Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Redes Bayesianas aplicadas a estimação da taxa de prêmio de seguro agrícola de produtividade

Lucas Polo

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de Concentração: Economia Aplicada

Piracicaba
2016

Lucas Polo
Engenheiro Agrônomo

Redes Bayesianas aplicadas a estimação da taxa de prêmio de seguro agrícola de produtividade

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
PROF. DR. VITOR AUGUSTO OZAKI

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de Concentração: Economia Aplicada

Piracicaba
2016

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Polo, Lucas
Redes Bayesianas aplicadas a estimação da taxa de prêmio de seguro agrícola de produtividade / Lucas Polo. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2016.
138 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura “Luiz de Queiroz”.

1. Seguro agrícola 2. Redes Bayesianas 3. Seleção de dados espaço-temporais 4. Regressão beta 5. Sensoriamento remoto 6. Meteorologia I. Título

CDD 338.17334
P778r

DEDICATÓRIA

Dedico este trabalho a minha família, meu pai Irineu, minha mãe Lourdes e minhas irmãs Marta, Mariana e Verônica.

AGRADECIMENTOS

Agradeço a Deus.

A minha família pelo apoio no que eu fiz ao longo de minha vida, inclusive nos estudos.

A meu orientador, Vitor, pela liberdade, sugestões e apoio no desenvolvimento desta dissertação.

A CAPES pela bolsa de mestrado que possibilitou minha dedicação integral na pesquisa.

A meus companheiros de GESER, em especial Rogério, Fábio, Henrique, Matheus, Denis e Carlos, pela amizade e conhecimentos compartilhados sobre sensoriamento remoto, geoprocessamento e TI.

Aos meus amigos Felipe Nazato, Sergio Zanon e Matheus Ottonelli Polo pelas discussões e contribuições durante o Mestrado.

À USP e ESALQ pelo acesso a infraestrutura e informação.

A todos os amigos que me apoiaram durante a pesquisa.

A Ronaldo Reis Jr. e José Eduardo Corrente pelo modelo TESALQ de teses em LaTeX com o qual a redação do documento da dissertação foi iniciado.

Ao INMET e ANA pela disponibilização de dados meteorológicos por meio do BDMEP e HidroWEB.

À NASA e entidades relacionadas pela disponibilização de dados de sensoriamento remoto de forma gratuita.

Por fim, agradeço às comunidades de desenvolvimento de bibliotecas de processamento de dados espaciais ”Open Source Geospatial Foundation”, aos projetos SciPy, Numpy, Scikit Learn e R, e as comunidades de desenvolvimento de software gratuitos, e open source.

EPÍGRAFE

”Everything should be made as simple as possible, but not simpler”

Roger Sessions (parafraseando Albert Einstein)

”...; just because we can handle a lot of complexity, it does not mean that we should.”

Noel Cressie e Christopher K. Wikle

8

SUMÁRIO

RESUMO

REDES BAYESIANAS APLICADAS A ESTIMAÇÃO DA TAXA DE PRÊMIO DE SEGURO AGRÍCOLA DE PRODUTIVIDADE

Informações que caracterizam o risco quebra de produção agrícola são necessárias para a precificação de prêmio do seguro agrícola de produção e de renda. A distribuição de probabilidade da variável rendimento agrícola é uma dessas informações, em especial aquela que descreve a variável aleatória rendimento agrícola condicionada aos fatores de risco climáticos. Este trabalho objetiva aplicar redes Bayesianas (grafo acíclico direcionado, ou modelo hierárquico Bayesiano) a estimação da distribuição de probabilidade de rendimento da soja em alguns municípios do Paraná, com foco na analise comparativa de riscos. Dados meteorológicos (ANA e INMET, período de 1970 a 2011) e de sensoriamento remoto (MODIS, período de 2000 a 2011) são usados conjuntamente para descrever espacialmente o risco climático de quebra de produção. Os dados de rendimento usados no estudo (COAMO, período de 2001 a 2011) requerem agrupamento de todos os dados ao nível municipal e, para tanto, a seleção de dados foi realizada nas dimensões espacial e temporal por meio de um mapa da cultura da soja (estimado por SVM – *support vector machine*) e os resultados de um algoritmo de identificação de ciclo de culturas. A interpolação requerida para os dados de temperatura utilizou uma componente de tendência estimada por dados de sensoriamento remoto, para descrever variações espaciais da variável que são ofuscadas pelos métodos tradicionais de interpolação. Como resultados, identificou-se relação significativa entre a temperatura observada por estações meteorológicas e os dados de sensoriamento remoto, apoiando seu uso conjunto nas estimativas. O classificador que estima o mapa da cultura da soja apresenta sobre-ajuste para safras das quais as amostras usadas no treinamento foram coletadas. Além da seleção de dados, a identificação de ciclo também permitiu obtenção de distribuições de datas de plantio da cultura da soja para o estado do Paraná. As redes bayesianas apresentam grande potencial e algumas vantagens quando aplicadas na modelagem de risco agrícola. A representação da distribuição de probabilidade por um grafo facilita o entendimento de problemas complexos, por suposições de causalidade, e facilita o ajuste, estruturação e aplicação do modelo probabilístico. A distribuição log-normal demonstrou-se a mais adequada para a modelagem das variáveis de ambiente (soma térmica, chuva acumulada e maior período sem chuva), e a distribuição beta para produtividade relativa e índices de estado (amplitude de NDVI e de EVI). No caso da regressão beta, o parâmetro de precisão também foi modelado com dependência das variáveis explicativas melhorando o ajuste da distribuição. O modelo probabilístico se demonstrou pouco representativo subestimando bastante as taxas de prêmio de seguro em relação a taxas praticadas no mercado, mas ainda assim apresenta contribui para o entendimento comparativo de situações de risco de quebra de produção da cultura da soja.

Palavras-chave: Seguro agrícola; Redes Bayesianas; Seleção de dados espaço-temporais; Regressão beta; Sensoriamento remoto; Meteorologia

ABSTRACT

BAYESIAN NETWORKS APPLIED TO ESTIMATION OF YIELD INSURANCE PREMIUM

Information that characterize the risk of crop losses are necessary to crop and revenue insurance underwriting. The probability distribution of yield is one of this information. This research applies Bayesian networks (direct acyclic graph, or hierarchical Bayesian model) to estimate the probability distribution of soybean yield for some counties in Paraná state (Brazil) with focus on risk comparative analysis. Meteorological data (ANA and INMET, from 1970 to 2011) and remote sensing data (MODIS, from 2001 to 2011) were used to describe spatially the climate risk of production loss. The yield data used in this study (COAMO, from 2001 to 2011) required grouping to county level and, for that, a process of data selection was performed on spatial and temporal dimensions by a crop map (estimated by SVM – support vector machine) and by the results of a crop cycle identification algorithm. The interpolation required to spatialize temperature required a trend component which was estimated by remote sensing data, to describe the spatial variations of the variable obfuscated by traditional interpolation methods. As results, a significant relation between temperature from meteorological stations and remote sensing data was found, sustaining the use of the supposed relation between the two variables. The soybean map classifier shown over-fitting for the crop seasons for which the training samples were collected. Besides the data collection, a seeding dates distribution of soybean in Paraná state was obtained from the crop cycle identification process. The Bayesian networks showed big potential and some advantages when applied to agronomic risk modeling. The representation of the probability distribution by graphs helps the understanding of complex problems, with causality suppositions, and also helps the fitting, structuring and application of the probabilistic model. The log-normal probability distribution showed to be the best to model environment variables (thermal sum, accumulated precipitation and biggest period without rain), and the beta distribution to be the best to model relative yield and state indexes (NDVI and EVI ranges). In the case of beta regression, the precision parameter was also modeled with explanation variables as dependencies increasing the quality of the distribution fitting. In the overall, the probabilistic model had low representativity underestimating the premium rates, however it contributes to understand scenarios with risk of yield loss for the soybean crop.

