparar modelos, bem como funções para visualização de dados sequenciais multicanal e modelos ocultos de Markov (HELSKE; HELSKE, 2019a). Todos os principais algoritmos são escritos em *C++* com suporte à computação paralela.

Existem pacotes de software desenvolvidos em outras linguagens como *C*, *JAVA* e *Python*, mas verificou-se que a maioria dos projetos foi descontinuado, não recebendo mais atualizações, inviabilizando o uso sobre plataformas atualizadas de sistemas operacionais e linguagens de programação. Apesar de não ser viável a utilização, a teoria utilizada para o desenvolvimento e a documentação são um legado importante deixado por essas ferramentas. A seguir, são apresentados alguns desses pacotes estudados.

A Biblioteca Geral para Modelo Oculto de Markov (GHMM, 2013) é uma biblioteca de funções, escrita em linguagem *C*, disponível gratuitamente e que implementa estruturas de dados e algoritmos eficientes para modelos ocultos de Markov com emissões discretas e contínuas. Disponibiliza *wrappers Python*, que fornecem uma interface muito mais agradável e funcionalidades adicionais. A *GHMM* está licenciada

[2]O S4 tem definições de classe formais, que descrevem a representação e a herança de cada classe.

sob a LGPL e a última atualização foi no ano de 2013.

HMMlib é uma biblioteca desenvolvida em *C++* para construir e analisar modelos de ocultos de Markov. Ela explora o poder computacional de *CPU* modernas com múltiplos núcleos, para obter acelerações significativas, (SAND *et al.*, 2010).

O software *EMMA*, desenvolvido em *JAVA*, fornece uma série de ferramentas que possibilitam o uso dos conceitos básicos de construção, validação e análise de modelos Markov (SENNE *et al.*, 2012). O *EMMA* foi descontinuado dando origem ao *PyEMMA*, que é uma biblioteca *Python* para a estimação, validação e análise de modelos de cinética molecular de Markov e outros modelos cinéticos e termodinâmicos a partir de dados de dinâmica molecular (SCHERER *et al.*, 2015).

O software *NVHMM* – modelo oculto de Markov multivariável não-homogêneo, disponibiliza algoritmos para modelar séries temporais multivariadas com modelos ocultos de Markov (KIRSHNER, 2019). O software está disponível nas plataformas *Linux*, *UNIX (Solaris)* e *MacOS X*. Existe um manual de informações básicas necessárias para instalar e executar o software e sua última atualização foi em 2007.

2.4 Clima

De acordo com a Classificação Climática Internacional de Köppen (ALVARES *et al.*, 2013), o clima na região onde se localiza a área utilizada para o estudo de caso desse trabalho é classificado como *Cfa*, subtropical úmido *(C)*, clima oceânico, sem estação seca *(f)* e com verão quente *(a)*, com temperatura média do mês mais frio entre $-3°C$ e $18°C$, temperatura média do mês mais quente acima de $22°C$, caracterizando-se por grande amplitude térmica, com temperaturas altas durante o dia e amenas à noite durante o verão e algumas temperaturas abaixo de zero no inverno, com ocorrência de geadas e alguns registros de neve.

Conforme Wrege *et al.* (2012, p49), especificamente, na região sudoeste do Rio Grande do Sul, o clima é semelhante ao do Uruguai. A precipitação é mais bem distribuída ao longo do ano e a evapotranspiração é maior no verão (dezembro, janeiro, fevereiro e março). O deficit hídrico nesses messes prejudica a fase de desenvolvimento vegetativo das plantas, principalmente nas culturas anuais.

Segundo o relatório da EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA (2019), as condições climáticas adversas são apontadas como as principais responsáveis pelo baixo rendimento nas lavouras de soja. Nos anos de 2015

e 2016 a ocorrência de condições desfavoráveis coincidiu com a fase de crescimento vegetativo, floração e enchimento de grãos, o que resultou no baixo desenvolvimento, abortamento de flores e vagens vazias. As secas severas, na fase vegetativa, reduzem o crescimento da planta e diminuem a área foliar e o rendimento de grãos.

O crescimento das culturas é um processo cumulativo e as condições climáticas em qualquer mês de crescimento afetam o rendimento das colheitas, por esse motivo a importância da utilização da serie de dados completas para o estudo da fase de crescimento (CAI *et al.*, 2013).

2.5 Compactação do solo

O solo da região, conforme Macedo (1984), em uma classificação mais detalhada, porém mais antiga, apresenta horizonte B textural e argila de atividade alta (não hidromórfico). O nome regional do solo é Bexigoso (*Bx*) e classificado como Brunizem raso textura argilosa, relevo ondulado e substrato granito. Segundo o Sistema Brasileiro de Classificação de Solos (SANTOS *et al.*, 2018), mais atual, é classificado como Planossolo Háplico Eutrófico (*Sx*e).

A compactação do solo é prejudicial ao desenvolvimento das plantas, pois causa a restrição ao crescimento e desenvolvimento das raízes, implicando em prejuízos na produtividade. O estudo desta variável objetiva o monitoramento da compactação do solo através do mapeamento de resistência à penetração. Conforme Mantovani (1987, p.52), "o solo está compactado quando a proporção do volume total de poros para o do solo é inadequada ao máximo desenvolvimento de uma cultura ou manejo eficiente do campo", ainda segundo o autor, "a compactação do solo pode ser considerada em relação à porosidade e densidade do solo e à resistência à penetração". A resistência à penetração é fortemente influenciada pela umidade e textura do solo, sendo os solos mais úmidos e com texturas diferentes mais suscetíveis a compactação.

Segundo Machado (2003), a mecanização agrícola resulta na compactação do solo, que consiste em uma diminuição do seu volume não saturado, decorrente de uma compressão que causa a expulsão do ar do solo e, consequentemente, o aumento da densidade. Ainda segundo o autor, o solo compactado resulta em diminuição do crescimento das raízes em profundidade, propensão à morte em períodos de seca, acúmulo de água na superfície que prejudica a respiração das raízes e favorece a erosão.

Wolkowski e Lowery (2008) afirmam que a compactação é a consolidação física

do solo por uma força aplicada que destrói a estrutura, reduz a porosidade, limita a infiltração de água e ar, aumenta a resistência à penetração das raízes e muitas vezes, reduz a produtividade das culturas. Grande parte da compactação que causa o limite de rendimento é causada pelo tráfego de equipamentos pesados sobre solos úmidos. O resultado do processo de compactação é um solo denso com poucos poros grandes, drenagem interna deficiente e aeração limitada. Para Akker e Soane (2005), a compactação do solo é um processo de densificação no qual a porosidade e a permeabilidade são reduzidas, a resistência aumentada e a estrutura do solo parcialmente destruída

Conforme FALKER (2009, p.3), de forma geral, a bibliografia indica que "com valores de 2000 Kpa de resistência à penetração, qualquer tipo de solo que se encontre em capacidade de campo, já existe restrições ao crescimento radicular". O relatório técnico FALKER indica os níveis críticos de compactação em função da classificação dos solos e seu teor de argila. Os solos leves com teor de argila abaixo de 20%, solos médios entre 20% e 50% e solos pesados acima de 50%, conforme a Tabela 3. Níveis toleráveis são aqueles em que não existe perda de produtividade em função da compactação. Níveis críticos são aqueles em que, de forma geral, ocorrerão perdas pela compactação. Nos níveis intermediários existe potencial para perdas de produtividade, mas essa depende de associação com outros fatores.

Tabela 3 Níveis críticos de compactação em função da classificação dos solos

Níveis	Leves	Médios	Pesados
Tolerável	< 2000 Kpa	< 2000 Kpa	< 2500 Kpa
Intermediário	> 2500 e < 3000 Kpa	> 2000 e < 3500 Kpa	> 2500 e < 4000 Kpa
Crítico	> 3000 Kpa	< 3500 Kpa	< 4000 Kpa

Fonte: (FALKER, 2018) e classificação em função do teor de argila, segundo EMBRAPA

.

Consta nos sistemas de produção da (EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA, 2014) que os solos arenosos (leves) apresentam teores de areia superiores a 70% e argila inferior a 15%, são permeáveis, de baixa capacidade de retenção de água e de baixo teor de matéria orgânica. Os solos medianos apresentam equilíbrio entre areia, silte e argila, são permeáveis, bem drenados, com média capacidade de retenção de água e médio índice de erodibilidade. Os solos argilosos (pesados), com teores de argila acima de 35%, apresentam baixa permeabilidade, alta capacidade de retenção de água, grande resistência à penetração, o que torna o solo muito suscetível à compactação.

2.6 Sistemas de informações geográficas

De modo geral, a condição fitossanitária das plantas, a influência do clima na produtividade e a condição física do solo podem ser observadas diretamente em campo pelos especialistas. Mas, para obter dados mais detalhados sobre essas condições, pode-se fazer uso de sensores que possam trazer informações mais detalhadas sobre o fenômeno estudado. Nesse contexto, o uso de técnicas de Agricultura de Precisão (AP) e Sistemas de Informações Geográficas (SIG) podem auxiliar na obtenção de análises mais detalhadas e precisas.

Os SIG sempre foram utilizados para o desenvolvimento da base cartográfica, manipulação e classificação de imagens e geração de mapas a partir de diversas bases georreferenciadas. É natural que alguns SIG tenham disponíveis ferramentas para o desenvolvimento de modelos de mudança da terra, integrados ou em forma de módulos. Conforme Christofoletti (1999), os procedimentos de modelagem por meio do uso de SIG são necessários, pois a espacialidade é característica inerente aos sistemas ambientais.

Por meio de um SIG pode-se analisar os fenômenos, variáveis ou eventos com referência espacial conhecida. Segundo (BURROUGH, 1986), um SIG pode ser definido como um conjunto de ferramentas para coletar, armazenar, transformar e utilizar dados sobre o mundo real. Estes sistemas disponibilizam várias funcionalidades por meio de um conjunto de funções nativas e de complementos, sendo possível visualizar, gerir, editar, analisar dados e criar mapas para impressão.

Existem vários SIG disponíveis que, apesar de terem o mesmo propósito, diferem em estrutura, licença, documentação, integração com outros sitemas, tipos de dados e funcionalidades que oferecem. Como ferramentas sem custo ao usuário, destacam-se o *Geographic Resources Analysis Support System* (GRASS GIS) (NETELER; MITASOVA, 2008) e o *QGIS* (QGIS, 2019), que são projetos oficiais da *Open Source Geospatial Foundation* (OPEN SOURCE GEOSPATIAL FOUNDATION, 2018). Ambos são softwares multiplataforma e suportam diversos formatos de dados vetoriais, *raster* e bases de dados. Entre as ferramentas comerciais, destacam-se o ArcGIS (ESRI, 2018) e o *IDRISI GIS Analysis* (EASTMAN, 2018a), ferramenta que integra o *TerrSet* (EASTMAN, 2018b), que é um sistema de software integrado para modelagem e monitoramento geoespacial.

No que tange ao armazenamento de dados espaciais, as ferramentas hoje disponíveis fazem o armazenamento dos dados matriciais e vetoriais em estruturas de

arquivos e diretórios utilizando formato próprio ou de intercâmbio. Os formatos de intercâmbio mais utilizados são *shapefile* (.shp) (ESRI, 1998) para o formato vetorial e *geotiff* (.tif) (GEOTIFF, 2018) para o formato raster. O armazenamento de atributos pode ser feito em arquivos texto ou em Sistemas de Gerenciamento de Banco de Dados (SGDB), sendo que a maioria das ferramentas oferece suporte às duas opções.

Neste trabalho, optou-se pelo uso do *QGIS* e pelo armazenamento de dados espaciais em shapefiles. O *QGIS* oferece os principais recursos de importação, exportação e manipulação de dados raster e vetorial necessários para o desenvolvimento deste trabalho. A integração com o *GRASS* acrescenta um conjunto importante de funções e módulos para a resolução de problemas.

O software *Idrisi*, através do módulo *CA-Markov*, e o *Dinamica EGO* oferecem funcionalidades para o cálculo de mudanças (matrizes de transição) através de métodos Markovianos. Na documentação disponível para o software não fica claro se o método disponível permite a construção de um modelo oculto de Markov.

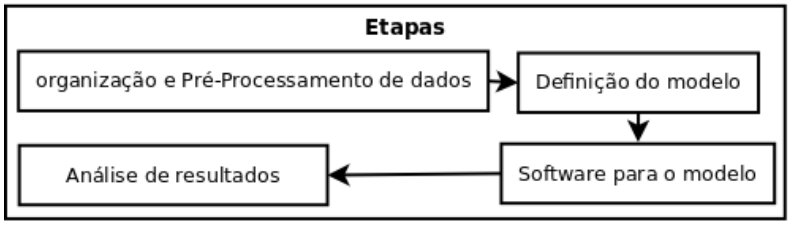
3 MATERIAL E MÉTODOS

3.1 Caracterização e etapas do método

O trabalho realizado caracteriza-se como uma pesquisa exploratória com estudo de caso delimitado por uma área de produção da EMBRAPA Pecuária Sul, apoiado em pesquisa bibliográfica relevante ao problema e uso técnicas quantitativas de modelagem e análise de dados.

Para atingir o objetivo proposto neste trabalho, a metodologia foi dividida em quatro etapas, apresentadas na Figura 3.

Figura 3 – Etapas da metodologia



Fonte: Autor (2019)

A primeira etapa diz respeito à organização dos dados de produtividade, meteorológicos e atributos físicos do solo, os quais estão detalhados, respectivamente, nas seções 3.2.1, 3.2.2 e 3.2.3.

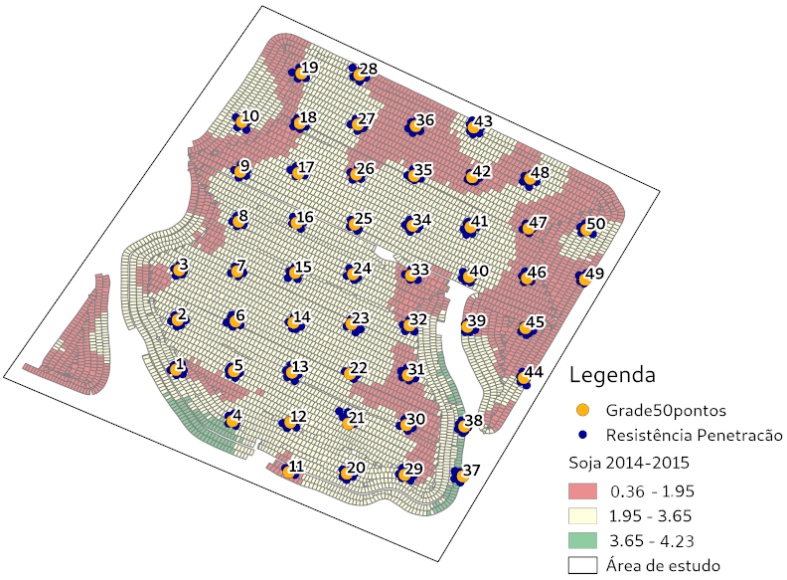
Na segunda etapa foi feita a definição de um modelo de Markov, com o objetivo de estudar a variabilidade e predizer a produtividade de soja a partir do histórico de produção, variáveis meteorológicas, dados de resistência à penetração do solo e dados de fertilidade química do solo. Na terceira etapa foi feita a definição dos recursos de software a serem utilizados e desenvolvidos para o processamento do modelo. Por fim, na quarta etapa foi feita a análise dos resultados a fim de entender a variabilidade da produtividade na área de estudo.

3.2 Organização e pré-processamento dos dados

Os dados utilizados neste trabalho são provenientes de duas fontes distintas: os dados meteorológicos provém da base dados do Instituto Nacional de Meteorologia

(INMET) e os dados de solo e de produção são provenientes de trabalhos de campo feitos pela equipe técnica da EMBRAPA Pecuária Sul. O formato dos dados e os processos usados para organizá-los como entrada do sistema desenvolvido neste trabalho são descritos a seguir.

Figura 4 – Grade de 50 pontos amostrais distribuídos no talhão de estudo



Fonte: Autor (2019)

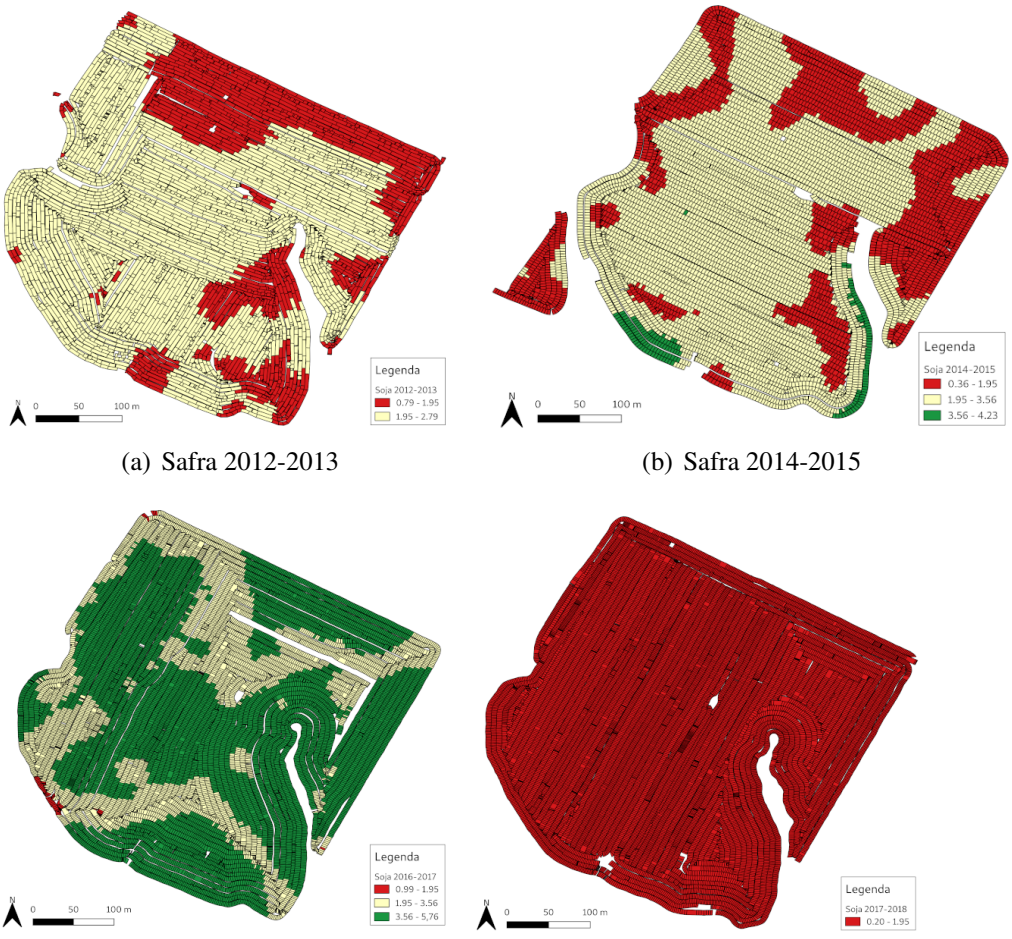
Os dados referentes ao solo, coletados em campo pela EMBRAPA, foram organizados em uma grade de cinquenta pontos, conforme mostra a Figura 4. Em cada um desses pontos foram coletados dados de resistência à penetração e de fertilidade química do solo. Os dados em planilha de cálculo foram importados no *LibreOffice Calc* (THE DOCUMENT FOUNDATION, 2019), reorganizados e, após esse procedimento, exportados para arquivos em formato de texto puro (ASCII) que foram posteriormente utilizados como entrada no *QGIS* (QGIS, 2019) e na linguagem *R* (R CORE TEAM, 2019).