Keywords: Crop insurance; Bayesian networks; Spatio-temporal data selection; Beta regression; Remote sensing; Meteorology

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANA – Agência Nacional de Águas

BDMEP – Banco de Dados Meteorológicos para Ensino e Pesquisa

COAMO – Cooperativa Agropecuária Mourãoense

EVI – Enhanced Vegetation Index

IAF – Índice de área foliar

INMET – Instituto Nacional de Meteorologia

MODIS – Moderate-Resolution Imaging Spectroradiometer

NASA – National Aeronautics and Space Administration

NDVI – Normalized Difference Vegetation Index

RBF – Radial Basis Funcion

SVM – Support Vector Machine

LISTA DE SÍMBOLOS

AEVI – Amplitude do EVI ao longo do ciclo da cultura

ANDVI – Amplitude do NDVI ao longo do ciclo da cultura

AP – Ambiente de produção

$diff$ – Diferença defasada de dados de EVI

DR – Desvio de rendimento agrícola em relação a tendência

GD – Soma térmica expressa em graus-dia

I – Indenização decorrente de um sinistro coberto pelo seguro

IAF – Índice de área foliar

IE – Índice de estado da cultura

IM – Índice meteorológicos

L – Limiar usado na identificação de mudança na série de EVI, por meio de $diff$

LMI – Limite máximo de indenização

NC – Nível de cobertura

P_a – Precipitação acumulada

R_o – Rendimento agrícola obtido

R_s – Rendimento agrícola segurado

ρ_i – Reflectância em uma banda i do espectro eletromagnético

s – Ponto no espaço

SP – Sistema de produção

t – Ponto no tempo

T_b – Temperatura basal de uma cultura agrícola

Y – Rendimento agrícola

1 INTRODUÇÃO

A incerteza é uma característica inerente ao processo de produção agrícola e consequentemente ao agronegócio. O sucesso de um sistema de produção agrícola depende de fatores que estão sob controle do produtor agrícola e de fatores que não estão sob seu controle. As práticas que compõe o manejo da cultura, como aplicação de fertilizantes e defensivos, escolha da época de semeadura e escolha da cultivar a ser implantada no campo são exemplos de variáveis que o agricultor pode escolher da forma mais adequada para seu sistema de produção. Variáveis meteorológicas, como temperatura, quantidade de chuva e ocorrência de fenômenos meteorológicos extremos estão fora do controle do agricultor.

A impossibilidade de controlar todos os fatores de produção de um sistema leva a incerteza sobre qual o desempenho da cultura durante seu ciclo . Um indicador de desempenho da cultura usual é o de rendimento, ou seja, quanto de biomassa (fibras, madeira, grãos, ou outro material energético) que a cultura dispõe para ser retirada da lavoura e ser utilizada para consumo ou comercialização. O rendimento, em certos casos, é relacionado à biomassa total ou a outra característica morfológica da cultura (Tan; Anderson; Huang; Zhang; Myneni, 2005), incluindo porções da planta que não são objeto de comercialização.

O risco do processo produtivo é em geral indesejado. Ele interfere no planejamento das atividades produtivas e no atingimento dos objetivos da produção, os quais variam entre os agricultores (subsistência da família, lazer, obtenção de lucro, diversificação de negócio, etc.). É possível dizer que os agricultores se importam com risco porque eles possuem compromissos firmados de alguma natureza, e estes dependem do rendimento obtido do sistema de produção agrícola para serem honrados. Dentre os principais compromissos existentes estão a sustentação financeira do negócio e da família do produtor e o pagamento de dívidas. Riscos climáticos, biológicos, de preço, políticos, e de saúde e trabalho (Brisolara, 2013) são alguns dos riscos aos quais o produtor está exposto.

Mesmo que o agricultor não consiga controlar todos os fatores de produção, ele pode adotar estratégias de mitigação, amenização, dos impactos negativos dos

fatores não controláveis. Tomando como exemplo o fenômeno meteorológico de seca (fornecimento insuficiente de água à cultura e que leva a danos no desenvolvimento e crescimento da mesma), seus efeitos sobre uma cultura podem ser mitigados se o agricultor adotar um sistema de irrigação e fornecer a água necessária para a cultura. Os efeitos do fenômeno de granizo (precipitação de água no estado sólido) podem ser mitigados pela adoção da cobertura da cultura agrícola com telas, as quais reduzem ou impedem o impacto direto do gelo com a cultura.

Os problemas trazidos com o risco da agricultura podem ser tratados com o objetivo de amenizá-los. Um problema financeiro pode ser tratado diretamente por meio de renegociação ou transferência de compromissos, ou pela redução da variabilidade da receita do negócio. Essa variabilidade da receita pode ser reduzida por ferramentas como fundo próprio ou cooperativo, mercados futuros, ou seguro agrícola. Mais de um desses mecanismos podem ser usados simultaneamente.

O seguro agrícola se apresenta como mecanismo de transferência de risco do agricultor segurado à seguradora em troca de um prêmio pago a ela. Um agricultor com produção segurada recebe uma indenização da seguradora se algum sinistro (quebra de rendimento da cultura decorrente de evento anômalo coberto pelo seguro - no caso do seguro agrícola) ocorrer. Em geral, apenas parte do risco de produção é transferido a seguradora, enquanto que o agricultor ainda arca com parte desse risco. Essa transferência parcial é necessária para redução do risco moral (risco de o produtor deixar de usar estratégias de mitigação por conta da transferência integral do risco), o qual tende a prejudicar a seguradora. A fração de risco que é transferido é determinado pelo nível de cobertura, o qual estabelece qual a fração da produção (ou receita) segurada que a seguradora garante que o produtor obterá; e o limite máximo de indenização, o qual indica qual a máxima indenização que a seguradora paga em caso de ocorrência de sinistro.

Tomando um exemplo, se um contrato de seguro firmado estabelece que o nível de cobertura é 70% sobre uma produção Y e que o limite máximo de indenização (LMI) é 40%, então: o rendimento garantido é $0,7Y$ e o LMI é $0,4Y$. Se o rendimento obtido é $0,8Y$, não há pagamento de indenização. Se a produção obtida é $0,6Y$, a indenização é $0,7Y - 0,6Y = 0,1Y$. Se o rendimento obtido é $0,2Y$, então a indenização é $min(0,4Y, 0,7Y - 0,2Y) = 0,4Y$, em decorrência do LMI. De forma genérica, conside-

rando um nível de cobertura (NC), um limite máximo de indenização, e um rendimento obtido (R_o) ambos sendo uma fração do rendimento segurado (R_s), a indenização paga pela seguradora é: $min(LMI, max((NC - R_o), 0)) \times R_s$. Outros exemplos e descrição de alguns tipos de seguro agrícola podem ser encontrados em Miqueleto, (2011) e Ozaki, (2005).

Diversos financiamentos bancários oferecidos a agricultores requerem que a cultura agrícola na qual o financiamento será empregado esteja segurada, reduzindo assim o risco de inadimplência nesse financiamento e destacando a importância do seguro com ferramenta de gestão de risco agrícola.