3.2.1 Dados de produtividade

Os dados de produtividade foram gerados durante a colheita e disponibilizados pela equipe técnica da EMBRAPA no formato *Shapefile* (.shp). Esses dados são oriundos do controlador de agricultura de precisão da Stara TOPPER 4500 (STARA, 2011), o

qual registra os dados no momento da colheita. Os arquivos shapefile contêm polígonos vetoriais associados a um banco de dados local no formato DBF (.dbf) com dados da colheita, que foram importados no QGIS e disponíveis na forma de *layers* (camadas). As Figuras 6(a), 6(b), 6(c) e 6(d) mostram a produtividade em cada umas das safras, categorizadas em três classes (baixa, média e alta). Note-se que a produtividade não é uniforme dentro da área. Algumas regiões consistentemente produziram mais do que outras, em todos os períodos. Algumas regiões internas, contudo, tiveram produção com alto índice de variação a cada safra.

Figura 5 – Produtividade da soja em t/ha nas safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018



(a) Safra 2012-2013 (b) Safra 2014-2015

(c) Safra 2016-2017 (d) Safra 2017-2018

Fonte: Autor (2019)

Juntamente aos dados de produtividade, o controlador de agricultura de precisão instalado na colheitadeira coleta os dados de altitude, por meio de um Sistema de Posicionamento Global (GPS - *global positioning system*). Todo GPS apresenta um erro

padrão associado e isso determina a diferença de altitude no mesmo ponto entre uma safra e outra.

3.2.2 Dados meteorológicos

Os dados meteorológicos, apresentados na Tabela 4, são referentes à estação meteorológica de Bagé (OMM 83980)[1], que opera desde 01/01/1912, localizada na latitude $-31.305661°$, longitude $-54.119352°$, com altitude de 245.66 metros, distante 12 quilômetros em linha reta da área de estudo e na mesma faixa de altitude.

Tabela 4 Variáveis diárias relacionadas ao clima

Nome	Unidade	Descrição
Precipitação	mm	acumulado da últimas 24 horas - coleta às 12 UTC
Insolação	h	medida da irradiação solar em uma superfície por unidade de tempo
Deficit	mm	Deficiência hídrica
Excedente	mm	Excesso hídrico

Fonte: Autor (2019)

Os dados históricos para o intervalo entre 01/01/2012 e 31/01/2019 são oriundos do Banco de Dados Meteorológicos para Ensino e Pesquisa (INSTITUTO NACIONAL DE METEOROLOGIA, 2019a), disponível no website do Instituto Nacional de Meteorologia (INMET). Esse dado foi coletado em estações meteorológicas convencionais da rede de estações do INMET com milhões de informações referentes às medições diárias, de acordo com as normas técnicas internacionais da Organização Meteorológica Mundial (INSTITUTO NACIONAL DE METEOROLOGIA, 2019b). Os dados de precipitação foram coletados as 12h UTC[2] (09 horas no horário de Brasília) e os dados de insolação as 00 UTC (21 horas no horário de Brasília).

O balanço hídrico de cultivo e perda de produtividade foi calculado com o Sistema de Suporte à Decisão na Agropecuária (SISDAGRO) (INSTITUTO NACIONAL DE METEOROLOGIA, 2019c), também desenvolvido pelo INMET. O balanço hídrico de cultivo específico de uma cultura visa calcular o balanço de água no solo e leva em consideração tanto o tipo de vegetação quanto a sua fase de crescimento

[1] código da Organização Meteorológica Mundial
[2] Coordinated Universal Time

e desenvolvimento. Os parâmetros utilizados foram: a) data de emergência (aproximadamente oito dias após o plantio), conforme a Tabela 5; b) cultura soja com ciclo de 150 dias; c) estação de Bagé-RS convencional; d) solo do tipo médio e e) capacidade de água disponível (CAD) igual a 40. O CAD representa a capacidade de água disponível quando a raiz estiver plenamente desenvolvida, sendo estimado em função da cultura e do solo em questão.

Tabela 5 Datas de plantio, emergência, colheita e número de dias das safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018

Plantio	Emergência	Colheita	Nº dias
09/12/12	17/12/12	05/06/13	178
11/12/14	19/12/14	08/05/15	148
21/11/16	29/12/16	25/04/17	155
21/12/17	29/12/17	18/05/18	148

Fonte: Autor (2019)

O modelo disponibilizado pelo INMET segue o referencial teórico dos seguintes autores: (ALLEN *et al.*, 1998), (DOORENBOS; KASSAM, 1979) e (THORNTHWAITE; MATHER, 1955). O cálculo do balanço hídrico diário segue a equação $ETc = ETP * Kc$, onde ETc é a evapotranspiração da cultura, ETP é a evapotranspiração potencial (ou de referência) e Kc o coeficiente de cultura (aumenta conforme o crescimento da cultura). A escolha do tipo de solo (arenoso, argiloso ou médio) e do cultivo (soja) determina a capacidade de água disponível (CAD). A diferença entre a evapotranspiração da cultura (ETc) e evapotranspiração real (ETr) determina a deficiência hídrica. As figuras do Apêndice E mostram os dados referentes ao deficit hídrico e excedente hídrico para cada uma das safras. Para a criação das séries temporais foi utilizado a linguagem *R* e as bibliotecas *ggplot2*, *forecast*, *tseries*, *scales* e *dplyr*.

3.2.3 Dados de solo

Os dados de resistência à penetração do solo foram coletados em campo pelos técnicos da EMBRAPA, por meio do medidor eletrônico de compactação do solo *penetroLOG*, desenvolvido pela empresa Falker (FALKER, 2018). O conjunto de dados corresponde a 50 observações, com o número de 4 a 8 amostras por observação, que registra a resistência à compactação do solo entre 0 e 40 centímetros. O registro foi realizado a cada centímetro, com unidade de medida em quilopascal (*kPa*). Para

cada ponto foi calculado o percentual da área nos níveis tolerável, intermediário e crítico, conforme a orientação especificada pela empresa fabricante do equipamento em (FALKER, 2009). Como a área apresenta baixa compactação, no nível crítico muitos pontos não apresentaram valor e no nível intermediário os percentuais são baixos, as duas classes foram unificadas para melhor representação dentro do modelo. Percentuais acima de 85% ficaram na classe baixa e percentuais abaixo de 85% ficaram na classe alta.

Os dados de fertilidade química do solo correspondem a 50 amostras georreferenciadas e 22 variáveis, que representam elementos físicos e químicos do solo, das quais 5 foram utilizadas, conforme a Tabela 6. O teor de argila (classe 3) foi utilizado para a definição do teor de fósforo no solo, extraído pelo método Mehlich-1, que é uma tentativa de reproduzir em laboratório o processo de absorção de fósforo (P) pelas plantas no campo. O teor de fósforo, originalmente com cinco classes (muito baixo, baixo, médio, alto e muito alto), foi agrupado em duas classes, baixo (muito baixo, baixo e médio) e alto (alto e muito alto), pois havia somente uma observação em cada uma das classes extremas.

A capacidade de troca de cátions (CTC) a pH 7,0 foi utilizado para a definição das classes do teor de potássio no solo. A matéria orgânica é dada em percentual. A interpretação dos resultados da análise de solos foi feita segundo SOCIEDADE BRASILEIRA DE CIÊNCIA DO SOLO (2004, p.50).

Tabela 6 Variáveis relacionadas ao solo

Nome	Unidade	Descrição
Argila	%	teor de argila no solo
P	mg/L	teor de fósforo
K	mg/L	teor de potássio
Mat. Orgânica	%	matéria orgânica
CTC pH 7	$cmol_c/dm^3$	capacidade de troca de cátions a pH 7,0

Fonte: Autor (2019)

3.3 Definição de classes

Para as variáveis em que não havia uma definição de valores de classes na bibliografia, essa definição foi feita através de uma distribuição de frequências, onde o número de elementos pertencentes a cada classe é chamado de frequência de classe (ASSIS; ARRUDA; PEREIRA, 1996, p.19). A Tabela 7 mostra os intervalos de classe

para cada variável usada neste trabalho.

Tabela 7 Frequências de classes das variáveis

Variável	Baixo	Médio	Alto
Peso (ton)	abaixo de 1,95	1,9 a 3,56	acima de 3,56
Precipitação (mm)	450,3 e 671,4		671,5 e 892,5
Insolação (h)	1093,6 e 1226,9		1227 e 1360,2
RP (Kpa)	abaixo de 2000		acima de 2000
Deficit (mm)	97,28 e 164,16		164,17 e 231,04
Excedente (mm)	181,36 e 488,545		488,546 e 795,730
Altitude (m)	231,3 e 243,55		243,56 e 255,80
M. Orgânica (%)	abaixo de 2,5	acima de 2,5	
Fósforo ()	abaixo de 12		acima de 12
Variável	Médio	Alto	Muito Alto
Potássio (%) (CTC médio 5,1 a 15)	41 a 60	61 a 120	acima de 120
Potássio (%) (CTC alto > 15)	61 a 90	91 a 180	

Fonte: Autor (2019)

Conforme orienta Spiegel e Stephens (2008, p. 38), a amplitude total foi definida calculando a diferença entre o maior e o menor valor do conjunto de dados. Após, a amplitude total foi dividida por três, utilizando o critério de mesmo tamanho de classe e número de classes pré-fixado, que definiu a amplitude de classe. O passo final foi determinar as frequências de classe ou intervalos de classe.

Para a definição das classes foi desenvolvida uma função na linguagem *R*, disponível no Apêndice D, que recebe um conjunto de dados e o número de classes desejado, e retorna os limites das classes. As variáveis peso, precipitação, insolação, excedente hídrico e deficit hídrico foram traduzidas em valores distintos conforme o cálculo de frequência de classes explicado acima.

3.4 Modelo oculto de Markov para análise da variabilidade espacial da produtividade

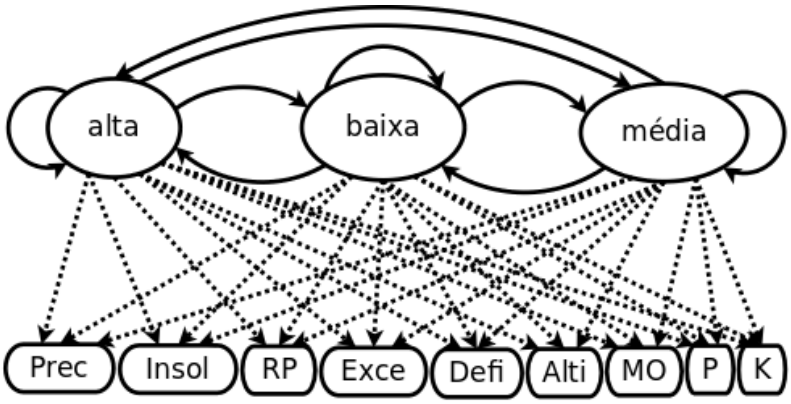
A estimativa feita por meio de modelos ocultos de Markov necessita de uma parametrização inicial: um conjunto de probabilidades iniciais, uma matriz de transição de probabilidades (parte oculta) e uma matriz de probabilidades de observação (ou emissão). Os valores iniciais também podem ser usados para definir restrições à estrutura do modelo.

As probabilidades iniciais fornecem a probabilidade de iniciar em um dado estado

oculto, neste caso, formam um vetor com a probabilidade do resultado de safra (alta, baixa, media), em cada posição do vetor. Uma probabilidade com valor maior significa uma maior chance de iniciar naquele estado.

No modelo proposto, cada *estado* representa um resultado da colheita em cada ponto, categorizado como um entre três possíveis resultados de produção: *alta*, *baixa* e *média*, constituindo a parte oculta do modelo, como mostrado na Figura 6.

Figura 6 – Modelo Oculto de Markov



Fonte: Autor (2019)

A *matriz de transição de estados*, necessária para o modelo, é calculada a partir dos dados de produção já existentes. O cálculo das probabilidades foi feito a partir dos dados de produção, referentes a 4 safras distintas, dentro das áreas demarcadas com dados de amostra do solo. Como esses dados somente estão disponíveis em 50 pontos da área, todos os dados de entrada do modelo foram extraídos das regiões correspondentes. Sendo assim, tem-se uma matriz de 50 linhas e 4 colunas, com os dados de produção de cada ponto, em cada uma das safras.

A matriz de transição resultante para os dados de produção da área objeto de estudo é dada na Equação 5. Cada intersecção de linha e coluna informa a probabilidade condicional de um resultado de produção, dado o resultado de produção do ano anterior. A soma dos valores em cada linha da matriz é 1, o que caracteriza uma distribuição probabilística discreta.

$$M = \begin{bmatrix} p_{a|a} & p_{a|b} & p_{a|m} \\ p_{b|a} & p_{b|b} & p_{b|m} \\ p_{m|a} & p_{m|b} & p_{m|m} \end{bmatrix} \tag{5}$$

As probabilidades restantes do modelo dizem respeito à relação existente entre a

variável oculta (peso produzido) e as variáveis observáveis: (i) precipitação acumulada, (ii) horas de insolação acumuladas, (iii) resistência á penetração do solo, (iv) excedente hídrico acumulado, (v) deficit hídrico acumulado, (vi) altitude, (vii) teor de matéria orgânica, (viii) teor de fósforo e (ix) teor de potássio, conforme a Equação 6 e representadas na parte inferior da Figura 6.

$$M = \begin{bmatrix} p_{a|Oa} & p_{a|Ob} & \cdots & p_{a|On} \\ p_{b|Oa} & p_{b|Ob} & \cdots & p_{b|On} \\ p_{m|Oa} & p_{m|Ob} & \cdots & p_{m|On} \end{bmatrix} \tag{6}$$

O modelo e o software proposto são flexíveis e permitem que, com poucas alterações no código, altere-se o número de variáveis observáveis, bem como os estados possíveis para cada variável. Também pode-se organizar o modelo em qualquer intervalo de tempo desejado, como por exemplo: safra, dia, fase, mês ou outro intervalo qualquer.

Para a construção do modelo optou-se pelo uso dos pacotes *markovchain* e *seqHMM*. A escolha por essas duas ferramentas se deu em função da facilidade de integração entre elas. Para as funções mais básicas, como a geração das matrizes de transição, foi utilizado o pacote *markovchain*. Para a organização dos dados em sequências, construção, análise e comparação do modelo foi utilizado o pacote *seqHMM*. O código *R* utilizado para o desenvolvimento do modelo é apresentado no Apêndice A.

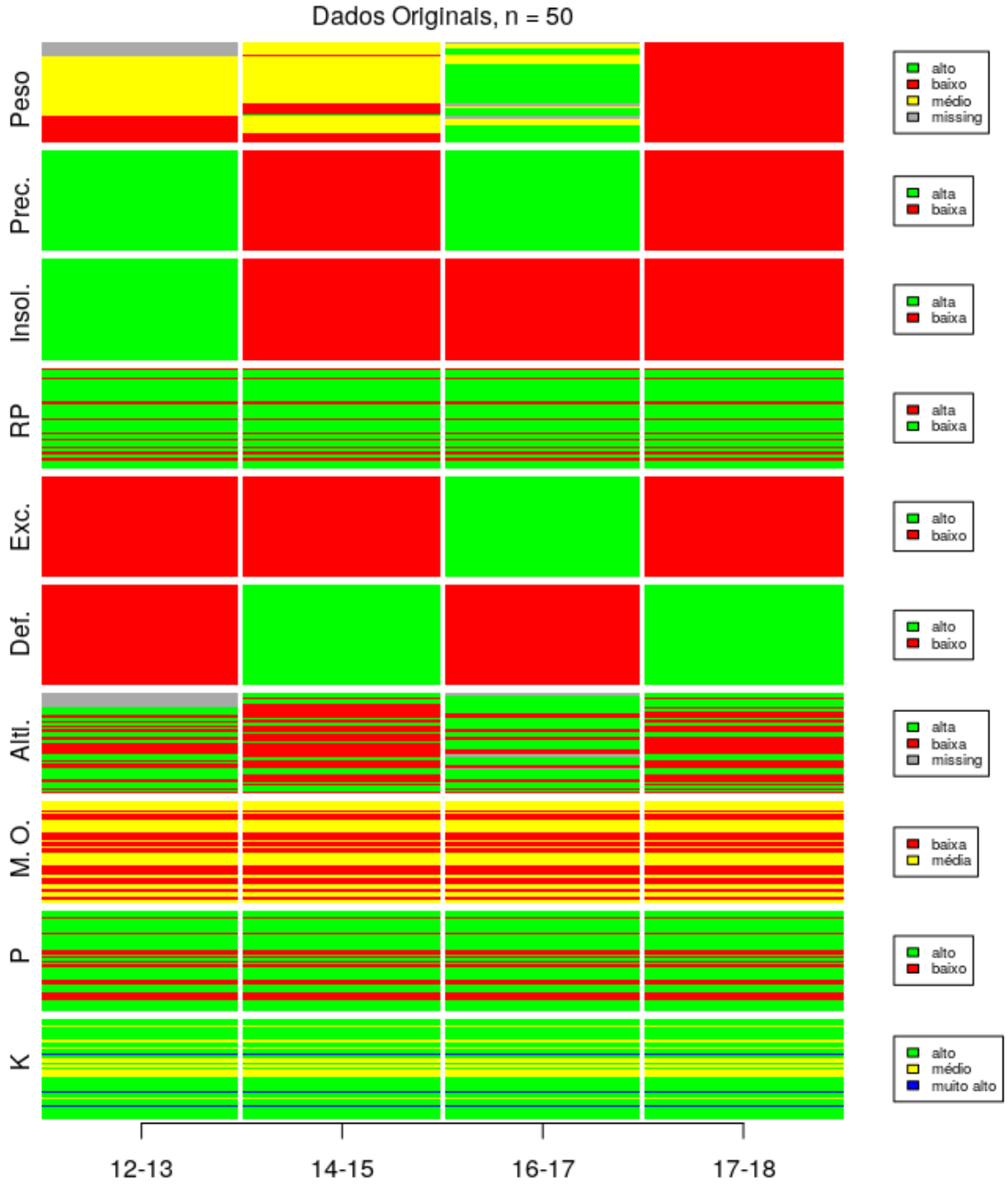
3.5 Simulação e comparação de modelos

Os dados de precipitação, insolação, excedente hídrico e deficit hídrico não apresentam variação em seu valor entre os pontos e considerados idênticos para todas as áreas, visto não ter havido medição local no períodos das safras. Os dados de resistência à penetração, matéria orgânica, fósforo e potássio, não apresentam variação entre as safras, pois foi feita somente uma coleta de dados em 2017. A Figura 7 mostra os dados originais.