De forma simplificada, na teoria econômica existem dois tipos de prêmio que um segurado paga a seguradora em troca da transferência de risco à mesma. Um deles, o prêmio puro, se refere ao montante pago a seguradora igual à esperança das indenizações decorrentes do seguro. O outro, prêmio de utilidade zero, corresponde ao montante que um segurado está disposto a pagar em troca da transferência de risco para a seguradora. No caso em que um produtor é avesso ao risco, o prêmio de utilidade zero é maior que o prêmio puro.

A contratação de seguro agrícola depende de um acerto entre seguradora e agricultor no qual são determinados o prêmio pago a seguradora, o montante de produção segurado e o nível de cobertura sobre esse montante. De forma geral, o produtor escolhe o nível de cobertura, enquanto a produção segurada é baseada na área cultivada com a cultura pelo produtor e em alguma referência de rendimento de produção. Com base nessas e outras informações a seguradora precifica o prêmio a ser cobrado. O contrato é firmado se o segurado e a seguradora concordarem com essas e outras condições do seguro em questão. De forma ideal, o valor de prêmio é estimado com base na distribuição de probabilidade da variável rendimento da cultura condicionada ao sistema de produção (relacionado ao agricultor), nos custos administrativos e atuariais, e também com base na estimativa da disposição de pagamento do agricultor (”se ele está disposto a pagar mais, a seguradora pode cobrar mais”).

Existem diversas metodologias disponíveis que podem ser usadas para precificação do prêmio. Algumas dessas metodologias buscam estabelecer uma relação entre a variável resposta rendimento e as variáveis independentes relacionadas ao risco (as quais podem levar a redução do rendimento), de maneira que a distribuição

de probabilidade relacionada ao rendimento possa ser estimada. De posse da distribuição de probabilidade é possível estimar a esperança de indenizações para um certo contrato de seguro (com determinado montante segurado, nível de cobertura e limite máximo de indenização) e o prêmio puro. Para estimação do prêmio de utilidade zero ainda seria necessário usar informações que caracterizem o comportamento do agricultor perante a decisão de contratação do seguro em situação de incerteza.

Diversas variáveis relacionadas a quebras de produção possuem os componentes temporal e espacial. Variáveis meteorológicas estão entre elas. Dessa forma, modelos temporais, espaciais, e espaço-temporais são recorrentemente usados nas modelagens de rendimento agrícola e apresentam potencial de uso no seguro agrícola (Melo; Fontana; Berlato; Ducati, 2008; Ozaki; Ghosh; Goodwin; Shirota, 2008). Dentre as abordagens adotadas estão análise de séries temporais múltiplas e de geoestatística.

Uma abordagem de modelagem de distribuição de probabilidade conjunta de variáveis que vêm sendo usada em trabalhos de inteligência artificial é a de redes Bayesianas (ou modelos hierárquicos Bayesianos). Redes Bayesianas são grafos direcionados acíclicos, ou DAG (”Direct Acyclic Graph”, em inglês). Elas buscam estimar a distribuição de probabilidade conjunta de certo conjunto de variáveis aproveitando-se de relações de independências entre elas para redução do número de parâmetros estimados. Embora haja essa finalidade para as redes Bayesianas, em muitos casos a representação diagramática do modelo é recorrentemente usada em diversas abordagens de modelagem para simples representação das relações existentes entre variáveis.

Embora várias abordagens metodológicas estejam disponíveis para solucionar problemas relacionados a precificação de prêmio, ainda existem dificuldades no que diz respeito a disponibilidade de dados e adequação do uso de metodologias com fontes de dados diversas. O uso de dados meteorológicos e de sensoriamento remoto [1] para caracterizar o risco da produção agrícola é um exemplo de integração complexa

[1] Segundo Colwell (1960, apud Jensen, 2009, p. 3), sensoriamento remoto define-se como ”*a medição ou aquisição de informação de alguma propriedade de um objeto ou fenômeno, por um dispositivo de registro que não esteja em contato físico ou íntimo com o objeto ou fenômeno em estudo*”. Um tipo de sensoriamento remoto é o orbital, realizado por sensores presente em plataformas de coleta de dados orbitais (satélites). O sensoriamento remoto mais comumente usado em pesquisas na agricultura é aquele que coleta dados da interação das culturas com a radiação solar. Essa forma de interação é em parte caracterizada pela reflectância do objeto de interesse (plantas). De forma simplificada, a reflectância retrata a fração refletida da radiação eletromagnética com certo comprimento de onda e

de dados para resolver o problema. O potencial desses dados está na caracterização do risco pela análise de séries temporais com dependência espacial (estimar prêmio), e no monitoramento de culturas implantadas (estimar indenização para um contrato corrente). Nesse contexto, a meteorologia relaciona-se às condições oferecidas pelo ambiente para a cultura se desenvolver e o sensoriamento remoto ao estado da cultura e à definição da dependência espacial de algumas variáveis meteorológicas.

Little; Schucking; Gartrell; Chen; Olson; Ross; Jenkerson; KcKellip, 2007 apresenta os esforços existentes e de cooperações entre agências de informação nos Estados Unidos da América para o uso de dados de sensoriamento na solução de problemas do seguro agrícola. Enfatizam a importância da agregação de dados em processos de ”data mining”para obtenção de informação útil para o seguro agrícola. O trabalho corrobora constatações de que dados de sensoriamento remoto podem explicar o comportamento e estado de uma cultura agrícola.

Para o caso de uso de dados meteorológicos e de sensoriamento remoto na precificação a nível municipal (ou outra forma de agrupamento) há necessidade de adotar uma forma de estimativa de dados representativos para produtores agrícolas do município. Isso se traduz na necessidade de considerar as dimensões temporal e espacial na seleção dos dados. Na literatura se verifica o uso de mapas de culturas para seleção de dados na dimensão espacial, mas a dimensão temporal ainda é deixada de lado. Nitidamente, há a necessidade de considerá-la por meio da definição ou identificação dos períodos de cultivo das culturas (ciclos). Esse é um problema que será abordado no presente trabalho, acompanhado da aplicação de redes Bayesianas para precificação de prêmio puro e para monitoramento de culturas.

1.1 Objetivo

O objetivo deste trabalho é a obtenção de uma distribuição de probabilidade condicional de rendimento agrícola (no sentido de produtividade agrícola) da cultura da soja no Paraná.

O objetivo secundário é mostrar a aplicação dessa distribuição de probabilidade na estimação do prêmio puro de um seguro agrícola de produtividade e estimação da indenização para um contrato firmado.

---

que incide sobre o objeto para uma direção de incidência da radiação e uma direção de reflexão.

1.1.1 Estrutura do documento

Para facilitar a localização do leitor no texto, uma breve descrição da estrutura do documento é apresentada a seguir.

O documento é iniciado com a introdução que brevemente apresenta alguns problemas do seguro agrícola e contextualiza a modelagem do rendimento das culturas na tentativa de solução desses problemas.

Em seguida os objetivos do trabalho são apresentados. Na mesma sessão, a metodologia do trabalho como um todo é contextualizada no processo do atingimento dos objetivos.