Essa falta de dados justificou a simulação com dados gerados de forma aleatória e distribuídos em percentual equivalente ao dos dados originais, conforme explicado e exemplificado na sequência.

Ao gerar os dados foi mantido um percentual, por variável, equivalente ao original. Para a variável precipitação, por exemplo, houve duas safras classificadas como alta (50%) e duas como baixa (50%), então esse percentual foi mantido ao gerar a sequência de dados para precipitação alta e precipitação baixa. Da mesma forma os dados foram

Figura 7 – Sequências de dados originais, com 50 registros por variável, para as safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018

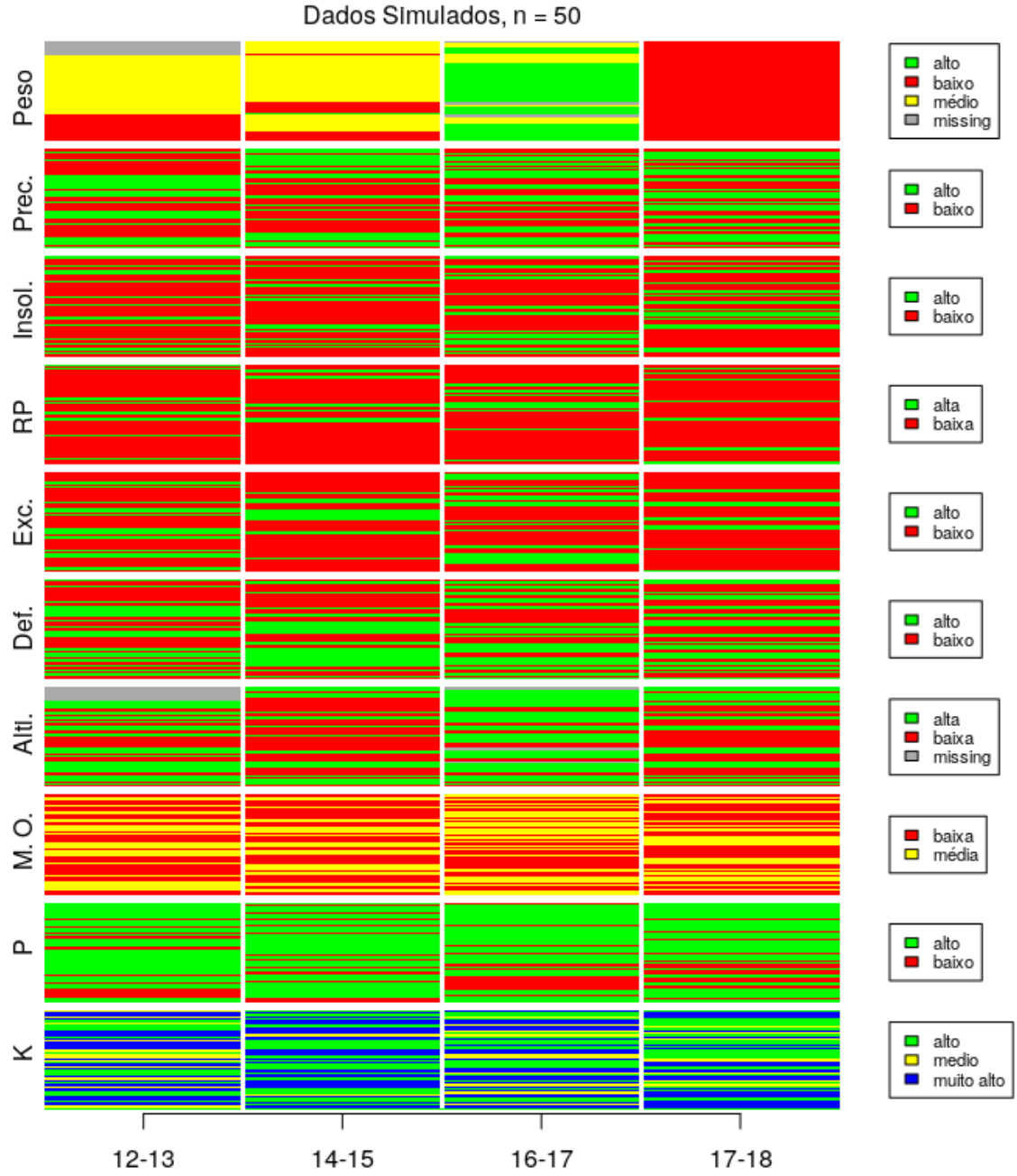


Fonte: Autor (2019)

gerados para as demais variáveis. Os dados de peso e altitude utilizados são da base original, pois continham suficiente detalhamento e variação. As probabilidades iniciais e a matriz de transição são derivadas dos dados de peso, logo, são os mesmos da base original. A Figura 8 mostra os dados simulados.

Após a simulação, foi feita uma comparação entre os modelos, com dados

Figura 8 – Sequência de dados simulados, com 50 registros por variável, para as safras 2012/2013, 2014/2015, 2016/2017 e 2017/2018



Fonte: Autor (2019)

originais e simulados: com nove, seis e cinco variáveis. No modelo com cinco variáveis foram removidos os dados meteorológicos. No modelo com seis variáveis foram removidos os dados de análise de solo. A comparação foi feita com uso do *Bayesian Information Criterion* (BIC), que permite comparar modelos com um número diferente de parâmetros. Para a interpretação do critério, quanto menor o valor, melhor o modelo

(HELSKE; HELSKE, 2019b, p.6). Conforme Bolano, Berchtold e Ritschard (2016, p.247), o número relevante de estados pode ser determinado somente em base teórica, mas como alternativa, pode-se definir este número a partir dos dados, ou seja, a partir da escolha do número de estados com base estatística. Como não existe uma regra que defina o número ótimo de estados, pode-se adotar um critério estatístico para escolher o modelo mais adequado à realidade que se quer representar. Ao aumentar o número de estados, aumenta também o número de probabilidades de emissão, probabilidades de transição e probabilidades iniciais, por consequência, aumenta a complexidade. O BIC penaliza a complexidade do modelo, sendo que a complexidade se refere ao número de parâmetros no modelo, (ROBLES *et al.*, 2012). Na etapa de escolha do modelo que melhor representa a realidade é importante a participação do especialista.

3.6 Espacialização dos resultados do modelo

Os resultados do modelo são apresentados como lista, gráficos sequenciais, gráficos percentuais, mapas espacializados em forma de pontos e mapas interpolados. Primeiramente, a sequência de dados de saída do modelo, disponível no código 1, foi recodificada como $(peso\_b = 1,\ peso\_m = 2,\ peso\_a = 3)$. Após, foram adicionadas as coordenadas geográficas para cada um dos pontos, criando um objeto geoespacial e um *grid raster* que serve de base para a interpolação dos resultados. Foi então criada uma função, para gerar o mapa com a camada limite da área plantada, limite geral do talhão e os pontos que informam a qual classe pertencem. Uma segunda função gera o mapa interpolado, que utiliza o inverso da potência das distâncias (*inverse distance weighting*), conforme (BIVAND; PEBESMA; GóMEZ-RUBIO, 2008, p.193). Esse método foi escolhido pelo número de pontos ser denso o suficiente para representar a variação espacial existente dentro da área de estudo. Conforme Neteler e Mitasova (2008, p.370), o método baseia-se na suposição de que o valor de um ponto não amostrado pode ser aproximado como uma média ponderada dos valores de pontos dentro de uma certa distância de corte ou a partir de um determinado número de pontos mais próximos, normalmente 10 a 30. A interpolação é definida a partir da distância dos pontos existentes e quanto mais longe, menos influência tem aquele ponto na interpolação. Nesse método, o resultado dos valores interpolados nunca fica fora da faixa de valores observados, diferente de outras técnicas, como o *Kriging*. Uma terceira função permite que os mapas de pontos e mapas interpolados sejam exportados para o formato *shapefile*.

Essa funcionalidade permite que os resultados possam ser utilizados como entrada em outro SIG ou sistema de agricultura de precisão embarcado ou acoplado em implemento agrícola.

Para o desenvolvimento dos mapas foi utilizada a linguagem *R* e as bibliotecas *sf*, *tmap*, *readODS*, *gstat*, *sp*, *raster* e *rgdal*. O código *R* das funções está disponível no Apêndice C.

4 RESULTADOS

4.1 O modelo oculto de Markov

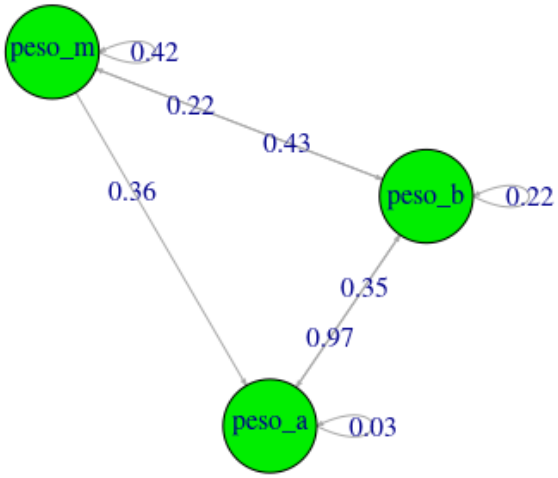
Na construção de um modelo oculto de Markov, tanto o vetor de probabilidades inicial para os estados ocultos como os valores iniciais da matriz de transição podem iniciar com valores aleatórios. Por questão de eficiência, optou-se por fornecer esses valores já calculados, a partir dos dados das safras inseridos no modelo. O vetor de probabilidades dos estados iniciais $[0.25, 0.5, 0.25]$ indica a probabilidade da sequência iniciar em um determinado estado. Nesse caso, alto com 25%, médio com 50% e baixo com 25% de probabilidade. Sendo assim, todos os estados podem ser iniciados. Essas probabilidades foram calculadas a partir dos dados das 4 safras disponíveis, onde duas delas tiveram resultados médios de produção, uma teve um valor médio alto e outra um valor médio baixo.

A matriz de transição para os estados ocultos do modelo, apresentados na Equação 7 e na Figura 9, também teve seus valores calculados a partir do conjunto de dados de peso da soja colhida em cada safra e em cada ponto.

$$M = \begin{bmatrix} & peso\_a & peso\_b & peso\_m \\ peso\_a & 0.02777778 & 0.9722222 & 0.0000000 \\ peso\_b & 0.34782609 & 0.2173913 & 0.4347826 \\ peso\_m & 0.35526316 & 0.2236842 & 0.4210526 \end{bmatrix} \tag{7}$$

Na matriz de transição as linhas representam os estados atuais e as colunas os estados futuros. Ao observar os dados pode-se verificar que a probabilidade de ter uma safra com produtividade alta e ter novamente uma safra com produtividade alta é somente de 2%. De forma oposta, a probabilidade de ter uma safra com produtividade alta e em seguida com baixa é de aproximadamente 97%. Já a probabilidade de ter uma safra com produtividade alta e após, uma média fica em 0%, ou seja, essa situação não ocorreria. De forma geral a probabilidade de transitar de uma safra com produtividade baixa para safras com produtividades alta e média, é maior do que para uma safra com produtividade baixa. Da mesma forma, a probabilidade de transitar entre uma safra com produtividade média e safras com produtividade média e alta, é maior do que para uma safra com produtividade baixa. A diagonal principal informa as probabilidades de permanecer no mesmo estado, alto 2%, baixo 21% e médio 42%. O estado oculto peso baixo é o mais

Figura 9 – Gráfico dirigido da matriz de transição



Fonte: Autor (2019)

comum, pois apresenta transições para os demais estados e para si mesmo. Salienta-se que os dados foram obtidos em uma abordagem frequentista (em oposição a uma abordagem bayesiana) a partir de somente quatro safras e à medida em que novos dados de safra forem incorporados ao sistema esses valores podem ser alterados.

As matrizes de emissão: precipitação (Prec), insolação (Insol), resistência à penetração (RP), excedente hídrico (Exce), deficit hídrico (Defi), altitude (Alti), matéria orgânica (MO), fósforo (Fosf) e potássio (K), são mostradas nas Equações 8, 9, 10, 11, 12, 13, 14, 15 e 16.

$$Prec = \begin{bmatrix} & prec\_a & prec\_b \\ peso\_a & 0.9722 & 0.0278 \\ peso\_b & 0.1733 & 0.8267 \\ peso\_m & 0.5256 & 0.4744 \end{bmatrix} \tag{8}$$

$$Insol = \begin{bmatrix} & insol\_a & insol\_b \\ peso\_a & 0.000 & 1.000 \\ peso\_b & 0.173 & 0.827 \\ peso\_m & 0.385 & 0.615 \end{bmatrix} \tag{9}$$

$$RP = \begin{bmatrix} & rp\_a & rp\_b \\ peso\_a & 0.167 & 0.833 \\ peso\_b & 0.200 & 0.800 \\ peso\_m & 0.154 & 0.846 \end{bmatrix} \tag{10}$$

$$Exce = \begin{bmatrix} & exce\_a & exce\_b \\ peso\_a & 0.9722 & 0.0278 \\ peso\_b & 0.0000 & 1.0000 \\ peso\_m & 0.1410 & 0.8590 \end{bmatrix} \tag{11}$$

$$Defi = \begin{bmatrix} & defi\_a & defi\_b \\ peso\_a & 0.0278 & 0.9722 \\ peso\_b & 0.8267 & 0.1733 \\ peso\_m & 0.4744 & 0.5256 \end{bmatrix} \tag{12}$$

$$Alti = \begin{bmatrix} & alti\_a & alti\_b \\ peso\_a & 0.722 & 0.278 \\ peso\_b & 0.467 & 0.533 \\ peso\_m & 0.500 & 0.500 \end{bmatrix} \tag{13}$$

$$MO = \begin{bmatrix} & mo\_b & mo\_m \\ peso\_a & 0.444 & 0.556 \\ peso\_b & 0.493 & 0.507 \\ peso\_m & 0.474 & 0.526 \end{bmatrix} \tag{14}$$

$$Fosf = \begin{bmatrix} & fosf\_a & fosf\_b \\ peso\_a & 0.611 & 0.389 \\ peso\_b & 0.760 & 0.240 \\ peso\_m & 0.756 & 0.244 \end{bmatrix} \tag{15}$$

$$K = \begin{bmatrix} & k\_a & k\_m & k\_ma \\ peso\_a & 0.6111 & 0.2778 & 0.1111 \\ peso\_b & 0.7733 & 0.1733 & 0.0533 \\ peso\_m & 0.6410 & 0.3077 & 0.0513 \end{bmatrix} \tag{16}$$

Ao analisar as matrizes de emissão do modelo completo e com dados reais pode-se verificar que o estado oculto peso alto *(peso_a)* está melhor associado com os seguintes

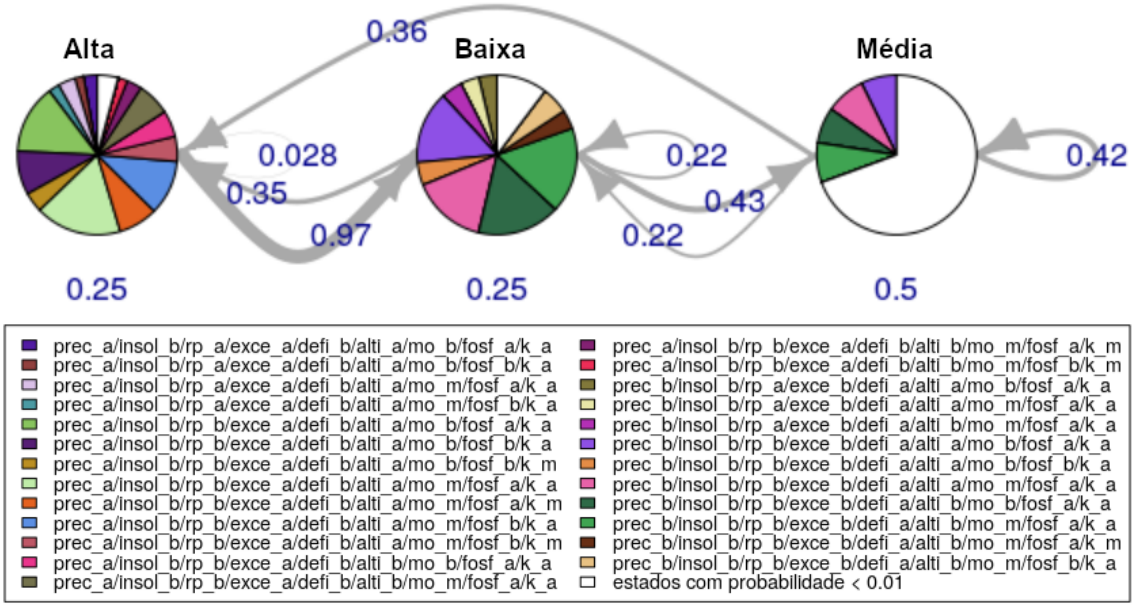
estados observáveis: precipitação alta *(prec_a)*, resistência à penetração alta *(rp_a)* ou baixa *(rp_b)*, excedente alto *(exce_a)*, deficit baixo *(defi_b)*, altitude alta *(alti_a)* ou baixa *(alti_b)*, matéria orgânica baixa *(mo_b)* ou média *(mo_m)*, fósforo alto *(fosf_a)* ou baixo *(fosf_b)* e potássio alto *(k_a)* ou médio *(k_m)*.

O estado oculto peso baixo *(peso_b)* está melhor associado com os observáveis: precipitação baixa *(prec_b)*, resistência à penetração alta *(rp_a)* ou baixa *(rp_b)*, excedente baixo *(exce_b)*, deficit alto *(defi_a)*, altitude alta *(alti_a)* ou baixa *(alti_b)*, matéria orgânica baixa *(mo_b)* ou média *(mo_m)*, fósforo alto *(fosf_a)* ou baixo *(fosf_b)* e potássio alto *(k_a)* ou médio *(k_m)*.

O estado oculto peso médio *(peso_m)* está melhor associado com os observáveis: precipitação baixa *(prec_b)*, resistência à penetração baixa *(rp_b)*, excedente baixo *(exce_b)*, deficit alto *(defi_a)*, altitude alta *(alti_a)* ou baixa *(alti_b)*, matéria orgânica baixa *(mo_b)* ou média *(mo_m)*, fósforo alto *(fosf_a)* e potássio alto *(k_a)*.

O estado observável insolação baixa *(insol_b)* apresenta probabilidades altas em relação a todos os estados ocultos. O potássio classificado como muito alto *(k_ma)* apresenta probabilidades baixas em relação aos estados ocultos.

Figura 10 – Gráfico dirigido do modelo oculto de Markov com probabilidades de emissão combinadas, probabilidades de transição e probabilidades iniciais



Fonte: Autor (2019)

Na Figura 10 é apresentado o modelo oculto de Markov como um grafo dirigido

com nove variáveis (completo) e dados reais. Os círculos/vértices representam os três estados ocultos (peso alto, peso baixo e peso médio) e os setores/partições as probabilidades de emissão das combinações dos estados observados. As setas/arestas mostram as probabilidades de transição e as probabilidades iniciais estão abaixo dos círculos/vértices. Estados observados com probabilidade menor do que 1% estão agrupados em um único setor (cor branca). As probabilidades estimadas como zero não são mostradas e para fins de legibilidade as probabilidades de transição foram arredondadas.

4.2 O caminho mais provável

O cálculo do caminho mais provável com o algoritmo Viterbi foi feito para todos os modelos. O Código 1, apresenta a sequência de estados mais provável em cada ponto, para o modelo com dados simulados e nove variáveis. Essa sequência foi utilizada para a construção dos mapas.

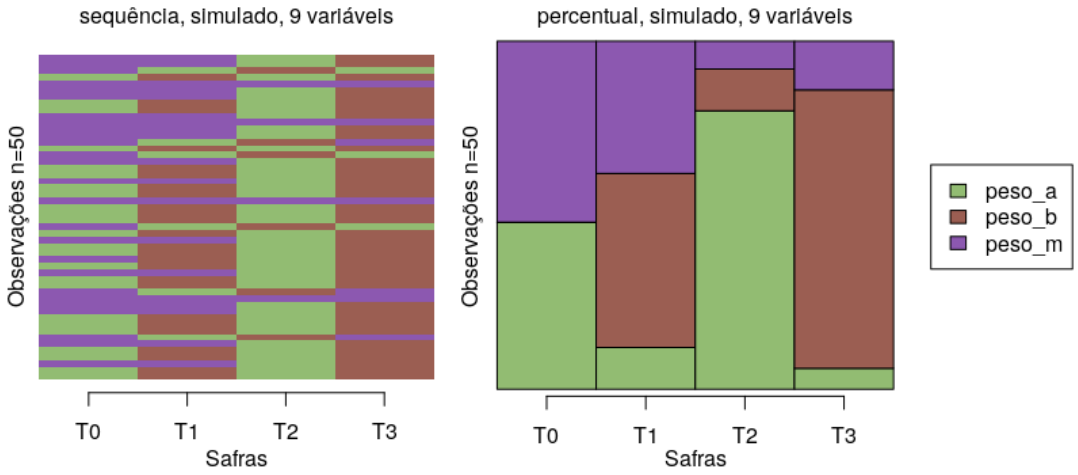
```
###Sequencia com o caminho mais provável para cada ponto de
 ↪ dados observados
1  peso_m-peso_m-peso_a-peso_b
2  peso_m-peso_m-peso_a-peso_b
3  peso_m-peso_a-peso_b-peso_a
4  peso_a-peso_b-peso_a-peso_b
5  peso_m-peso_m-peso_m-peso_m
6  peso_m-peso_m-peso_a-peso_b
7  peso_m-peso_m-peso_a-peso_b
8  peso_a-peso_b-peso_a-peso_b
9  peso_a-peso_b-peso_a-peso_b
10 peso_m-peso_m-peso_a-peso_b
11 peso_m-peso_m-peso_m-peso_m
12 peso_m-peso_m-peso_a-peso_b
13 peso_m-peso_m-peso_a-peso_b
14 peso_m-peso_a-peso_b-peso_m
15 peso_a-peso_b-peso_a-peso_b
16 peso_m-peso_a-peso_b-peso_a
17 peso_m-peso_m-peso_a-peso_b
18 peso_a-peso_b-peso_a-peso_b
```

```
19 peso_a-peso_b-peso_a-peso_b
20 peso_m-peso_m-peso_a-peso_b
21 peso_a-peso_b-peso_a-peso_b
22 peso_a-peso_b-peso_a-peso_b
23 peso_m-peso_m-peso_m-peso_m
24 peso_a-peso_b-peso_a-peso_b
25 peso_a-peso_b-peso_a-peso_b
26 peso_a-peso_b-peso_a-peso_b
27 peso_m-peso_a-peso_b-peso_a
28 peso_a-peso_b-peso_a-peso_b
29 peso_m-peso_m-peso_a-peso_b
30 peso_a-peso_b-peso_a-peso_b
31 peso_a-peso_b-peso_a-peso_b
32 peso_m-peso_b-peso_a-peso_b
33 peso_a-peso_b-peso_a-peso_b
34 peso_m-peso_m-peso_a-peso_b
35 peso_a-peso_b-peso_a-peso_b
36 peso_a-peso_b-peso_a-peso_b
37 peso_m-peso_a-peso_b-peso_m
38 peso_m-peso_m-peso_m-peso_m
39 peso_m-peso_m-peso_a-peso_b
40 peso_m-peso_m-peso_a-peso_b
41 peso_a-peso_b-peso_a-peso_b
42 peso_a-peso_b-peso_a-peso_b
43 peso_a-peso_b-peso_a-peso_b
44 peso_m-peso_a-peso_b-peso_m
45 peso_m-peso_m-peso_a-peso_b
46 peso_a-peso_b-peso_a-peso_b
47 peso_a-peso_b-peso_a-peso_b
48 peso_m-peso_m-peso_a-peso_b
49 peso_a-peso_b-peso_a-peso_b
50 peso_a-peso_b-peso_a-peso_b
```

Na Figura 11 é apresentado, de forma sequencial e percentual, o caminho mais provável. O resultado aponta uma primeira safra com produtividade alta (48%), baixa (0%) e média (52%). A segunda safra com produtividade alta (12%), baixa (50%) e média (38%). A terceira safra com produtividade alta (80%), baixa (12%) e média (8%).

A quarta safra com produtividade alta (6%), baixa (80%) e média (14%).

Figura 11 – Dados simulados, nove variáveis



Fonte: Autor (2019)

O gráfico com a sequência de dados (50 observações) e distribuição percentual permite avaliar a variabilidade espacial da produtividade dentro da área de estudo e também ao longo do tempo.

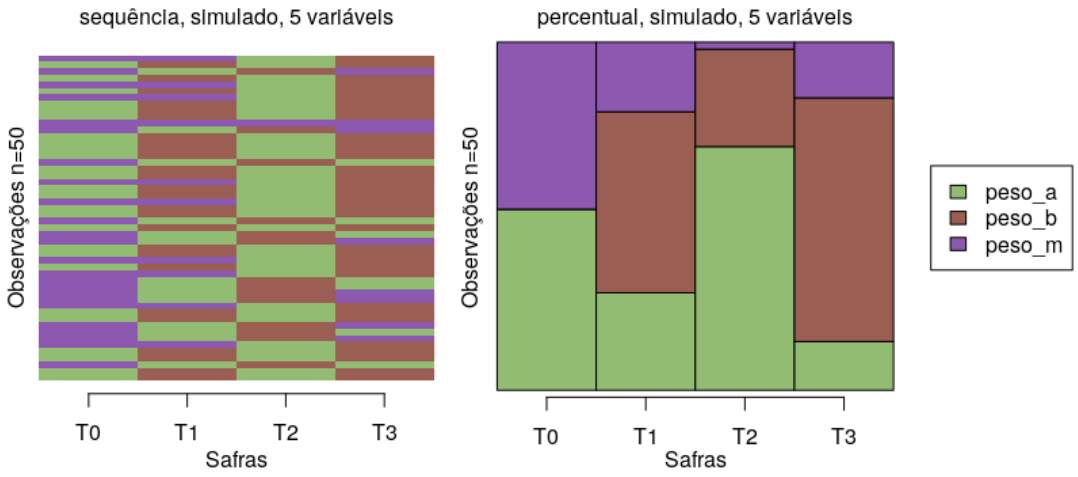
4.3 Comparação de modelos

Entre os modelos com dados simulados, conforme o *Bayesian Information Criterion* (BIC), o modelo com cinco variáveis (1476.884) é melhor, em seguida o modelo com seis variáveis (1587.394) e por último o modelo com nove variáveis (2564.739).

A Figura 12, com dados simulados, cinco variáveis e dados de solo, aponta o resultado de uma primeira safra com produtividade alta (52%) e média (48%). A segunda safra com produtividade alta (28%), baixa (52%) e média (20%). A terceira safra com produtividade alta (70%), baixa (28%) e média (2%). A quarta safra com produtividade alta (14%), baixa (70%) e média (16%).

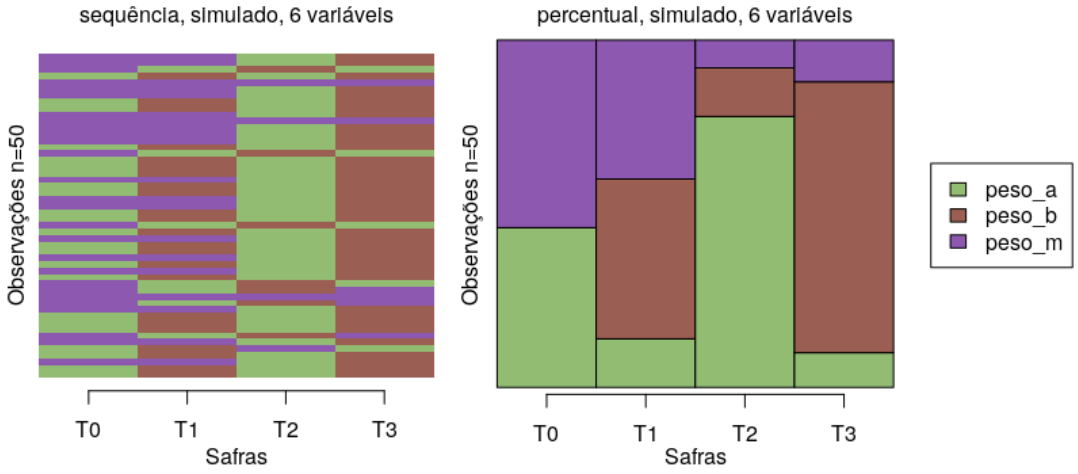
A Figura 13, com dados simulados, seis variáveis e dados meteorológicos, aponta o resultado de uma primeira safra com produtividade alta (46%) e média (54%). A segunda safra com produtividade alta (14%), baixa (46%) e média (40%). A terceira safra com produtividade alta (78%), baixa (14%) e média (8%). A quarta safra com produtividade alta (10%), baixa (78%) e média (12%).

Figura 12 – Dados simulados, cinco variáveis, dados de solo



Fonte: Autor (2019)

Figura 13 – Caminho mais provável com dados simulados, seis variáveis e dados meteorológicos



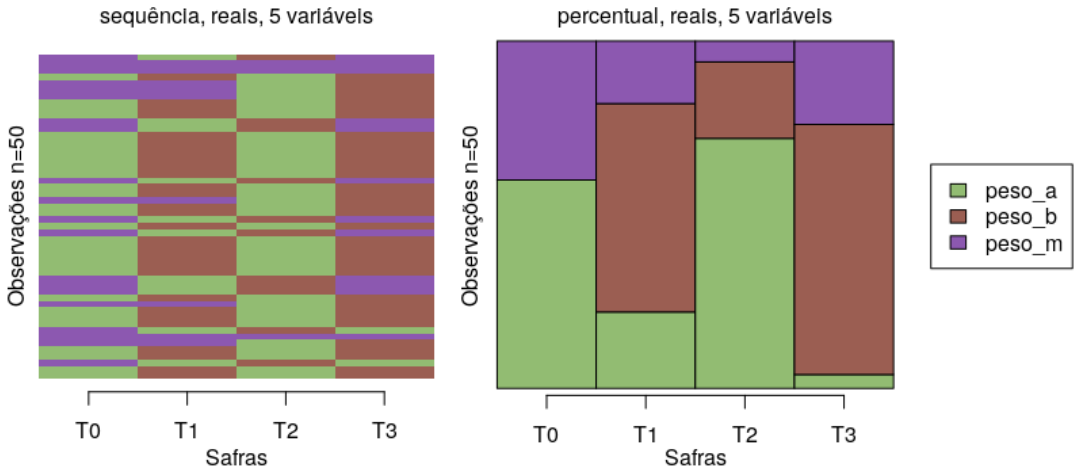
Fonte: Autor (2019)

Entre os modelos com dados reais, conforme o *BIC*, o modelo com seis variáveis (1207.326) é melhor, em seguida o modelo com cinco variáveis (1398.319) e por fim o modelo com nove variáveis (2090.064). De forma geral, os modelos com mais variáveis apresentaram o pior valor *BIC* quando comparados aos modelos com menos variáveis.

A Figura 14, com cinco variáveis e dados relativos ao solo, aponta o resultado de uma primeira safra com produtividade alta (52%) e média (48%). A segunda safra com produtividade alta (28%), baixa (52%) e média (20%). A terceira safra com produtividade

alta (70%), baixa (28%) e média (2%). A quarta safra com produtividade alta (14%), baixa (70%) e média (16%).

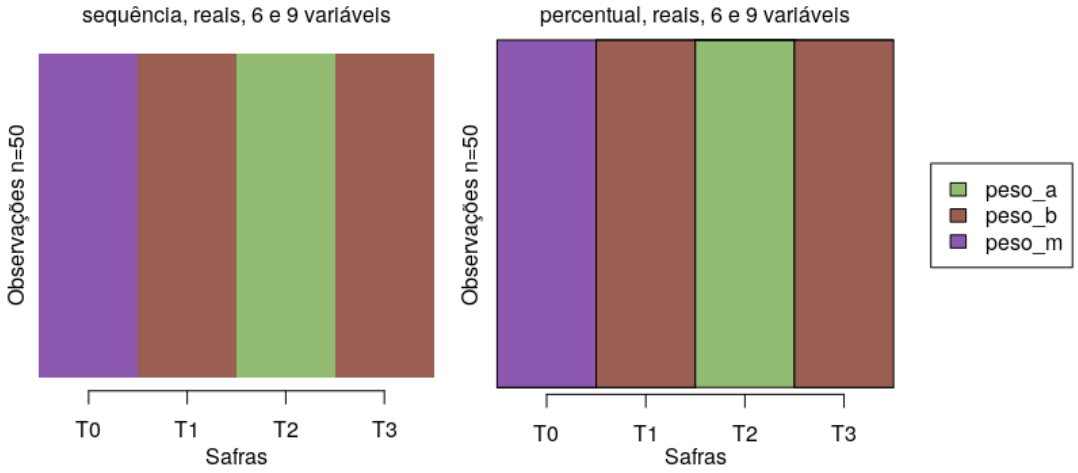
Figura 14 – Cinco variáveis, dados de solo



Fonte: Autor (2019)

A Figura 15, representa os modelos com nove variáveis (modelo completo) e com seis (dados meteorológicos). Não apresenta variabilidade entre os pontos observados, somente a variação entre as safras. Os dois modelos utilizam os dados meteorológicos, onde, por falta de dados detalhados, foi feita uma repetição para cada uma das observações.

Figura 15 – Nove variáveis e seis variáveis, dados meteorológicos

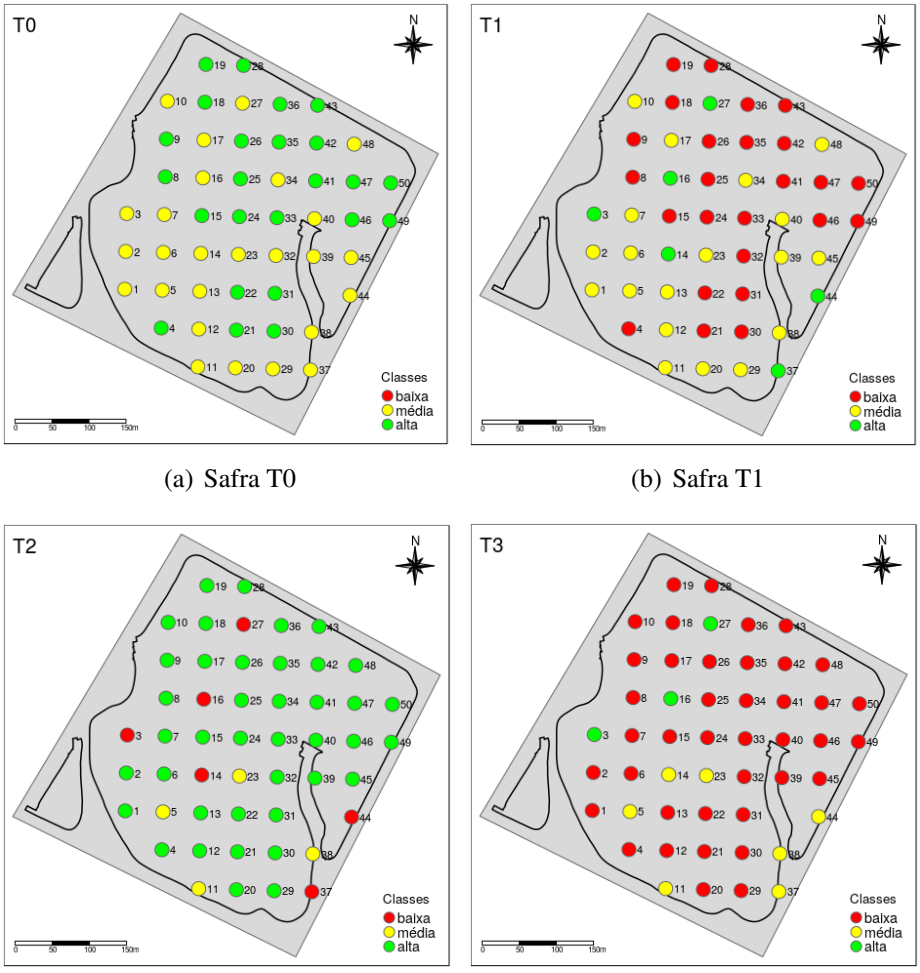


Fonte: Autor (2019)

4.4 Espacialização dos dados

A partir da sequência de dados, mostrada no Código 1, foi possível criar os mapas para cada uma das safras. Nas Figuras 16(a), 16(b), 16(c) e 16(d), é apresentada a produtividade predita para quatro safras (T0, T1, T2 e T3), em cada um dos pontos observados, utilizando o modelo completo (nove variáveis) e dados simulados. Este tipo de mapa permite identificar o local onde foram coletados os dados de peso, amostras de solo e resistência à penetração.