O desenvolvimento do trabalho foi dividido em dois artigos (duas partes). O primeiro artigo trata da seleção de dados espaço-temporais de culturas agrícolas por meio de mapa de culturas e de identificação e ciclo. O segundo artigo trata da estimação da distribuição de probabilidade de rendimento agrícola por redes bayesianas usando dados espaço-temporais de culturas agrícolas. Ambos os artigos apresentam uma introdução própria, metodologia, resultados e discussão. Parte do processo de obtenção dos dados usados no segundo artigo são obtidos com o uso das metodologias apresentadas no primeiro.

Um fluxograma dos processos que compõe as metodologias dos dois artigos é apresentado na figura 1.1.

Por fim, as considerações finais gerais são feitas em uma sessão separada.

1.1.2 Contextualização da metodologia

Como o atingimento do objetivo no contexto de modelagem agronômica é um problema complexo, há a necessidade de contextualizar a metodologia usada para uma melhor leitura do documento pelo leitor. A figura 1.2 representa essa contextualização de forma hierárquica.

Parte-se da necessidade de se obter um modelo ajustado que represente a distribuição de probabilidade condicional e uma forma de validá-lo. O modelo requer a especificação de uma estrutura, um conjunto de variáveis relevantes para o problema e um conjunto de dados para ajustá-lo. A validação requer um conjunto de dados para gerar indicadores de qualidade de ajuste e critérios de validação para verificar se o

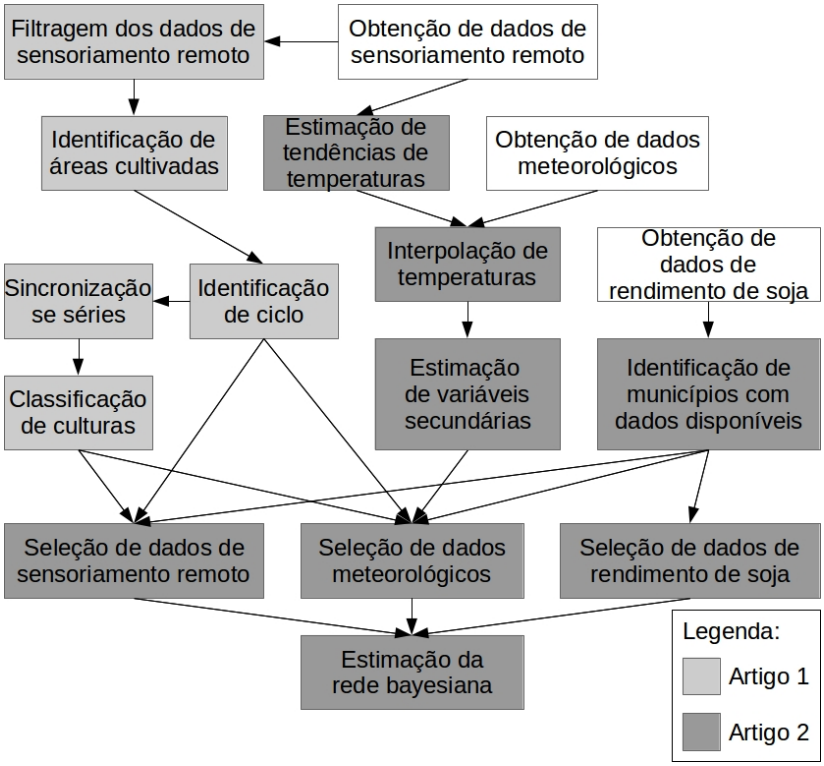


Figura 1.1 – Fluxograma da metodologia da pesquisa

modelo é adequado.

No contexto agronômico, dados de culturas agrícolas são dados espaço-temporais. Esses dados se referem ao ambiente de produção, ao sistema de produção e aos estados de desenvolvimento e de crescimento da cultura ao longo de seu ciclo. Dados meteorológicos, do solo e de manejo são exemplos de dados que caracterizam o ambiente e sistema de produção. Altura de plantas, número de nós, massa seca e rendimento de produção são exemplos de dados que caracterizam estados da cultura medidos *in situ*. O sensoriamento remoto é uma fonte de dados que podem caracterizar o estado de uma cultura, mas sem precisar entrar em contato com ela para fazer observações. Todos esses dados mencionados possuem uma componente espacial e outra temporal (onde e quando foram observados).

Para usar os dados espaço-temporais na modelagem de culturas agrícolas (em especial a variável rendimento agrícola) é preciso levar em consideração a representatividade dos dados para a cultura de interesse no ajuste do modelo e, ainda mais, na validação do mesmo. No contexto de dados espaço-temporais a seleção de dados no espaço e no tempo é crítica para que culturas de interesse sejam representadas e modeladas adequadamente. A seleção dos dados de culturas no espaço pode ser re-

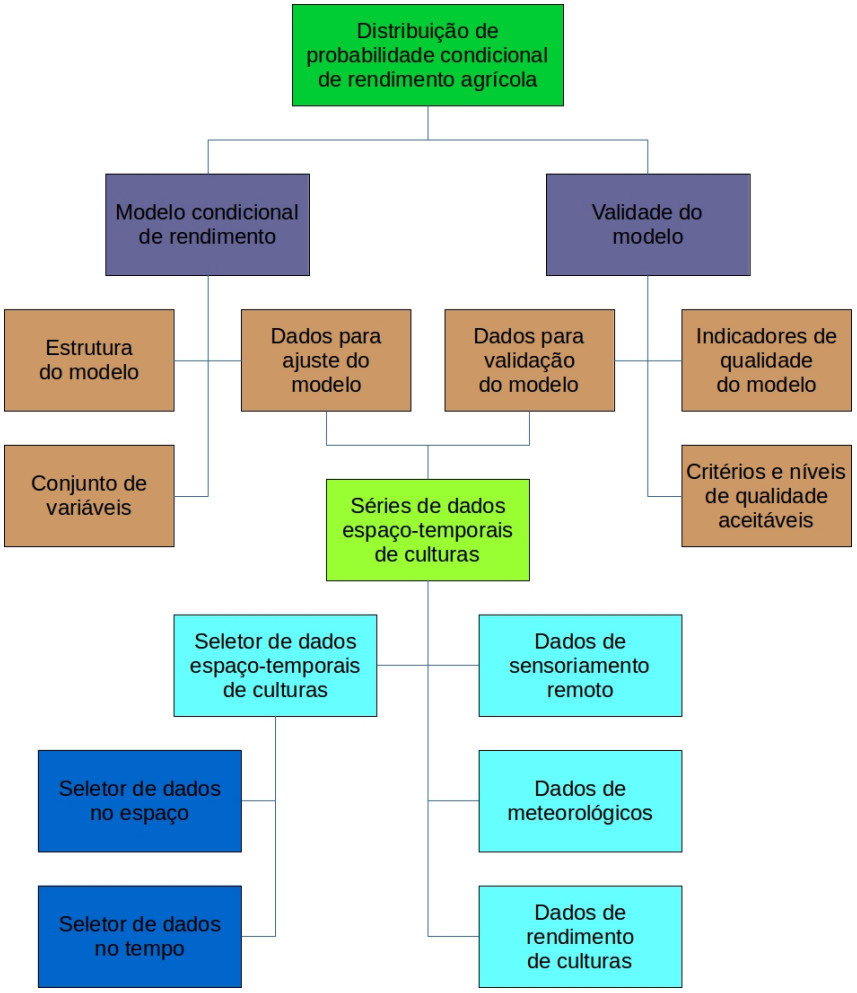


Figura 1.2 – Esquema que representa a contextualização da metodologia para atingimento do objetivo do trabalho (Fonte: autor)

alizada por meio de mapas de culturas. A seleção de dados de culturas no tempo pode ser realizada considerando o período em que a cultura foi cultivada, desde o início até o término de seu ciclo (semeadura à colheita, por exemplo).