Figura 16 – Mapa de Pontos da Produtividade Simulada da Soja



(a) Safra T0 (b) Safra T1

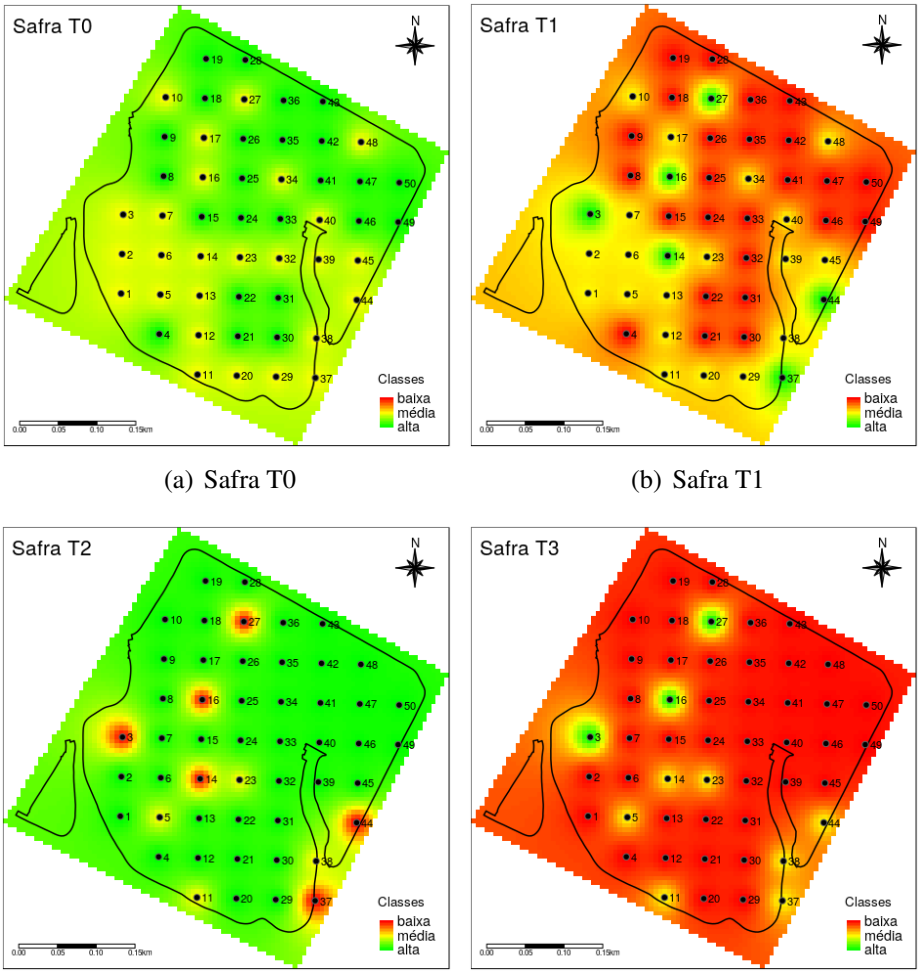
(c) Safra T2 (d) Safra T3

Fonte: Autor (2019)

Nas Figuras 17(a), 17(b), 17(c) e 17(d), é apresentada a produtividade predita interpolada (IDW) para quatro safras (T0, T1, T2 e T3), utilizando o modelo completo (nove variáveis) e dados simulados. Este tipo de mapa, além de identificar o local onde

foram coletados os dados de peso, amostras de solo e resistência à penetração, permite obter uma estimativa com valores intermediários entre dois ou mais pontos existentes.

Figura 17 – Mapa Interpolado da Produtividade Simulada da Soja



(a) Safra T0 (b) Safra T1

(c) Safra T2 (d) Safra T3

Fonte: Autor (2019)

O conjunto de mapas apresentado pode ser utilizado como subsídio para a tomada de decisão. Via simulação pode-se definir o percentual de cada classe para cada variável, projetando diferentes cenários desejados. Os mapas podem ser utilizados como entrada de dados em um SIG ou importados em algum sistema de agricultura de precisão que esteja acoplado ou embarcado em implementos agrícolas, para a aplicação de insumos conforme a especificação do mapa. Assim como os gráficos apresentados anteriormente, a sequência de mapas também permite apreciar a variabilidade espacial da produtividade dentro da área de estudo e também ao longo do tempo.

5 CONCLUSÃO

5.1 Discussão dos resultados obtidos

A variabilidade espacial da produtividade é uma questão complexa e várias abordagens têm sido utilizadas para tentar resolver ou atenuar seus efeitos no resultado final de uma ou de um conjunto de safras, bem como, em locais específicos dentro de uma área. Os estudos comumente realizados focam somente na relação entre as variáveis em um determinado instante e deixam de lado a evolução do processo ao longo do tempo. Os modelos ocultos de Markov estabelecem um relacionamento probabilístico entre os estados ocultos e as observações e o processo evolui no tempo por meio das transições entre seus estados, as quais são responsáveis pela emissão dos observáveis. Como discutido na revisão bibliográfica, no modelo proposto, a única exigência é que o estado futuro dependa apenas do estado presente e não dos estados passados, sem a necessidade de atender suposições mais rigorosas como em outros tipos de modelo, dependendo basicamente da qualidade na discretização dos dados.

É necessário que a discretização dos dados seja definida ou validada por especialistas, pois tanto as categorias definidas na bibliografia como a discretização por frequências podem não ser adequadas para a área de estudo em questão. A discretização por frequências deve ser utilizada somente na falta de outro critério mais especializado. Em caso de falta de dados o uso de repetição deve ser evitado, usando-se outros métodos de estimação de dados para complementar o conjunto faltante. Outra abordagem possível, quando existe insuficiência de dados para uma abordagem frequentista, é a utilização da abordagem Bayesiana, em que as probabilidade do modelo estão associadas à crença de um especialista sobre a relação entre as variáveis. Desde que o conjunto de crenças possua as mesmas propriedades de uma função de probabilidade, o modelo pode ser usado exatamente da mesma forma.

O modelo mostrou-se adequado para predizer a produtividade ao longo das safras, mas a estimativa da variabilidade dentro de uma determinada área é mais sensível à disponibilidade e discretização dos dados de entrada.

Para que o resultado do modelo seja confiável é necessário um conjunto maior de dados. Em um conjunto pequeno de dados, como utilizado nesse trabalho, as probabilidades podem ser estimadas de forma imprecisa. Na matriz de transição apresentada no trabalho, existe probabilidade com valor zero, o que talvez não ocorresse

se houvesse dados disponíveis de outras safras.

As matrizes de transição e emissão conseguem apresentar dados primários interessantes sobre o comportamento das safras no tempo e em relação as variáveis observáveis. O cálculo do caminho mais provável permite que se tenha a predição probabilística de forma rápida e em formato inteligível, o qual foi utilizado para a geração de mapas.

A organização e análise dos dados sequenciais se mostrou uma importante ferramenta para o entendimento do comportamento das variáveis antes de serem utilizadas no modelo. Essa inspeção e análise permitem a identificação de algum problema na preparação dos dados.

A organização dos dados meteorológicos na forma de séries temporais é uma forma prática e adequada para a manipulação deste tipo de dado. Essa organização permite a construção de subconjuntos temporais e provê uma série de funções específicas para o estudo dos dados ao longo do tempo.

Os softwares utilizados – planilha *LibreOficce Calc*, *QGIS* e a linguagem *R* – atenderam de forma plena as necessidades do trabalho, desde a organização das sequencias de dados, construção do modelos de Markov e finalmente apresentação dos resultados em forma gráfica e em forma de mapas.

5.2 Trabalhos Futuros

Ao longo do trabalho foram percebidas lacunas que podem ser desenvolvidas a partir deste trabalho. A seguir, discute-se algumas dessas questões.

Pode-se acrescentar dados de outras safras e novas variáveis. São necessários dados que agreguem informação qualificada ao modelo, principalmente os dados meteorológicos, os quais foram utilizados de forma genérica para os 50 pontos de observação.

Seria interessante a implementação de uma base de dados simulados para culturas de interesse. Essa base poderia ser composta, em parte, por dados reais e o restante da série completada com dados simulados.

A adição de variáveis topográficas como declividade, curvatura e acumulação do fluxo de água, dentre outras, em combinação com informações do solo, poderiam ser úteis para explicar a variabilidade da produtividade dentro da área e informar melhor sobre o balanço hídrico em cada ponto.

A resposta espectral a partir de dados de sensoriamento remoto pode ser utilizada para acompanhar os vários estágios de desenvolvimento da cultura em estudo. Os dados de sensoriamento remoto são mais fáceis de ser coletados, pois não dependem de trabalho de campo, têm uma boa resolução espacial e temporal, com a vantagem de já estarem armazenados em forma de matrizes de dados.

Os fenômenos que ocorrem dentro de cada fase fenológica são elementos importantes para o entendimento da variabilidade na produtividade, logo, pode-se estender o modelo, para que o tempo seja dividido em fases de uma safra. Para um modelo desse nível é importante a coleta de dados que façam referência a cada uma das fases de interesse. As modificações no sistema não seriam extensas, apenas o número de variáveis aumentaria para dar conta das diversas fases do período de safra. Os resultados, nesse caso, poderiam ser muito mais próximos da realidade, visto que, a distribuição da precipitação nas diversas fases da safra – e não apenas uma média final, sem a perspectiva temporal – com certeza permitiria a análise muito mais detalhada dos valores de produtividade e suas relações com as variáveis observadas.

A partir do conjunto de funções em *R*, desenvolvidas nesse trabalho, pode-se implementar um módulo que permita a integração com um SIG. Nesse caso, todo o processo de entrada de dados, modelagem e resultados seria feito dentro do SIG. O cálculo de probabilidades e caminho mais provável em Linguagem *R*. Outra opção é o desenvolvimento completo do módulo (*plugin*) utilizando um conjunto de ferramentas e bibliotecas de um SIG. Nesse caso tanto QGIS como GRASS GIS oferecem a infraestrutura necessária para esse tipo de implementação. A licença código aberto, execução multiplataforma, integração com outras linguagens de programação, *Application Programming Interface* (API) robusta e uma comunidade ativa de desenvolvedores são características que qualificam esses sistemas como candidatos para implementação da ferramenta.

REFERÊNCIAS

ACOSTA, J. J. B. *et al.* Variabilidade espacial da produtividade, perdas na colheita e lucratividade da cultura de soja. **Revista Agrogeoambiental**, IFSULDEMINAS, v. 10, p. 27–46, 03 2019. ISSN 2316-1817.

AKKER, J. van den; SOANE, B. Compactation. In: HILLEL, D. (Ed.). **Encyclopedia of Soils in the Environment**. Oxford: Elsevier, 2005. p. 285–293. ISBN 978-0-12-348530-4.

AL-OMRAN, A. *et al.* Spatial variability for some properties of the wastewater irrigated soils. **Journal of the Saudi Society of Agricultural Sciences**, v. 12, n. 2, p. 167 – 175, 2013. ISSN 1658-077X.

ALLEN, R. G. *et al.* **Crop evapotranspiration: guidelines for computing crop water requirements**. 1. ed. Roma: Food and Agriculture Organization of the United Nations, 1998. (FAO irrigation and drainage paper, 56). ISBN 9789251042199.

ALVARES, C. A. *et al.* Köppen's climate classification map for brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.

ASSIS, F. N.; ARRUDA, H. V. de; PEREIRA, A. R. **Aplicações de Estatística à Climatologia - Teoria e Prática**. 8th. ed. Pelotas: Editora Universitária UFPEL, 1996. 161 p.

BERQUO, E. S.; SOUZA, J. M. P. de; GOTLIEB, S. L. D. **Bioestatística**. 1ª edição. ed. São Paulo: EPU, 1981. 360 p.

BIVAND, R. S.; PEBESMA, E.; GóMEZ-RUBIO, V. **Applied Spatial Data Analysis with R**. 1. ed. New York, NY: Springer, 2008. 374 p.

BOLANO, D.; BERCHTOLD, A.; RITSCHARD, G. A discussion on hidden markov models for life course data. In: RITSCHARD, G.; STUDER, M. (Ed.). **Proceedings of the International Conference on Sequence Analysis and Related Methods**. Lausanne: NCCR LIVES, 2016. p. 241–260.

BURROUGH, P. **Principles of Geographical Information Systems for Land Resources Assessment**. 1. ed. Oxford: Clarendon Press, 1986. 194 p. (Monographs on Soil and Resources Survey). ISBN 9780198545637.

CAI, R. *et al.* Using a climate index to measure crop yield response. **Journal of Agricultural and Applied Economics**, v. 45, n. 4, p. 18, 2013.

CHRISTOFOLETTI, A. **Modelagem de Sistemas Ambientais**. 1. ed. São Paulo, SP, Brasil: Blucher, 1999. ISBN 9788521201779.

COMPANHIA NACIONAL DE ABASTECIMENTO. **A produtividade da soja: análises e perspectivas**. Brasília, DF, Brasil, 2017.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da Safra Brasileira de Grãos**. Brasília, DF, Brasil, 2018.

DINIZ, J. A. F. **Geografia da Agricultura**. 1. ed. São Paulo: Difel, 1984. 278 p.

DOORENBOS, J.; KASSAM, A. H. **Yield response to water**. 1. ed. Rome: Food and Agriculture Organization of the United Nations, 1979. 193 p. (FAO irrigation and drainage paper).

EASTMAN, J. R. **IDRISI GIS Analysis in TerrSet**. 2018. Disponível em: https://clarklabs.org/terrset/idrisi-gis/. Acesso em: 20 mar. 2018.

EASTMAN, J. R. **TerrSet Geospatial Monitoring and Modeling Software**. 2018. Disponível em: https://clarklabs.org/terrset/. Acesso em: 29 mar. 2018.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Cultivo do Algodão Irrigado**. 2014. Disponível em: https://www.spo.cnptia.embrapa.br/home. Acesso em: 10 jan. 2019.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Relatório solicitado pela Aprosoja e elaborado pela Embrapa Agrossilvipastoril - Março de 2016**. 2019. Disponível em: http://www.aprosoja.com.br/storage/site/files/comunicacao/arquivos/Situacao_Lavouras_Embrapa.pdf. Acesso em: 25 jan. 2019.

EMPRESA DE ASSISTÊNCIA TÉCNICA E EXTENSÃO RURAL. **Estimativa de Verão Safra 2018 2019**. 2019. Disponível em: http://www.emater.tche.br/site/arquivos_pdf/safra/safraTabela_27032019.pdf. Acesso em: 15 mar. 2019.

ESRI. **ESRI Shapefile Technical Description**. 1998. Disponível em: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf. Acesso em: 4 mai. 2018.

ESRI. **About ArcGIS Mapping & Analytics Plataform**. 2018. Disponível em: https://www.esri.com/en-us/arcgis/about-arcgis/overview. Acesso em: 29 jan. 2018.

FALKER. **Parâmentros para avaliação da resistência à penetração**. Porto Alegre, RS, Brasil, 2009.

FALKER. **Falker - Inovando a Agricultura**. 2018. Disponível em: http://falker.com.br/. Acesso em: 15 jan. 2018.

FINK, G. A. **Markov Models for Pattern Recognition - From Theory to Applications**. 8th. ed. Londres: Springer-Verlag, 2008. 248 p.

FORNEY, G. D. The viterbi algorithm. **Proceedings of IEEE**, v. 61, n. 3, p. 268–278, 1973.

FRANCHINI, J. C. *et al.* **Variabilidade espacial e temporal da produção de soja no Paraná e definição de ambientes de produção**. 1. ed. Londrina: Embrapa Soja, 2016. (Documentos). ISSN 2176-2937.

GEOTIFF. **OSGEO/libgeotiff**. 2018. Disponível em: https://github.com/OSGeo/libgeotiff. Acesso em: 4 mai. 2018.

GERARDI, L. H. de O.; SILVA, B. C. M. N. **Quantificação em Geografia**. São Paulo, SP: DIFEL, 1981. 161 p.

GHMM. **The General Hidden Markov Model library**. 2013. Disponível em: http://ghmm.sourceforge.net. Acesso em: 10 jan. 2019.

GREGO, C. R.; OLIVEIRA, R. P. de; VIEIRA, S. R. Geoestatística aplicada a agricultura de precisão. In: BERNARDI, A. C. de C. *et al.* (Ed.). **Agricultura de Precisão - Resultados de um Novo Olhar**. 2. ed. Brasília, DF: EMBRAPA, 2014. v. 1, p. 74–83.

GUEDES-FILHO, O. **Variabilidade espacial e temporal de mapas de colheita e atributos do solo em um sistema de semeadura direta**. 114 p. Dissertação (Mestrado) — Instituto Agronômico, Campinas, São Paulo, 2009.

HELSKE, J.; HELSKE, S. **seqHMM: Mixture hidden Markov models for social sequence data and other multivariate, multichannel categorical time series**. [S.l.], 2019. R package version 1.0.14. Disponível em: https://www.rdocumentation.org/packages/seqHMM/versions/1.0.14.

HELSKE, S.; HELSKE, J. Mixture hidden Markov models for sequence data: The seqHMM package in R. **Journal of Statistical Software**, v. 88, n. 3, p. 1–32, 2019.

HIMMELMANN, L. **Package HMM**. Vienna, Austria, 2015. Disponível em: https://cran.r-project.org/web/packages/HMM/HMM.pdf. Acesso em: 15 jan. 2019.

INSTITUTO NACIONAL DE METEOROLOGIA. **INMET - Instituto Nacional de Meteorologia**. 2019. Disponível em: http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep. Acesso em: 25 ago. 2018.

INSTITUTO NACIONAL DE METEOROLOGIA. **INMET - Instituto Nacional de Meteorologia**. 2019. Disponível em: http://www.inmet.gov.br/portal/. Acesso em: 25 ago. 2018.

INSTITUTO NACIONAL DE METEOROLOGIA. **SISDAGRO**. 2019. Disponível em: http://sisdagro.inmet.gov.br/sisdagro/app/index. Acesso em: 25 mar. 2018.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 2018. Disponível em: https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf. Acesso em: 12 mar. 2019.