A distribuição de probabilidade condicional de rendimento agrícola (condicionada aos fatores de produção, ambiente de produção, e estado da cultura) permite estimação do prêmio puro do seguro agrícola de produtividade e estimação da indenização a ser paga para um contrato de seguro firmado.

1.1.3 Seguro agrícola

Considera-se um indivíduo que faz escolhas a fim de maximizar sua utilidade esperada. Esse indivíduo está exposto a uma condição de risco, em que sua renda é condicionada a eventos incertos. Diz-se que a renda do indivíduo é estado-contingente,

pois depende da realização de certos estados da natureza.

Está disponível para o indivíduo um bem que garante a ele um nível de risco menor. Para ter acesso a esse bem ele deve pagar um prêmio, em troca da redução de seu risco. O indivíduo irá adquiri-lo se a utilidade esperada com a aquisição é maior que a utilidade esperada sem a aquisição. Essa função de utilidade é do tipo Von Neumann-Morgenstern.

Um indivíduo com uma dotação de renda w é submetido a uma situação de risco a qual uma variável aleatória X está associada e representa uma perda de renda. O indivíduo pode adquirir um seguro que lhe garante uma indenização I em caso de perda X, sob o pagamento prévio de um premio. Representa-se a função de utilidade de Von Neumann-Morgenstern para esse indivíduo por U.

$$E[U(w - X)] < E[U(w - X + I(X) - premio)] \tag{1.1}$$

Se a inequação 1.1 é satisfeita, o indivíduo contrata o seguro.

Retomando o que já foi mencionado, no contexto do seguro é possível distinguir dois tipos de prêmios: o prêmio puro e prêmio de utilidade zero. O primeiro corresponde a esperança das indenizações, e representa o mínimo que a seguradora deve cobrar de um indivíduo de modo que a esperança de prejuízo seja zero. O prêmio de utilidade zero é definido como o valor a ser cobrado do indivíduo de modo que ele fique indiferente entre escolher a situação de maior risco e a situação de menor risco, com pagamento do prêmio neste caso.

premio puro:

$$premio = E[I(X)] \tag{1.2}$$

premio de utilidade zero:

$$premio \in \{premio : E[U(w - X)] = E[U(w - X + I(X) - premio)]\} \tag{1.3}$$

Mesmo em um mercado competitivo de seguro, dificilmente a seguradora cobra o prêmio puro. Isso se deve basicamente a existência de custos para operação do seguro. Esses custos compreendem os custos administrativos da seguradora e a remuneração da corretora que comercializa o seguro. Ainda existe um acréscimo no valor do prêmio decorrente de riscos como o risco moral.

A subvenção ao prêmio consiste em uma forma de subsídio do prêmio por parte do governo. De acordo com condições impostas para a subvenção, o governo (federal ou estadual) paga uma fração do prêmio respeitando um limite de subvenção estabelecido.

Tomando esses pontos destaca-se que um contrato de seguro é firmado se o prêmio cobrado é maior que a esperança de indenizações somada aos custos administrativos (garante lucro para a seguradora) e acréscimo pelo risco moral, e é menor que o prêmio de utilidade zero somado ao subsídio. Qualquer prêmio entre esses dois valores possibilita uma satisfação maior do segurado e lucro esperado positivo para a seguradora.

1.1.3.1 Seleção adversa

A seleção adversa é um problema verificado no mercado de seguros decorrente da assimetria de informação. A assimetria de informação ocorre quando uma das partes envolvidas em um contrato de seguro possui mais informação sobre o risco de interesse. A literatura em geral assume que o segurado possui mais informação/conhecimento sobre o risco que a seguradora.

A seleção adversa ocorre quando é cobrado prêmio correspondente a uma situação de risco de um indivíduo com risco maior que aquele. Ela pode ocorrer em seguros cujo prêmio é estimado com base em médias de indenizações de grupos de indivíduos. Pela característica da estimação, a condição de alguns indivíduos reflete esperança de indenização menor que a de outros indivíduos. Pela média, o prêmio cobrado é sobre-estimado para alguns e subestimado para outros. Espera-se que apenas os indivíduos para os quais o prêmio cobrado é subestimado preferem a contratação do seguro. No entanto, a esperança das indenizações para esse grupo que contrata o seguro é maior que o prêmio cobrado, levando a esperança de prejuízo maior que zero para a seguradora. Ressalta-se que as escolhas pelos indivíduos dependerão de suas preferências.

Uma forma de contenção da seleção adversa consiste na estimação do prêmio cobrado para cada indivíduo usando o maior e melhor conjunto de dados e metodologias de modelagem do risco desse indivíduo. Essa medida não é definitiva, já que sempre há a possibilidade de o indivíduo deter informação que indique de forma mais

precisa (e exata) a condição de risco que o contrato do seguro abrange.

A subvenção do prêmio pode ser usada como mecanismo para tornar o prêmio arrecadado maior ou igual a esperança das indenizações de um grupo de indivíduos.

Um fato comentado na literatura é a relação entre formas de mitigação da seleção adversa e transferência de renda. Esse fenômeno é fácil de ser verificado com um exemplo. Supõe-se que dois indivíduos são obrigados a contratar seguro de renda. O prêmio do seguro foi estimado com base na esperança das indenizações médias dos dois indivíduos. Um deles possui esperança das perdas maior que do outro. Pela natureza da estimação, o prêmio cobrado de ambos (que é o mesmo) é maior que o que deveria ser cobrado de um e menor que do outro. Nesse caso, a obrigação da contratação do seguro por ambos faz com que o indivíduo de menor risco pague por parte do risco do indivíduo de maior risco. Como aquele de maior risco recebe mais indenizações, isso caracteriza uma transferência forçada de renda. O subsídio do prêmio, nesse contexto, gera a transferência de renda dos que financiam o subsídio (contribuintes) para os segurados que contrataram o seguro com sub-precificação.

1.1.3.2 Risco moral

O risco moral é outro problema existente no mercado de seguro. Ele se caracteriza pela mudança de atitude do segurado quanto a condução da cultura por ele ter contratado o seguro. Um produtor pode deixar de adotar medidas de redução de risco de perdas por ter contratado o seguro. Assim como a seleção adversa, o risco moral leva a uma esperança de indenização maior que aquela considerada na estimação do prêmio pela seguradora.

Uma forma de reduzir o risco moral a seguradora acompanhar a condução da lavoura do produtor a fim de verificar irregularidades em relação ao concordado entre as partes no seguro. Embora esse acompanhamento seja possível, em geral ele é inviável em larga escala e frequente, pois os custos são altos.

Há muito sugere-se a implantação do monitoramento remoto de lavouras pelo uso de informações de satélites, mas existem limitações metodológicas e de disponibilidade de dados que barram o uso dessa tecnologia em larga escala. Um sistema de monitoramento desse tipo possibilitaria identificar as lavouras seguradas com maior chance de uma anomalia ter ocorrido e, com isso, mobilizar peritos para uma avaliação.

1.1.3.3 Antecipação de indenização

Não bastando os problemas já mencionados, as seguradoras ainda necessitam resolver dificuldades operacionais relacionadas com o pagamento de indenizações e auditoria de sinistros. Prazos para pagamentos e mobilização de fiscais são estabelecidos no contrato e exigidos por leis. Esses prazos são relativamente curtos para seguradoras com efetivo de auditores disponíveis limitado. Grandes catástrofes climáticas que afetam a produção de muitos produtores ocasiona insuficiência de auditores.