KING, G. **Unifying Political Methodology: The Likelihood Theory of Statistical Inference**. 1. ed. [S.l.]: University of Michigan Press, 1998. 288 p. (Political Analysis). ISBN 9780472085545.

KIRSHNER, S. **Multi-Variate Non-homogeneous Hidden Markov Model**. 2019. Disponível em: http://www.sergeykirshner.com/software/mvnhmm. Acesso em: 25 jan. 2019.

KONOPATZKI, M. R. S. *et al.* Spatial variability of yield and other parameters associated with pear trees. **Engenharia Agrícola**, v. 32, p. 381–392, 04 2012. ISSN 0100-6916.

KRAVCHENKO, A. N.; BULLOCK, D. G. Correlation of corn and soybean grain yield with topography and soil properties. **Agronomy Journal**, American Society of Agronomy, v. 92, n. 1, p. 75–83, 1 2000. ISSN 0002-1962.

MACEDO, W. **Levantamento de reconhecimento dos solos do município de Bagé**. 1. ed. Brasília, DF: EMBRAPA - Departamento de Difusão Tecnológica, 1984. 69 p. EMBRAPA-UEPAE de Bagé.

MACHADO, P. L. O. de A. **Compactação do solo e crescimento de plantas: como identificar, evitar e remediar**. 1. ed. Rio de Janeiro, RJ: Embrapa Solos, 2003. (Documentos, 56). ISSN 1517-2627.

MANTOVANI, E. C. Compactação do solo. **Informe Agropecuário**, Empresa de Pesquisa Agropecuária de Minas Gerais - EPA, v. 13, n. 147, p. 52–55, 1987. ISSN 010-3364.

MATIS, J. *et al.* A markov chain approach to crop yield forecasting. **Agricultural Systems**, v. 18, n. 3, p. 171 – 187, 1985. ISSN 0308-521X. Disponível em: http://www.sciencedirect.com/science/article/pii/0308521X85900307.

MATIS, J. H.; BIRKETT, T.; BOUDREAUX, D. An application of the markov chain approach to forecasting cotton yields from surveys. **Agricultural Systems**, v. 29, n. 4, p. 357–370, 1989.

MATTIONI, N. M.; SCHUCH, L. O. B.; VILLELA, F. A. Variabilidade espacial da produtividade e da qualidade das sementes de soja em um campo de produção. **Revista Brasileira de Sementes**, scielo, v. 33, p. 608 – 615, 00 2011. ISSN 0101-3122.

MILLER, M. P.; SINGER, M. J.; NIELSEN, D. R. Spatial variability of wheat yield and soil properties on complex hill. **Soil Science Society of American Journal**, v. 52, p. 1133–1141, 12 1998. ISSN 0361-5995.

MINISTÉRIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO. **Indicadores Gerais Agrostat**. 2019. Disponível em: http://indicadores.agricultura.gov.br/agrostat/index.htm. Acesso em: 10 jan. 2019.

MURPHY, K. P. **Machine Learning - A Probabilistic Perspective**. 1. ed. Cambridge, MA: MIT Press, 2012. 1067 p.

NETELER, M.; MITASOVA, H. **Open Source GIS: A GRASS GIS Approach**. 1. ed. New York: Springer, 2008. 428 p. ISBN 978-0387357676.

OPEN SOURCE GEOSPATIAL FOUNDATION. **The Open Source Geospatial Foundation Project**. 2018. Disponível em: http://www.osgeo.org. Acesso em: 25 mar. 2018.

PAEGELOW, M.; OLMEDO, M. T. C. **Modelling Environmental Dynamics - Advances in Geomatic Solutions**. 1. ed. [S.l.]: Springer-Verlag Berlin Heidelberg, 2008. (Environmental Science). ISBN 978-3-540-68489-3.

POISSON-CAILLAULT, E.; TERNYNCK, P. **uHMM: Construct an Unsupervised Hidden Markov Model**. Vienna, Austria, 2019. Disponível em: https://cran.r-project.org/web/packages/uHMM/index.html. Acesso em: 15 jan. 2018.

QGIS. **QGIS Geographic Information System**. 2019. Disponível em: http://qgis.osgeo.org. Acesso em: 25 mar. 2018.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. 5. ed. Vienna, Austria: R Foundation for Statistical Computing, 2019. Disponível em: http://www.R-project.org. Acesso em: 15 jan. 2019.

RABINER, L. R.; JUANG, B. H. An introduction to hidden markov models. **IEEE ASSP Magazine**, v. 3, p. 4–16, 1986.

REIMANN, C. *et al.* **Statistical Data Analysis Explained: Applied Environmental Statistics with R**. 1. ed. West Sussex, England: John Wiley & Sons, Ltd., 2008. 362 p. ISBN 978-0-470-98581-6.

ROBLES, B. *et al.* Methods to choose the best hidden markov model topology for improving maintenance policy. In: **International Conference of Modeling, Optimization and Simulation**. Bordeaux, France: HAL, 2012.

SAND, A. *et al.* Hmmlib: A c++ library for general hidden markov models exploiting modern cpus. **2010 Ninth International Workshop on Parallel and Distributed Methods in Verification, and Second International Workshop on High Performance Computational Systems Biology**, p. 126–134, 2010.

SANTI, A. L. *et al.* Análise de componentes principais de atributos químicos e físicos do solo limitantes à produtividade de grãos. **Pesquisa Agropecuária Brasileira**, v. 47, p. 1346–1357, 09 2012.

SANTOS, H. G. dos *et al.* **Sistema Brasileiro de Classificação de Solos**. 5. ed. Brasília, DF: EMBRAPA, 2018. 365 p. ISBN 978-85-7035-800-4.

SCHERER, M. K. *et al.* Pyemma 2: A software package for estimation, validation, and analysis of markov models. **Journal of Chemical Theory and Computation**, v. 11, n. 11, p. 5525–5542, 2015.

SENNE, M. *et al.* Emma: A software package for markov model building and analysis. **Journal of Chemical Theory and Computation**, v. 8, n. 7, p. 2223–2238, 2012.

SILVA Éder David Borges da. **Estimando a produtividade na cultura da soja**. 2019. Disponível em: http://www.pioneersementes.com.br/blog/46/estimandoa-produtividadenaculturadasoja. Acesso em: 5 mai. 2019.

SIQUEIRA, T. V. de. O ciclo da soja: desempenho da cultura da soja entre 1961 e 2003. **BNDES artigos**, n. 20, p. 127–222, 2004.

SOCIEDADE BRASILEIRA DE CIÊNCIA DO SOLO. **Manual de Adubação e Calagem para os estados do Rio Grande do Sul e Santa Catarina**. Porto Alegre, Rio Grande do Sul, Brasil, 2004. 400 p.

SPEDICATO, G. A. **Package markovchain**. Vienna, Austria, 2019. Disponível em: https://cran.r-project.org/web/packages/markovchain/markovchain.pdf. Acesso em: 15 jan. 2019.

SPEDICATO, G. A. *et al.* **The markovchain Package: A Package for Easily Handling Discrete Markov Chains in R**. 2014. 70 p. Disponível em: https://cran.r-project.org/web/packages/markovchain/vignettes/an_introduction_to_markovchain_package.pdf. Acesso em: 25 ago. 2019.

SPIEGEL, M. R.; STEPHENS, L. J. **Theory and Problems of Statistics**. 4th. ed. New York: McGRAW-HILL, 2008. 577 p.

STARA. **Manual de Instruções TOPPER 4500VT**. Não-Me-Toque, Rio Grande do Sul, Brasil, 2011.

TAHA, H. A. **Operations Research: An Introduction**. 8th. ed. Upper Saddle River, New Jersey: Pearson Prentice Hall, 2007. 840 p.

THE DOCUMENT FOUNDATION. **Calc LibreOffice Free Office Suite Fun Project Fantastic People**. 2019. Disponível em: https://www.libreoffice.org/discover/calc/. Acesso em: 25 mar. 2018.

THIRUNAVUKKARASU, M. **Stochastic modeling in agricultural production**. 209 p. Tese (Doutorado) — Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, Índia, 2015.

THORLEY, J. **mcmcr: Manipulate MCMC Samples**. Vienna, Austria, 2019. Disponível em: https://cran.r-project.org/web/packages/mcmcr/index.html. Acesso em: 15 jan. 2019.

THORNTHWAITE, C. W.; MATHER, J. R. **The water balance**. 1. ed. Centerton, NJ: Drexel institute of technology, 1955. VIII. 104 p. (Publications in Climatology, 1).

UNITED STATES DEPARTAMENT OF AGRICULTURE. **USDA ERS - Brazil**. 2019. Disponível em: https://www.ers.usda.gov/topics/international-markets-us-trade/countries-regions/brazil. Acesso em: 12 mar. 2019.

USOWICZ, B.; LIPIEC, J. Spatial variability of soil properties and cereal yield in a cultivated field on sandy soil. **Soil and Tillage Research**, v. 174, p. 241 – 250, 2017. ISSN 0167-1987.

VIEIRA, S. R.; GONZALEZ, A. P. Analysis of the spatial variability of crop yield and soil properties in small agricultural plots. **Bragantia**, scielo, v. 62, p. 127 – 138, 00 2003. ISSN 0006-8705.

VITERBI, A. J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. **IEEE Transactions on Information Theory**, v. 13, n. 2, p. 260–269, April 1967.

WOLKOWSKI, R.; LOWERY, B. A3367, **Soil Compactation: Causes, Concerns and Cures**. Madson, Wisconsin, US: Cooperative Extension Publishing, 2008. Disponível em: https://learningstore.uwex.edu/Assets/pdfs/A3367.pdf. Acesso em: 10 jan. 2019.

WREGE, M. S. *et al.* **Atlas climático da Região Sul do Brasil: Estados do Paraná, Santa Catarina e Rio Grande do Sul**. 1. ed. Pelotas: Embrapa Clima Temperado, 2012. 333 p. ISBN 978-85-7383-519-9.

APÊNDICE A – CÓDIGO R UTILIZADO PARA O MODELO OCULTO DE MARKOV

```
## PACOTES ##
library(TraMineR)
library(gmodels)
library(seqHMM)
library(markovchain)

## DADOS ##
df_peso <- read.table(file = "Peso.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_prec <- read.table(file = "Precipitacao.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_insol <- read.table(file = "Insolacao.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_rp30 <- read.table(file = "RP30.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_exce <- read.table(file = "Excedente.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_defi <- read.table(file = "Deficit.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_alti <- read.table(file = "Altitude.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_rp <- read.table(file = "RP.csv", header=TRUE,sep="\t")
df_morg <- read.table(file = "MateriaOrganica.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_fosf <- read.table(file = "Fosforo.csv",
  ↪ header=TRUE,sep="\t",dec=".")
df_pot <- read.table(file = "Potassio.csv",
  ↪ header=TRUE,sep="\t")

##  ESTADOS E SEQUÊNCIA DE DADOS ##
##Cria sequencia de dados com alfabeto, paleta de cores e
  ↪ rótulos
```

```
l_peso <- ifelse(df_peso < 1.95, "peso_b", ifelse(df_peso <
  ↪ 3.56, "peso_m","peso_a"))
peso.seq <- seqdef(l_peso, start=0, alphabet=c("peso_a",
  ↪ "peso_b", "peso_m"), cpal=c("green", "red", "yellow"),
  ↪ labels=c("alto", "baixo", "médio"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

l_prec <- ifelse(df_prec < 671.4, "prec_b", "prec_a")
prec.seq <- seqdef(l_prec, start=0,
  ↪ alphabet=c("prec_a","prec_b"), cpal=c("green", "red"),
  ↪ labels=c("alta", "baixa"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

l_insol <- ifelse(df_insol < 1226.9, "insol_b", "insol_a")
insol.seq <- seqdef(l_insol, start=0, alphabet=c("insol_a",
  ↪ "insol_b"), cpal=c("green", "red"), labels=c("alta",
  ↪ "baixa"), cnames = c("12-13","14-15","16-17","17-18"))

l_exce <- ifelse(df_exce < 488.545, "exce_b", "exce_a")
exce.seq <- seqdef(l_exce, start=0, alphabet=c("exce_a",
  ↪ "exce_b"), cpal=c("green", "red"), labels=c("alto",
  ↪ "baixo"), cnames = c("12-13","14-15","16-17","17-18"))

l_defi <- ifelse(df_defi < 164.16, "defi_b", "defi_a")
defi.seq <- seqdef(l_defi, start=0, alphabet=c("defi_a",
  ↪ "defi_b"), cpal=c("green", "red"), labels=c("alto",
  ↪ "baixo"), cnames = c("12-13","14-15","16-17","17-18"))

l_alti <- ifelse(df_alti < 243.55, "alti_b", "alti_a")
alti.seq <- seqdef(l_alti, start=0, alphabet=c("alti_a",
  ↪ "alti_b"), cpal=c("green", "red"), labels=c("alta",
  ↪ "baixa"), cnames = c("12-13","14-15","16-17","17-18"))

l_rp <- ifelse(df_rp == "rp_b", "rp_b", "rp_a")
```

```
rp.seq <- seqdef(df_rp, start=0, alphabet=c("rp_a", "rp_b"),
  ↪ cpal=c("red", "green"), labels=c("alta", "baixa"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

l_morg <- ifelse(df_morg < 2.5, "mo_b", "mo_m")
morg.seq <- seqdef(l_morg, start=0, alphabet=c("mo_b", "mo_m"),
  ↪ cpal=c("red", "yellow"), labels=c("baixa", "média"), cnames
  ↪ = c("12-13","14-15","16-17","17-18"))

l_fosf <- ifelse(df_fosf < 12, "fosf_b", "fosf_a")
fosf.seq <- seqdef(l_fosf, start=0, alphabet=c("fosf_a",
  ↪ "fosf_b"), cpal=c("green", "red"), labels=c("alto",
  ↪ "baixo"), cnames = c("12-13","14-15","16-17","17-18"))

l_pot <- ifelse(df_pot == "k_a", "k_a", ifelse(df_pot ==
  ↪ "k_m","k_m","k_ma"))
pot.seq <- seqdef(df_pot, start=0, alphabet=c("k_a", "k_m",
  ↪ "k_ma"), cpal=c("green", "yellow", "blue"), labels=c("alto",
  ↪ "médio", "muito alto"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

## GRÁFICOS ##
##dados empilhados
ssplot(list("Peso" = peso.seq, "Prec." = prec.seq, "Insol." =
  ↪ insol.seq, "RP" = rp.seq, "Exc." = exce.seq, "Defi." =
  ↪ defi.seq, "Alt." = alti.seq, "M. O." = morg.seq, "P" =
  ↪ fosf.seq, "K" = pot.seq), title = "Dados Empilhados",
  ↪ legend.prop = 0.4, cex.legend = 0.7)

##dados em sequência
ssplot(list(peso.seq, prec.seq, insol.seq, rp.seq, exce.seq,
  ↪ defi.seq, alti.seq, morg.seq, fosf.seq, pot.seq), type =
  ↪ "I",title = "Dados em Sequência", sortv = "from.start",
  ↪ sort.channel = 1, with.legend = TRUE, legend.prop = 0.4,
  ↪ cex.legend = 0.6, ylab = c("Peso", "Prec.", "Insol.", "RP",
  ↪ "Exc.", "Def.", "Alti.", "M. O.", "P", "K"))
```

```
##estados combinados
sc_data <- mc_to_sc_data(list(prec.seq, insol.seq, rp.seq,
 ↪  exce.seq, defi.seq, alti.seq, morg.seq, fosf.seq, pot.seq))
ssplot(sc_data, type = "d", ylab = "Proporção", yaxis = TRUE,
 ↪  xlab = "Safras", title = "Estados combinados", legend.prop =
 ↪  0.75, cex.legend = 0.7, ncol.legend =2)

##Gráficos por variável
##type=I sequência
##type=d empilhado
ssplot(peso.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Peso")
ssplot(prec.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Precipitação")
ssplot(insol.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Insolação")
ssplot(rp.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Resistência à Penetração")
ssplot(exce.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Excedente Hídrico mm")
ssplot(defi.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Deficit Hídrico mm")
ssplot(alti.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Altitude m")
ssplot(morg.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Matéria Orgânica")
ssplot(fosf.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Fósforo")
ssplot(pot.seq, type="d", xlab="Safras", ylab="Observações",
 ↪  title="Potássio")


##  PROBABILIDADE INICIAL  ##
##(alta, baixa, média)
v_prob_inicial <- c(0.25,0.25,0.5)
```

```
##  MATRIZ DE TRANSICAO - ESTADOS OCULTOS - 3x3 ##
mcfit <- markovchainFit(data=l_peso, parallel = TRUE, sanitize =
  ↪ TRUE)
m_trans <- as(mcfit$estimate, "matrix")
mcPeso <- new("markovchain", transitionMatrix = m_trans,
  ↪ name="MC Peso", states = c("peso_a", "peso_b", "peso_m"))

#Constrói o gráfico de transição
plot(mcPeso,edge.arrow.size = 0.2, vertex.size = 40)

## MATRIZES EMISSÃO - OBSERVADO - ##
#Tabulação cruzada entre oculto e observado
peso.prec <- CrossTable(l_peso,l_prec)
peso.insol <- CrossTable(l_peso,l_insol)
peso.rp <- CrossTable(l_peso,l_rp)
peso.exce <- CrossTable(l_peso,l_exce)
peso.defi <- CrossTable(l_peso,l_defi)
peso.alti <- CrossTable(l_peso,l_alti)
peso.morg <- CrossTable(l_peso,l_morg)
peso.fosf <- CrossTable(l_peso,l_fosf)
peso.pot <- CrossTable(l_peso,l_pot)

##Probabilidade Condicional p(A|B) = P(A,B)/P(B) - direto do
  ↪ crosstable prop.row
emiss.prec <- peso.prec$prop.row
emiss.insol <- peso.insol$prop.row
emiss.rp <- peso.rp$prop.row
emiss.exce <- peso.exce$prop.row
emiss.defi <- peso.defi$prop.row
emiss.alti <- peso.alti$prop.row
emiss.morg <- peso.morg$prop.row
emiss.fosf <- peso.fosf$prop.row
emiss.pot <- peso.pot$prop.row

## HMM ##
```

```
mc_hmm <- build_hmm(observations = list(prec.seq, insol.seq,
  ↪ rp.seq, exce.seq, defi.seq, alti.seq, morg.seq, fosf.seq,
  ↪ pot.seq), initial_probs = v_prob_inicial, transition_probs =
  ↪ m_trans, emission_probs = list(emiss.prec, emiss.insol,
  ↪ emiss.rp, emiss.exce, emiss.defi, emiss.alti, emiss.morg,
  ↪ emiss.fosf, emiss.pot ), channel_names = c("Prec.",
  ↪ "Insol.", "RP", "Exc.", "Defi", "Alti", "M.O.", "P", "K"))

##sumário do modelo
summary(mc_hmm)
print(mc_hmm)

##Gráfico dirigido do HMM com probabilidades de emissão
  ↪ combinadas
plot(mc_hmm, edge.curved = c(0, -1.3, 0.6, 0.8, -0.7, -0.7,
  ↪ 1.3), ncol.legend = 2, legend.prop = 0.5, vertex.label =
  ↪ "initial.probs", combined.slice.label = "estados com
  ↪ probabilidade < 0.01", loops=TRUE, combine.slices=0.01,
  ↪ cex.legend = 0.9)

##Probabilidade de emissão combinadas
sc_hmm <- mc_to_sc(mc_hmm)
print(sc_hmm)

prec_a/insol_a/rp_a/exce_b/defi_b/alti_a/mo_b/fosf_a/k_a
peso_a  0.000000, peso_b  0.000141, peso_m  0.001615
prec_a/insol_a/rp_a/exce_b/defi_b/alti_a/mo_b/fosf_a/k_m
peso_a  0.00e+00, peso_b  3.16e-05, peso_m  7.75e-04
prec_a/insol_a/rp_a/exce_b/defi_b/alti_a/mo_b/fosf_a/k_ma
peso_a  0.00e+00, peso_b  9.72e-06, peso_m  1.29e-04
prec_a/insol_a/rp_a/exce_b/defi_b/alti_a/mo_m/fosf_a/k_a
peso_a  0.000000, peso_b  0.000145, peso_m  0.001790
prec_a/insol_a/rp_a/exce_b/defi_b/alti_a/mo_m/fosf_b/k_a
peso_a  0.00e+00, peso_b  4.57e-05, peso_m  5.76e-04
prec_a/insol_a/rp_a/exce_b/defi_b/alti_b/mo_b/fosf_b/k_a
peso_a  0.00e+00, peso_b  5.09e-05, peso_m  5.20e-04
```