A antecipação de quebras de produção, em especial as que afetam grande número de produtores, é de interesse para as seguradoras para que possam alocar seus esforços de forma a conseguir mobilizar o dinheiro para indenizações, e auditores para averiguação das perdas sofridas pelos produtores.

Também neste caso, informações de monitoramento remoto tem grande potencial de contribuição. A identificação e estimação da magnitude de perdas pela modelagem de culturas agrícolas é interessante por possibilitar planejamento de alocação de auditores em regiões críticas e mobilização de dinheiro para as indenizações futuras.

1.1.4 Estimação do prêmio

É possível indicar duas abordagens de estimação de prêmio puro no contexto de seguro agrícola. Uma delas toma a definição de prêmio puro (esperança das indenizações a serem pagas) e o estima com base em bancos de dados de contratos de seguros e de indenizações pagas. Esta é uma forma direta de estimação e caracteriza o prêmio puro referente àqueles indivíduos cujos contratos constam no banco de dados. Evidentemente, esta forma de estimação pressupõe a disponibilidade e um banco de dados representativo para as estimativas. Uma abordagem similar a essa é usada em Ozaki, (2009), onde modelos hierárquicos Bayesianos são aplicados a modelagem de prêmio considerando séries históricas de rendimento agrícola de municípios do Paraná.

Outra forma de estimação de prêmio puro é por meio da modelagem dos fatores geradores dos riscos cobertos pelo seguro e dos efeitos deles sobre o objeto de interesse no seguro (produção agrícola). Em relação ao método anteriormente citado, este possui a desvantagem da complexidade e requerimento da disponibilidade de bancos

de dados referentes aos riscos de interesse, porém há a vantagem da existência de modelos físicos e séries históricas longas que permitem estimar chances de ocorrência de eventos catastróficos ou de pequena probabilidade de ocorrência. Em especial destacam-se os riscos relacionados a escassez e excesso de chuva, e a temperaturas extremas. Além disso, essa forma de abordagem de modelagem de prêmio puro permite o aproveitamento de modelos dinâmicos para estimação de indenizações em uma escala de tempo inferior ao comprimento do ciclo da cultura.

1.1.4.1 Rendimento agrícola, meteorologia e sensoriamento remoto

Índices que caracterizam o ambiente de produção, sistema de produção ou a condição da cultura agrícola servem como proxy para estimativa de rendimento (Skees, 2008). A estimação desses índices possui vantagem sobre a variável rendimento por estarem associados de forma mais próxima a fenômenos físicos muito estudados e com modelos de estimação consolidados. Além disso, índices derivados de variáveis meteorológicas podem ser facilmente estimados quando dados meteorológicos estão disponíveis.

A importância de índices (meteorológicos, climáticos) já tem destaque na literatura sobre seguro agrícola, seja por sugerir o uso de índices para modelagem de produtividade agrícola ou pela sugestão de produto de seguro baseado no índice (World Bank, 2011; Leblois; Quirion, 2013).

Quando o risco de interesse é o climático, os componentes de dependência espacial e temporal devem ser considerados. Esses componentes devem-se a dinâmica do fenômeno gerador da condição climática e meteorológica. Um estado do clima em um local e momento depende das condições daquele mesmo local e dos locais vizinhos naquele momento e em anteriores. Muito dessa dinâmica se deve a fluxos de massa e energia no ambiente. Dessa forma o uso de modelagem espaço temporal torna-se algo natural, dependendo apenas do nível de complexidade a ser abordado.

Embora diversas abordagens de modelagem de variáveis meteorológicas estejam disponíveis (séries temporais multivariadas e modelos climáticos dinâmicos, por exemplo), elas requerem disponibilidade de quantidade e qualidade de dados adequada para as estimações. No que se refere a dados meteorológicos, as fontes de dados usuais são as estações meteorológicas. A não integração de bancos de dados

e o custo de implantação e manutenção das estações levam a limitação da disponibilidade dos dados meteorológicos. Essa limitação requer estratégias de contorno ao problema. O sensoriamento remoto pode ser usado de forma conjunta com dados de estações meteorológicas a fim de descrever as relações espaço-temporais das variáveis de interesse. São exemplos de uso conjunto desses dados os trabalhos de Kilibarda; Hengl; Heuvelink; Gräler; Pebesma; Tadi'c; Bajat, (2014) e Hengl; Heuvelink; Tadie; Pebesma, (2012).

Séries de dados meteorológicos costumam estar disponíveis para períodos longos, com até mais de uma observação diária. Dados de sensoriamento remoto possuem grande resolução espacial [2] (muitos pontos observados), mas em geral possuem intervalos entre aquisições longos. Um problema típico de dados de sensoriamento remoto relaciona-se interferência da atmosfera na qualidade dos dados, especialmente pela presença de nuvens.

Existem propostas de modelos para estimar rendimento de culturas agrícolas com uso de índices meteorológicos e de sensoriamento remoto. Alguns desses modelos são os denominados agrometeorológicos e os agrometeorológicos-espectrais. O modelo descritos em Melo; Fontana; Berlato; Ducati, (2008), por exemplo, estima o rendimento agrícola da soja por uma relação entre rendimento, déficit hídrico (representado pela razão entre evapotranspiração real e a potencial), e o índice de vegetação NDVI (Normalized Difference Vegetation Index). Nesse trabalho é possível perceber que os componentes meteorológico e espectrais tem efeitos significativos para explicar o rendimento da cultura e, além disso, a introdução do componente espectral resultou em melhora das estimativas quando comparado ao modelo sem o componente (modelo agrometeorológico).

Considerando as características do sensoriamento remoto, essa melhora das estimativas de modelos agrometeorológicos-espectrais é esperada. Dados de sensoriamento remoto conseguem descrever características locais da cultura agrícola e do ambiente de produção onde ela está implantada. Essa característica tem mais impor-

[2] No contexto de sensoriamento remoto, resolução espacial *"é uma medida da menor separação angular ou linear entre dois objetos que pode ser determinada pelo sistema de sensoriamento remoto"* Jensen, (2009, p 17-18). Está relacionado a densidade de pontos distintos de observação em uma superfície, e ao nível de detalhamento da informação na dimensão espacial. É associada, em geral, ao tamanho quadrado de um pixel projetado na superfície terrestre (30m x 30m, por exemplo). Quanto melhor a resolução espacial (alto *poder* de resolução), menor é a área representada por um pixel no imageamento e mais detalhada é a imagem

tância na modelagem de rendimento agrícola quando não há estação meteorológica próxima a lavoura. No caso em que variáveis meteorológicas precisam ser interpoladas (por métodos da geoestatística, por exemplo) dados de sensoriamento remoto podem ser usados como o objetivo extra de melhorar as estimativas dessas variáveis.

Referências

Brisolara, C. S. **Proposições para o desenvolvimento do seguro de receita agrícola no Brasil:** do modelo teórico ao cálculo das taxas de prêmio. 2013. 239 p. Tese (Doutorado em Economia Aplicada) — Escola Superior de Agricultura ”Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2013.

Colwell, R. N. History and place of photographic interpretation. In: Colwell, R. N. **Manual of Photographic Interpretation**. American Society of Photogrammetry, 1960. p. 33–48.

Hengl, T.; Heuvelink, G. B. M.; Tadie, M. P.; Pebesma, E. J. Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. **Theoretical and Applied Climatology**, Wien, AT, v. 107, n. 1-2, p. 265–277, 2012.

Jensen, J. R. **Sensoriamento remoto do ambiente:** uma perspectiva em recursos terrestres. São José dos Campos: Parêntese Editora, 2009. p. 598.

Kilibarda, M.; Hengl, T.; Heuvelink, G.; Gräler, B.; Pebesma, E.; Tadi’c, M. P.; Bajat, B. Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. **Journal of Geophysical Research: Atmospheres**, Washington, US, v. 119, n. 5, p. 2294–2313, 2014.

Leblois, A.; Quirion, P. Agricultural insurances based on meteorological indices: Realizations, Methods and research challenges. **Meteorological Applications**, Hoboken, US, v. 20, n. 1, p. 1–9, 2013.

Little, B.; Schucking, M.; Gartrell, B.; Chen, B.; Olson, S.; Ross, K.; Jenkerson, C.; KcKellip, R. Remote sensing and US crop insurance program integrity: data mining satellite and agricultural data. **WIT Transactions on Information and Communication Technologies**, Southampton, UK, v. 38, p. 151–159, 2007.

Melo, R. W. D.; Fontana, D. C.; Berlato, M. A.; Ducati, J. R. An agrometeorological–spectral model to estimate soybean yield, applied to southern Brazil. **International Journal of Remote Sensing**, Abingdon, UK, v. 29, n. 14, p. 4013–4028, 2008.

Miqueleto, G. J. **Contribuições para o desenvolvimento do seguro agrícola de renda para o Brasil:** evidências teóricas e empíricas. 2011. 204 p. Tese (Doutorado em Economia Aplicada) — Escola Superior de Agricultura ”Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2011.

Ozaki, V. A. **Métodos atuariais aplicados à determinação da taxa de prêmio de contratos de seguro agrícola:** um estudo de caso. 2005. 347 p. Tese (Doutorado em Economia Aplicada) — Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2005.

Ozaki, V. A. Pricing farm-level agricultural insurance: a Bayesian approach. **Empirical Economics**, Heildenberg, DEU, v. 36, n. 2, p. 231–242, 2009.

Ozaki, V. A.; Ghosh, S. K.; Goodwin, B. K.; Shirota, R. Spatio-Temporal Modeling of Agricultural Yield Data With an Application To Pricing Crop Insurance Contracts. **American journal of agricultural economics**, Cari, US, v. 90, n. 4, p. 951–961, 2008.

Skees, J. Challenges for use of index-based weather insurance in lower income countries. **Agricultural Finance Review**, Bingley, UK, v. 68, n. 1, p. 197–217, 2008.

Tan, B.; Anderson, B.; Huang, D.; Zhang, P.; Myneni, R. Potential monitoring of crop production using a satellite-based Climate-Variability Impact Index. **Agricultural and Forest Meteorology**, Amsterdam, NL, v. 132, n. 3-4, p. 344–358, 2005.

World Bank. Weather Index Insurance for Agriculture : Guidance for Development Practitioners. **Agriculture and rural development discussion paper**, Washington, DC, v. 50, p. 116, 2011.

2 SELEÇÃO DE DADOS ESPAÇO-TEMPORAIS DE CULTURAS AGRÍCOLAS POR MEIO DE MAPA DE CULTURAS E IDENTIFICAÇÃO DE CICLO

Resumo

A seleção de dados espaço-temporais é de fundamental importância para que eles representem uma cultura agrícola modelada. Essa seleção se dá no espaço por meio de um mapa de culturas, e no tempo por meio do período que compreende o ciclo da cultura. Dados de sensoriamento remoto provenientes do sensor MODIS com resolução temporal de 16 dias (produto MOD13Q1.005) são usados. Um algoritmo simples é proposto para a identificação do ciclo considerando diferença de valores sucessivos na série de dados. Com o ciclo, as séries de dados são selecionadas no tempo e usadas na classificação de culturas (soja ou milho) com Support Vector Machine (SVM). Desse processo, obteve-se a distribuição de datas de plantio no Estado do Paraná, bem como uma máscara das principais culturas agrícolas no Estado. Os resultados podem ser usados em seleção de dados de sensoriamento remoto no espaço e tempo para modelagem de produtividade das culturas agrícolas.

Palavras-chave: Seleção de dados espaço-temporais; Sensoriamento remoto; Mapeamento de culturas

Abstract

Spatio-temporal data selection is of main importance to ensure representativeness of the data for crop modelling. The selection is done in space by a crop map, and in time by the crop cycle (growth window). Remote sensing data from MODIS with temporal resolution of 16 days (MOD13Q1.005) are used. A simple algorithm is suggested to identify the crop cycle based on difference of successive values in the image time series. The cycle identification is then used to select reflectance series to be used in crop classification (soybean or corn) by Support Vector Machine (SVM). This process resulted in a seeding distribution for Paraná state, and a crop map for the main crops as well. The results can be used in spatio-temporal data selection of remote sensing data to be used in crop yield modelling.

Keywords: Spatio-temporal data selection; Remote sensing; Crop mapping

2.1 Introdução

O monitoramento de culturas agrícolas pela modelagem de rendimento agrícola e de área cultivada com o uso de dados de sensoriamento remoto e dados meteoroló-

gicos é recorrentemente abordado em pesquisas. Esses dados são denominados de espaço-temporais, ou seja, são observações de variáveis que possuem uma componente espacial (onde foi observado) e temporal (quando foi observado).

A importância das componentes espacial e temporal está relacionada à dinâmica da interação entre a cultura e o ambiente de produção (fatores de produção, generalizando). Secas que ocorrem nos estádios vegetativos tem efeitos diferentes sobre a produção que secas que ocorrem durante o enchimento de grãos na cultura milho, por exemplo. Lavouras localizadas em locais diferentes podem ter condição ambiental (solo, clima, relevo, dentre outros) distintos e podem, assim, definir potenciais produtivos diferentes para uma mesma cultura.

Para bancos de dados coletados remotamente, as componentes espacial e temporal possuem uma grande importância para a modelagem de culturas agrícolas. Elas são fundamentais para a relação de uma cultura com os dados observados da mesma. O ajuste ou treinamento de um modelo de rendimento requer dados de rendimento coletados em campo. Um modelo ajustado caracterizará a relação entre o rendimento e covariáveis de interesse. No entanto, as observações dessas covariáveis devem ser representativas para descrever os fatores de produção que resultaram naquele rendimento. Para isso, a localização da cultura no espaço e o período do tempo em que ela se desenvolveu devem ser consideradas para a obtenção dos dados e relação com o rendimento a ser modelado.Tal representatividade dos dados espaço-temporais para a modelagem de culturas de interesse cria um problema de seleção de dados espaço-temporais (SDET).

Comumente na literatura de modelagem de culturas, há grande destaque para a seleção de dados no espaço. Ela é feita por meio de mapas de culturas, o qual é obtido por classificação (manual, supervisionada, ou não supervisionada) de imagens provenientes de sensoriamento remoto. Muitos trabalhos dão ênfase a medição de áreas cultivadas.