prec_a/insol_a/rp_a/exce_b/defi_b/alti_b/mo_m/fosf_b/k_m
peso_a 0.00e+00, peso_b 1.17e-05, peso_m 2.77e-04
prec_a/insol_a/rp_b/exce_b/defi_b/alti_a/mo_b/fosf_a/k_a
peso_a 0.000000, peso_b 0.000564, peso_m 0.008883
prec_a/insol_a/rp_b/exce_b/defi_b/alti_a/mo_b/fosf_b/k_a
peso_a 0.000000, peso_b 0.000178, peso_m 0.002861
prec_a/insol_a/rp_b/exce_b/defi_b/alti_a/mo_b/fosf_b/k_m
peso_a 0.00e+00, peso_b 3.99e-05, peso_m 1.37e-03
prec_a/insol_a/rp_b/exce_b/defi_b/alti_a/mo_m/fosf_a/k_a
peso_a 0.000000, peso_b 0.000579, peso_m 0.009843
prec_a/insol_a/rp_b/exce_b/defi_b/alti_a/mo_m/fosf_a/k_m
peso_a 0.00000, peso_b 0.00013, peso_m 0.00472
prec_a/insol_a/rp_b/exce_b/defi_b/alti_a/mo_m/fosf_b/k_a
peso_a 0.000000, peso_b 0.000183, peso_m 0.003170
prec_a/insol_a/rp_b/exce_b/defi_b/alti_b/mo_b/fosf_a/k_a
peso_a 0.000000, peso_b 0.000644, peso_m 0.008883
prec_a/insol_a/rp_b/exce_b/defi_b/alti_b/mo_b/fosf_b/k_ma
peso_a 0.000000, peso_b 0.000014, peso_m 0.000229
prec_a/insol_a/rp_b/exce_b/defi_b/alti_b/mo_m/fosf_a/k_a
peso_a 0.000000, peso_b 0.000662, peso_m 0.009843
prec_a/insol_a/rp_b/exce_b/defi_b/alti_b/mo_m/fosf_a/k_m
peso_a 0.000000, peso_b 0.000148, peso_m 0.004725
prec_a/insol_a/rp_b/exce_b/defi_b/alti_b/mo_m/fosf_b/k_m
peso_a 0.00e+00, peso_b 4.68e-05, peso_m 1.52e-03
prec_a/insol_a/rp_b/exce_b/defi_b/alti_b/mo_m/fosf_b/k_ma
peso_a 0.00e+00, peso_b 1.44e-05, peso_m, 2.54e-04
prec_a/insol_b/rp_a/exce_a/defi_b/alti_a/mo_b/fosf_a/k_a
peso_a 0.018360, peso_b 0.000000, peso_m 0.000424
prec_a/insol_b/rp_a/exce_a/defi_b/alti_a/mo_b/fosf_a/k_m
peso_a 0.008345, peso_b 0.000000, peso_m 0.000204
prec_a/insol_b/rp_a/exce_a/defi_b/alti_a/mo_b/fosf_a/k_ma
peso_a 3.34e-03, peso_b 0.00e+00, peso_m 3.39e-05
prec_a/insol_b/rp_a/exce_a/defi_b/alti_a/mo_b/fosf_b/k_a
peso_a 0.011684, peso_b 0.000000, peso_m 0.000137
prec_a/insol_b/rp_a/exce_a/defi_b/alti_a/mo_m/fosf_a/k_a
peso_a 0.02295, peso_b 0.00000, peso_m 0.00047

prec_a/insol_b/rp_a/exce_a/defi_b/alti_a/mo_m/fosf_b/k_a
peso_a 0.014605, peso_b 0.000000, peso_m 0.000151
prec_a/insol_b/rp_a/exce_a/defi_b/alti_a/mo_m/fosf_b/k_m
peso_a 6.64e-03, peso_b 0.00e+00, peso_m 7.27e-05
prec_a/insol_b/rp_b/exce_a/defi_b/alti_a/mo_b/fosf_a/k_a
peso_a, 0.09180, peso_b 0.00000, peso_m 0.00233
prec_a/insol_b/rp_b/exce_a/defi_b/alti_a/mo_b/fosf_bk_a
peso_a 0.058418, peso_b 0.000000, peso_m 0.000751
prec_a/insol_b/rp_b/exce_a/defi_b/alti_a/mo_b/fosf_b/k_m
peso_a 0.026554, peso_b 0.000000, peso_m 0.000361
prec_a/insol_b/rp_b/exce_a/defi_b/alti_a/mo_m/fosf_a/k_a
peso_a 0.11475, peso_b 0.00000, peso_m 0.00259
prec_a/insol_b/rp_b/exce_a/defi_b/alti_a/mo_m/fosf_a/k_m
peso_a 0.05216, peso_b 0.00000, peso_m 0.00124
prec_a/insol_b/rp_b/exce_a/defi_b/alti_a/mo_m/fosf_b/k_a
peso_a 0.073023, peso_b 0.000000, peso_m 0.000833
prec_a/insol_b/rp_b/exce_a/defi_b/alti_a/mo_m/fosf_b/k_m
peso_a 0.0332, peso_b 0.0000, peso_m 0.0004
prec_a/insol_b/rp_b/exce_a/defi_b/alti_b/mo_b/fosf_a/k_a
peso_a 0.03531, peso_b 0.00000, peso_m 0.00233
prec_a/insol_b/rp_b/exce_a/defi_b/alti_b/mo_b/fosf_b/k_ma
peso_a 4.09e-03, peso_b 0.00e+00, peso_m 6.01e-05
prec_a/insol_b/rp_b/exce_a/defi_b/alti_b/mo_m/fosf_a/k_a
peso_a 0.04413, peso_b 0.00000, peso_m 0.00259
prec_a/insol_b/rp_b/exce_a/defi_b/alti_b/mo_m/fosf_a/k_m
peso_a 0.02006, peso_b 0.00000, peso_m 0.00124
prec_a/insol_b/rp_b/exce_a/defi_b/alti_b/mo_m/fosf_b/k_m
peso_a 0.0128, peso_b 0.0000, peso_m 0.0004
prec_a/insol_b/rp_b/exce_a/defi_b/alti_b/mo_m/fosf_b/k_ma
peso_a 5.11e-03, peso_b 0.00e+00, peso_m 6.66e-05
prec_b/insol_b/rp_a/exce_b/defi_a/alti_a/mo_b/fosf_a/k_a
peso_a 4.28e-07, peso_b 1.53e-02, peso_m 2.10e-03
prec_b/insol_b/rp_a/exce_b/defi_a/alti_a/mo_b/fosf_a/k_m
peso_a 1.95e-07, peso_b 3.43e-03, peso_m 1.01e-03
prec_b/insol_b/rp_a/exce_b/defi_a/alti_a/mo_m/fosf_a/k_a
peso_a 5.35e-07, peso_b 1.57e-02, peso_m 2.33e-03

prec_b/insol_b/rp_a/exce_b/defi_a/alti_b/mo_b/fosf_a/k_ma
peso_a 2.99e-08, peso_b 1.20e-03, peso_m 1.68e-04
prec_b/insol_b/rp_a/exce_b/defi_a/alti_b/mo_b/fosf_b/k_a
peso_a 1.05e-07, peso_b 5.52e-03, peso_m 6.78e-04
prec_b/insol_b/rp_a/exce_b/defi_a/alti_b/mo_m/fosf_a/k_a
peso_a 2.06e-07, peso_b 1.79e-02, peso_m 2.33e-03
prec_b/insol_b/rp_a/exce_b/defi_a/alti_b/mo_m/fosf_b/k_a
peso_a 1.31e-07, peso_b 5.67e-03, peso_m 7.51e-04
prec_b/insol_b/rp_a/exce_b/defi_a/alti_b/mo_m/fosf_b/k_m
peso_a 5.96e-08, peso_b, 1.27e-03, peso_m 3.60e-04
prec_b/insol_b/rp_b/exce_b/defi_a/alti_a/mo_b/fosf_a/k_a
peso_a 2.14e-06, peso_b 6.12e-02, peso_m 1.16e-02
prec_b/insol_b/rp_b/exce_b/defi_a/alti_a/mo_b/fosf_b/k_a
peso_a 1.36e-06, peso_b 1.93e-02, peso_m 3.73e-03
prec_b/insol_b/rp_b/exce_b/defi_a/alti_a/mo_b/fosf_b/k_m
peso_a, 6.19e-07, peso_b 4.33e-03, peso_m 1.79e-03
prec_b/insol_b/rp_b/exce_b/defi_a/alti_a/mo_m/fosf_a/k_a
peso_a 2.68e-06, peso_b, 6.28e-02, peso_m 1.28e-02
prec_b/insol_b/rp_b/exce_b/defi_a/alti_a/mo_m/fosf_b/k_m
peso_a 7.74e-07, peso_b 4.45e-03, peso_m 1.98e-03
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_b/fosf_a/k_a
peso_a 8.24e-07, peso_b 6.99e-02, peso_m 1.16e-02
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_b/fosf_b/k_m
peso_a 2.38e-07, peso_b 4.95e-03, peso_m 1.79e-03
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_b/fosf_b/k_ma
peso_a 9.53e-08, peso_b 1.52e-03, peso_m 2.98e-04
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_m/fosf_a/k_a
peso_a 1.03e-06, peso_b 7.18e-02, peso_m 1.28e-02
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_m/fosf_a/k_m
peso_a 4.68e-07, peso_b 1.61e-02, peso_m 6.16e-03
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_m/fosf_b/k_a
peso_a 6.55e-07, peso_b 2.27e-02, peso_m 4.13e-03
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_m/fosf_b/k_m
peso_a 2.98e-07, peso_b 5.08e-03, peso_m 1.98e-03
prec_b/insol_b/rp_b/exce_b/defi_a/alti_b/mo_m/fosf_b/k_ma
peso_a 1.19e-07, peso_b 1.56e-03, peso_m 3.30e-04

```
## CAMINHO MAIS PROVÁVEL -  VITERBI ##
mpp <- hidden_paths(mc_hmm)

##Traçar caminhos mais provável para os 50 pontos - sequencial e
  ↪ percentual
graf_mpp_stac<-ssp(mpp, type = "I", tlim = 50:1, xlab="Safras",
  ↪ ylab="Observações n=50", title="sequência, reais, 9
  ↪ variáveis", title.n=FALSE, with.legend=FALSE)
graf_mpp_perc<-ssp(mpp, type = "d", tlim = 50:1, xlab="Safras",
  ↪ ylab="Observações n=50", title="percentual, reais, 9
  ↪ variáveis", title.n=FALSE)
gridplot(list(graf_mpp_stac, graf_mpp_perc), with.legend =
  ↪ FALSE, ncol = 2, col.prop = c(0.4, 0.6))
```

APÊNDICE B – CÓDIGO R UTILIZADO NA SIMULAÇÃO

```
## SIMULACÃO ##
#9 variáveis
#Peso e altitude utilizados da base original
#probabilidades iniciais e matriz de transição da base original
#cria sequências randômicas, com probabilidade de estados para
  ↪ as demais variáveis
sim_seq_prec <- seqgen(50, 4, alphabet=c("prec_a", "prec_b"),
  ↪ p=c(0.5, 0.5))
sim_seq_prec1 <- seqdef(sim_seq_prec, start=0,
  ↪ alphabet=c("prec_a", "prec_b"), cpal=c("green", "red"),
  ↪ labels=c("alto", "baixo"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

sim_seq_insol <- seqgen(50, 4, alphabet=c("insol_a", "insol_b"),
  ↪ p=c(0.25, 0.75))
sim_seq_insol1 <- seqdef(sim_seq_insol, start=0,
  ↪ alphabet=c("insol_a", "insol_b"), cpal=c("green", "red"),
  ↪ labels=c("alto", "baixo"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

sim_seq_rp <- seqgen(50, 4, alphabet=c("rp_a", "rp_b"),
  ↪ p=c(0.18, 0.82))
sim_seq_rp1 <- seqdef(sim_seq_rp, start=0, alphabet=c("rp_a",
  ↪ "rp_b"), cpal=c("green", "red"), labels=c("alta", "baixa"),
  ↪ cnames = c("12-13","14-15","16-17","17-18"))

sim_seq_exce <- seqgen(50, 4, alphabet=c("exce_a", "exce_b"),
  ↪ p=c(0.25, 0.75))
sim_seq_exce1 <- seqdef(sim_seq_exce, start=0,
  ↪ alphabet=c("exce_a", "exce_b"), cpal=c("green", "red"),
  ↪ labels=c("alto", "baixo"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))
```

```
sim_seq_defi <- seqgen(50, 4, alphabet=c("defi_a", "defi_b"),
  ↪ p=c(0.5, 0.5))
sim_seq_defi1 <- seqdef(sim_seq_defi, start=0,
  ↪ alphabet=c("defi_a", "defi_b"), cpal=c("green", "red"),
  ↪ labels=c("alto", "baixo"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

sim_seq_morg <- seqgen(50, 4, alphabet=c("mo_b", "mo_m"),
  ↪ p=c(0.54, 0.46))
sim_seq_morg1 <- seqdef(sim_seq_morg, start=0,
  ↪ alphabet=c("mo_b", "mo_m"), cpal=c("red", "yellow"),
  ↪ labels=c("baixa", "média"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

sim_seq_fosf <- seqgen(50, 4, alphabet=c("fosf_a", "fosf_b"),
  ↪ p=c(0.74, 0.26))
sim_seq_fosf1 <- seqdef(sim_seq_fosf, start=0,
  ↪ alphabet=c("fosf_a", "fosf_b"), cpal=c("green", "red"),
  ↪ labels=c("alto", "baixo"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

sim_seq_pot <- seqgen(50, 4, alphabet=c("k_a", "k_m", "k_ma"),
  ↪ p=c(0.70, 0.24, 0.6))
sim_seq_pot1 <- seqdef(sim_seq_pot, start=0, alphabet=c("k_a",
  ↪ "k_m", "k_ma"), cpal=c("green", "yellow", "blue"),
  ↪ labels=c("alto", "medio", "muito alto"), cnames =
  ↪ c("12-13","14-15","16-17","17-18"))

#Tabulacao cruzada entre oculto e observado
sim_peso.prec <- CrossTable(l_peso,sim_seq_prec)
sim_peso.insol <- CrossTable(l_peso,sim_seq_insol)
sim_peso.rp <- CrossTable(l_peso,sim_seq_rp)
sim_peso.exce <- CrossTable(l_peso,sim_seq_exce)
sim_peso.defi <- CrossTable(l_peso,sim_seq_defi)
sim_peso.morg <- CrossTable(l_peso,sim_seq_morg)
sim_peso.fosf <- CrossTable(l_peso,sim_seq_fosf)
```

```
sim_peso.pot <- CrossTable(l_peso,sim_seq_pot)

#Probabilidade Condicional p(A|B) = P(A,B)/P(B)
sim_emiss.prec <- sim_peso.prec$prop.row
sim_emiss.insol <- sim_peso.insol$prop.row
sim_emiss.rp <- sim_peso.rp$prop.row
sim_emiss.exce <- sim_peso.exce$prop.row
sim_emiss.defi <- sim_peso.defi$prop.row
sim_emiss.morg <- sim_peso.morg$prop.row
sim_emiss.fosf <- sim_peso.fosf$prop.row
sim_emiss.pot <- sim_peso.pot$prop.row

#construção do modelo
sim_mc_hmm <- build_hmm(observations = list(sim_seq_prec1,
  ↪ sim_seq_insol1, sim_seq_rp1, sim_seq_exce1, sim_seq_defi1,
  ↪ alti.seq, sim_seq_morg1, sim_seq_fosf1, sim_seq_pot1),
  ↪ initial_probs = v_prob_inicial, transition_probs = m_trans,
  ↪ emission_probs = list(sim_emiss.prec, sim_emiss.insol,
  ↪ sim_emiss.rp, sim_emiss.exce, sim_emiss.defi, emiss.alti,
  ↪ sim_emiss.morg, sim_emiss.fosf, sim_emiss.pot),
  ↪ channel_names = c("Prec.", "Insol.", "RP", "Exc.", "Def.",
  ↪ "Alt.", "M.O.", "P", "K"))

plot(sim_mc_hmm, edge.curved = c(0, -0.7, 0.6, 0.8, 0, -0.7, 0),
  ↪ ncol.legend = 2, legend.prop = 0.4, vertex.label =
  ↪ "initial.probs", combined.slice.label = "estados com
  ↪ probabilidade < 0.05")


#Computa os caminhos de estado oculto mais prováveis, conforme
  ↪ os dados e o modelo
sim_mpp <- hidden_paths(sim_mc_hmm)

##Traça caminhos ocultos para os 50 pontos
ssplot(sim_mpp, type = "I", tlim = 50:1, xlab="Safras")as
```

```
graf_mpp_stac<-ssp(sim_mpp, type = "I", tlim = 50:1,
  ↪  xlab="Safras", ylab="Observações")
graf_mpp_perc<-ssp(sim_mpp, type = "d", tlim = 50:1,
  ↪  xlab="Safras", ylab="Observações")
gridplot(list(graf_mpp_stac, graf_mpp_perc), with.legend =
  ↪  FALSE, ncol = 2, col.prop = c(0.5, 0.5))

summary(sim_mpp)
seqstatd(sim_mpp)
print(sim_mpp)


##Sequencia com o caminho mais provável para cada ponto de dados
  ↪  observados

1  peso_m-peso_m-peso_a-peso_b
2  peso_m-peso_m-peso_a-peso_b
3  peso_m-peso_a-peso_b-peso_a
4  peso_a-peso_b-peso_a-peso_b
5  peso_m-peso_m-peso_m-peso_m
6  peso_m-peso_m-peso_a-peso_b
7  peso_m-peso_m-peso_a-peso_b
8  peso_a-peso_b-peso_a-peso_b
9  peso_a-peso_b-peso_a-peso_b
10 peso_m-peso_m-peso_a-peso_b
11 peso_m-peso_m-peso_m-peso_m
12 peso_m-peso_m-peso_a-peso_b
13 peso_m-peso_m-peso_a-peso_b
14 peso_m-peso_a-peso_b-peso_m
15 peso_a-peso_b-peso_a-peso_b
16 peso_m-peso_a-peso_b-peso_a
17 peso_m-peso_m-peso_a-peso_b
18 peso_a-peso_b-peso_a-peso_b
19 peso_a-peso_b-peso_a-peso_b
20 peso_m-peso_m-peso_a-peso_b
21 peso_a-peso_b-peso_a-peso_b
```

```
22 peso_a-peso_b-peso_a-peso_b
23 peso_m-peso_m-peso_m-peso_m
24 peso_a-peso_b-peso_a-peso_b
25 peso_a-peso_b-peso_a-peso_b
26 peso_a-peso_b-peso_a-peso_b
27 peso_m-peso_a-peso_b-peso_a
28 peso_a-peso_b-peso_a-peso_b
29 peso_m-peso_m-peso_a-peso_b
30 peso_a-peso_b-peso_a-peso_b
31 peso_a-peso_b-peso_a-peso_b
32 peso_m-peso_b-peso_a-peso_b
33 peso_a-peso_b-peso_a-peso_b
34 peso_m-peso_m-peso_a-peso_b
35 peso_a-peso_b-peso_a-peso_b
36 peso_a-peso_b-peso_a-peso_b
37 peso_m-peso_a-peso_b-peso_m
38 peso_m-peso_m-peso_m-peso_m
39 peso_m-peso_m-peso_a-peso_b
40 peso_m-peso_m-peso_a-peso_b
41 peso_a-peso_b-peso_a-peso_b
42 peso_a-peso_b-peso_a-peso_b
43 peso_a-peso_b-peso_a-peso_b
44 peso_m-peso_a-peso_b-peso_m
45 peso_m-peso_m-peso_a-peso_b
46 peso_a-peso_b-peso_a-peso_b
47 peso_a-peso_b-peso_a-peso_b
48 peso_m-peso_m-peso_a-peso_b
49 peso_a-peso_b-peso_a-peso_b
50 peso_a-peso_b-peso_a-peso_b
```

APÊNDICE C – CÓDIGO R UTILIZADO PARA A GERAÇÃO DO MAPAS

```r
##Carrega bibliotecas
library(sf)
library(tmap)
library(readODS)
library(gstat)
library(sp)
library(raster)
library(rgdal)

##Importa coordenadas e feições geográficos
df_coord <- read_ods("MarkovMapas.ods", col_names = TRUE, sheet
  ↪ =1)
talhao.sf <- st_read("mask_talhao.shp")
masc.sf <- st_read("soja_mask_area.shp")
masc.r <- raster("soja_mask_area.tif")

##Converte símbolos em números
df_hmm_mpp_result <- ifelse(sim_mpp == "peso_a", 3,
  ↪ ifelse(sim_mpp == "peso_b", 1, ifelse(sim_mpp == "peso_m",
  ↪ 2, NA)))

##Une coordenadas e resultados
hmm_mpp_pontos <- cbind(df_coord, df_hmm_mpp_result)

##Cria o geobjeto
mpp_pontos.sf <- st_as_sf(hmm_mpp_pontos, coords = c("lat",
  ↪ "long"), crs = 4326)
mpp_pontos <- as(mpp_pontos.sf, 'Spatial')

##Cria o grid para a interpolação
grid <- as(masc.r, 'SpatialGridDataFrame')

##Altera modos de visualização do mapa
tmap_mode("plot")
```

```
#tmap_mode("view")

##Função que cria os mapas de pontos
def_mapa_pontos <- function (col, titulo){
  tm_shape(talhao.sf, unit = "m", unit.size=1000) +
    ↪ tm_polygons() + tm_shape(masc.sf) + tm_borders("black",
    ↪ lwd = 1.5) +
  tm_shape(mpp_pontos.sf) + tm_symbols(col=col, breaks =
    ↪ c(1,2,3, Inf), labels = c("baixa","média","alta"), palette
    ↪ = c("red","yellow","green"), title.col = "Classes") +
    ↪ tm_text("cod", just="left", xmod=.5, size = 0.7) +
  tm_compass(type = "8star", position = c("right", "top"), size
    ↪ = 3) +
  tm_scale_bar(text.size = 0.5, position = c("left", "bottom"),
    ↪ width = 0.20) +
  tm_layout(title=titulo, legend.title.size = 1,
    ↪ legend.text.size = 0.9, legend.position =
    ↪ c("right","bottom"), legend.bg.color = "white",
    ↪ legend.bg.alpha = 1)
}

def_mapa_pontos(col="T0", titulo="T0")
def_mapa_pontos(col="T1", titulo="T1")
def_mapa_pontos(col="T2", titulo="T2")
def_mapa_pontos(col="T3", titulo="T3")

##Função que cria os mapas interpolados raster
def_mapa_idw <- function (formula, titulo){
  mpp.idw <- gstat::idw(formula, mpp_pontos, newdata=grid,
    ↪ idp=2.0)
  r.mpp.idw  <- raster(mpp.idw)
  tm_shape(r.mpp.idw) + tm_raster(n=3,style = "cont", breaks =
    ↪ c(1,2,3), title="Classes", labels =
    ↪ c("baixa","média","alta"), palette =
    ↪ c("red","yellow","green")) + tm_shape(masc.sf) +
    ↪ tm_borders("black", lwd = 1.5) +
```

```
  tm_shape(mpp_pontos.sf, unit = "m", unit.size=1000) +
    ↪ tm_symbols(col="black", size = .2) + tm_text("cod",
    ↪ just="left", xmod=.3, size = 0.7)  +
  tm_compass(type = "8star", position = c("right", "top"), size
    ↪ = 3) +
  tm_scale_bar(text.size = 0.5, position = c("left", "bottom"),
    ↪ width = 0.20) +
  tm_layout(title=titulo, legend.title.size = 1,
    ↪ legend.text.size = 0.9, legend.position =
    ↪ c("right","bottom"), legend.bg.color = "white",
    ↪ legend.bg.alpha = 1)
}

def_mapa_idw(formula=T0~1, titulo="Safra T0")
def_mapa_idw(formula=T1~1, titulo="Safra T1")
def_mapa_idw(formula=T2~1, titulo="Safra T2")
def_mapa_idw(formula=T3~1, titulo="Safra T3")


##Função que exporta arquivos vetoriais para shapefile (.shp)
def_exporta_shp <- function (camada, nome){
  rgdal::writeOGR(camada, dsn = "/home/jean", layer = nome,
    ↪ driver = "ESRI Shapefile" )
}


def_exporta_shp(camada = mpp_pontos, nome = "mpp_export1")
def_exporta_shp(camada = mpp_idw, nome = "mpp_export")
```

APÊNDICE D – FUNÇÕES AUXILIARES

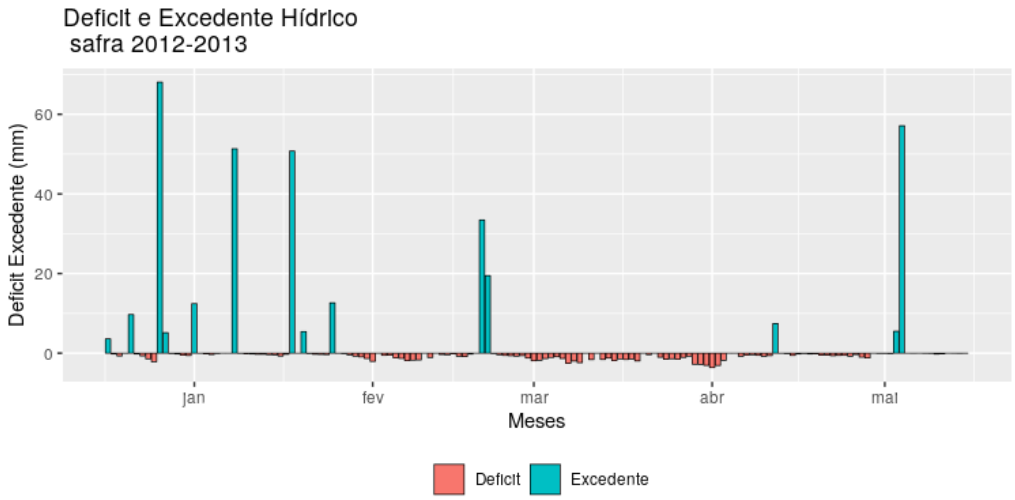
```
##  FREQUÊNCIAS DE CLASSES  ##
##Calcula os limites de classes de um dataset
def_classes <- function(n_classes, df_nome) {
  val_max <- max(df_nome, na.rm = TRUE)
  val_min <- min(df_nome, na.rm = TRUE)
  amp_total <- val_max - val_min
  amp_classe <- amp_total / n_classes
  classes <- vector()
  classes <- val_min + amp_classe
  for (i in 1:(n_classes -1)) {
    classes[i+1] <- classes[i] + amp_classe; classes
    }
  nome <- deparse(substitute(df_nome))
  message(nome, " - Limites de Classes")
  return(classes)
}
def_classes(2,df_prec)
def_classes(2,df_insol)
def_classes(2,df_exce)
def_classes(2,df_defi)
def_classes(2,df_alti)


##  MATRIZES DE EMISSÂO  ##
##Probabilidade Condicional p(A|B) = P(A,B)/ P(B) - Calculada
##pconj: probabilidade conjunta
##pmarg: probabilidade marginal
##nlin: numero de linhas da matriz
##ncol: numero de colunas da matriz
##pab: 1 = P(A|B), 2 = P(B|A)
def_prob_cond <- function(pconj, pmarg, nlin, ncol, pab) {
  m_pcond <- matrix(, nrow = nlin, ncol = ncol)
  for(col in 1:ncol) {
    for(row in 1:nlin) {
```

```
      if (pab == 1){
        m_pcond[row,col]<-pconj[row,col]/pmarg[row]
      }else if (pab == 2){
        m_pcond[row,col]<-pconj[row,col]/pmarg[col]
      }
    }
  }
return(m_pcond)
}

emiss1.prec <- def_prob_cond(pcj.prec, pmA_peso.prec, 3, 2, 1)
emiss1.insol <- def_prob_cond(pcj.insol, pmA_peso.insol, 3, 2,
 ↪ 1)
emiss1.rp30 <- def_prob_cond(pcj.rp30, pmA_peso.rp30, 3, 2, 1)
emiss1.exce <- def_prob_cond(pcj.exce, pmA_peso.exce, 3, 2, 1)
emiss1.defi <- def_prob_cond(pcj.defi, pmA_peso.defi, 3, 2, 1)
emiss1.alti <- def_prob_cond(pcj.alti, pmA_peso.alti, 3, 2, 1)

emiss2.prec <- def_prob_cond(pcj.prec, pmB_peso.prec, 3, 2, 2)
emiss2.insol <- def_prob_cond(pcj.insol, pmB_peso.insol, 3, 2,
 ↪ 2)
emiss2.rp30 <- def_prob_cond(pcj.rp30, pmB_peso.rp30, 3, 2, 2)
emiss2.exce <- def_prob_cond(pcj.exce, pmB_peso.exce, 3, 2, 2)
emiss2.defi <- def_prob_cond(pcj.defi, pmB_peso.defi, 3, 2, 2)
emiss2.alti <- def_prob_cond(pcj.alti, pmB_peso.alti, 3, 2, 2)


## BIC - Bayesian information criterion ##
##Dados reais 9, 5 e 6 variáveis
BIC(mc_hmm)
2090.064
BIC(mc5_hmm)
1398.319
BIC(mc6_hmm)
1207.326
```

```
##Dados simulados 9, 5 e 6 variáveis
BIC(sim_mc_hmm)
2564.739
BIC(sim5_mc_hmm)
1476.884
BIC(sim6_mc_hmm)
1587.394
```

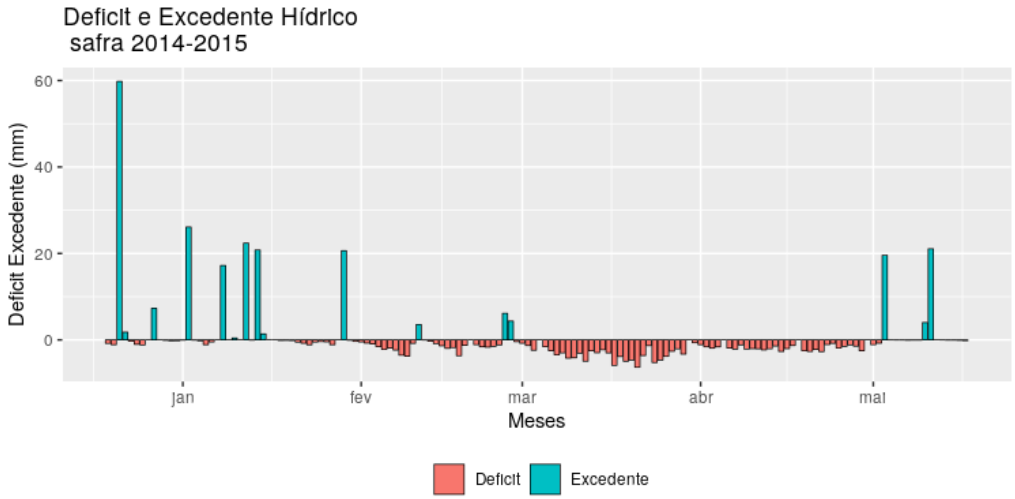
APÊNDICE E – GRÁFICOS DE DEFICIT E EXCEDENTE HÍDRICO

Figura 18 – Deficit e excedente hídrico na safra 2012-2013



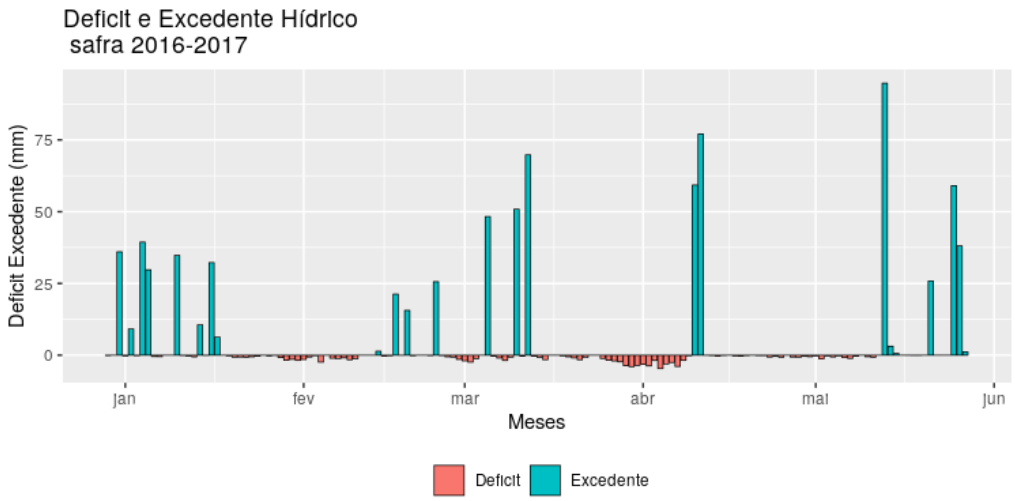
Fonte: autor, dados SISDAGRO-INMET

Figura 19 – Deficit e excedente hídrico na safra 2014-2015



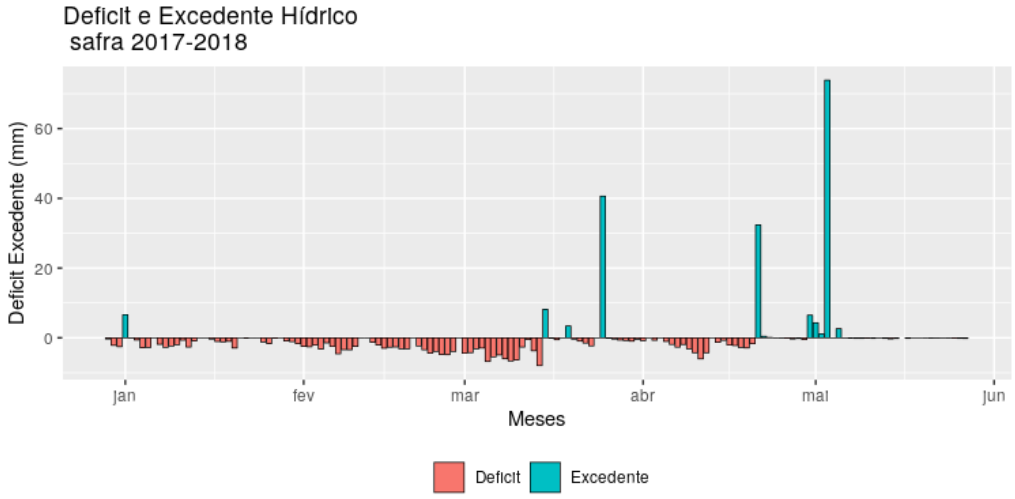
Fonte: autor, dados SISDAGRO-INMET

Figura 20 – Deficit e excedente deficit hídrico na safra 2016-2017



Fonte: autor, dados SISDAGRO-INMET

Figura 21 – Deficit e excedente hídrico na safra 2017-2018



Fonte: autor, dados SISDAGRO-INMET