Em contraste com a enorme quantidade de trabalhos e metodologias de classificação de culturas aplicados a seleção espacial de dados, pouco se aprofunda em metodologias de seleção de dados no tempo. Isso deve-se, em parte, pela complexidade do problema, já que a seleção de dados no tempo para culturas agrícolas deve considerar o período em que ela foi cultivada. Os dados selecionados devem ter sua

referência temporal no intervalo de tempo que vai da implantação da cultura (semeadura, plantio, ou emergência), até a remoção da cultura ou extração da produção. Esse período pode ser definido, com ressalvas, como o ciclo da cultura.

Em geral, quando a componente temporal é importante na modelagem de culturas os pesquisadores assumem um momento de início e outro de término de ciclo padrões para a cultura de interesse em uma região. Com isso os dados são selecionados para esse período invariável entre diferentes localizações no espaço. Obviamente esta abordagem é falha, já que os agricultores não implantam suas culturas todos em um único momento e também não colhem ao mesmo tempo.

Existem trabalhos que tratam da identificação do momento de emergência, de colheita e de outras fases de desenvolvimento da cultura (Zhao; Yang; Di; Li; Zhu, 2009). Essas metodologias podem ser usadas para identificar o início e o término do ciclo da cultura, de acordo com as definições do que cada um é. Com o ciclo identificado, os dados podem ser selecionados pela componente temporal.

Além das contribuições já mencionadas da seleção de dados no tempo, ela ainda possibilita a classificação de séries temporais para criação de mapas de culturas.

Tomando esse problema de seleção de dados nas dimensões espacial e temporal, este estudo tem como objetivo apresentar uma metodologia de seleção de dados espaço-temporais de culturas agrícolas usando um mapa de culturas e identificação de ciclo das culturas.

2.2 Metodologia

Tomando o objetivo do trabalho, a metodologia foi elaborada e organizada em uma sequência de operações interdependentes e com possibilidade de automação. A aplicação desta metodologia para identificação de ciclo e classificação de culturas requer disponibilidade de dados de sensoriamento remoto.

2.2.1 Definição do problema

Seja X a variável aleatória observável de interesse. Essa variável pode ser observada em um momento do tempo $t \in D_t$, e em um local $s \in D_s$ do espaço. O produto cartesiano $D_{st} = D_s \times D_t$ representa o conjunto dos pontos, representados por pares ordenados (s, t), no espaço e no tempo em que observações podem ser

feitas. A variável indexada X_{st} representa a variável de interesse que é observada no momento t e no local s. Para simplificação, a mesma notação será usada quando t e s são contínuos ou discretos.

Considera-se a existência de um conjunto Υ cujos elementos representam os genótipos (em certo nível de especificidade) de culturas disponíveis para serem cultivadas por produtores agrícolas. Uma determinada cultivar da cultura da soja é um elemento de Υ, por exemplo.

Define-se c como uma cultura agrícola anual $\upsilon \in \Upsilon$ que é cultivada por um certo período e em algum lugar. Define-se o conjunto $\mathbb{C}$ como aquele cujos elementos são os diferentes c existentes. Para distinção, c pode ser indexado de forma que c_1 e c_2 são elementos diferentes pela cultura a que se referem, pelos momentos quando a cultura é cultivada ou pelo lugar onde ela é cultivada. c_i é cultivada por um período denominado ciclo da cultura, sendo tratado aqui como o período compreendido entre emergência ou plantio (início do ciclo) até a colheita ou retirada do campo de produção (término do ciclo), sendo representados respectivamente por T_i e T_t. O lugar onde uma cultura agrícola é cultivada é representado em geral por um talhão (lavoura), e é possível considerar um talhão como o conjunto de pontos s tratados pelo agricultor como uma unidade de produção de uma cultura com possibilidade de ter o mesmo manejo dentro do talhão. Com isso, c_i pode ser representada por $c_i = (\upsilon, D_t^{c_i}, D_s^{c_i})$, em que $\upsilon \in \Upsilon$, $D_t^{c_i} = [T_i, T_t] \subset D_t$ representa o ciclo da cultura e $D_s^{c_i} \subset D_s$ representa o talhão onde a cultura é cultivada.

Assume que $c_i \neq c_j \Rightarrow (D_s^{c_i} \times D_t^{c_i}) \cap (D_s^{c_j} \times D_t^{c_j}) = \emptyset$, ou seja, não há cultivos consorciados. Esta suposição é questionável, mas é possível assumir que culturas consorciadas constituem uma única cultura (cultura “composta”).

Toma-se o interesse de modelar alguma característica da cultura c_i por meio da variável X, para o qual existe um banco de dados de observações x_{st} dessa variável. Existe o problema de encontrar o conjunto de observações x_{st} tal que $(s,t) \in D_s^{c_i} \times D_t^{c_i}$, com $D_s^{c_i} \times D_t^{c_i} \subset D_s \times D_t$.

Considerando que em uma região $R \subset D_s$ diversas culturas são cultivadas em um momento t. O conjunto das culturas sendo cultivadas nesta região e neste momento é definido como sendo $\mathbb{C}_{Rt} \subset \mathbb{C}$. Esse conjunto é definido de tal forma que $\mathbb{C}_{Rt} = \{c_i : t \in D_t^{c_i}, (\exists s \in D_s^{c_i}, s \in R)\}$. O conjunto $\{(\upsilon, s) : \upsilon \in c_i, c_i \in \mathbb{C}_{Rt}, s \in D_s^{c_i} \cap R\}$

é um mapa de culturas sendo cultivadas no momento t na região R, em que $g(c_i)$ é uma função que resulta na cultura $v \in \Upsilon$ associada a c_i.

Para determinado $c_i \in \mathbb{C}_{Rt}$ com um determinado t a seleção de dados da variável X_{st} para a cultura de interesse segue pela definição do conjunto de dados selecionados $\{x_{st} : t \in D_t^{c_i}, s \in D_s^{c_i} \cap R, c_i \in \mathbb{C}_{Rt}\}$.

O objetivo deste trabalho traduz-se na proposição de uma metodologia para obter $\mathbb{C}_{Rt}$, para $\Upsilon = \{soja\}$.

A figura 2.1 apresenta um fluxograma que resume a metodologia usada no trabalho e descrita a diante.

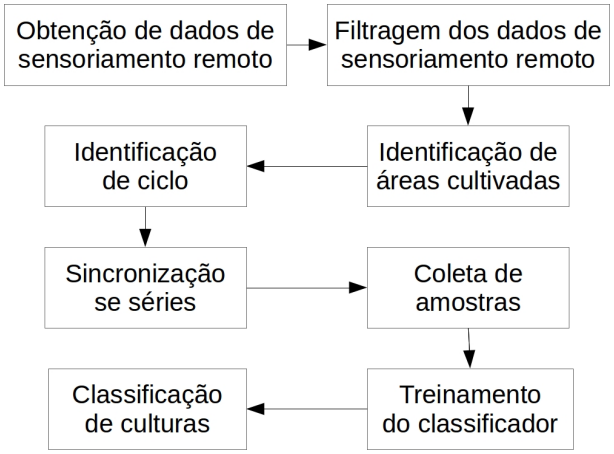


Figura 2.1 – Fluxograma da metodologia (Fonte: autor)

2.2.2 Descrição dos dados

Este trabalho tomou como área de estudo o Estado do Paraná, e como cultura a soja. O período de interesse compreende as safras de verão entre as datas 01/06/2001 e 01/06/2011, totalizando 10 safras.

Para execução deste estudo optou-se pelo uso de dados de sensoriamento remoto provenientes do sensor MODIS (Moderate Resolution Imaging Spectroradiometer), desenvolvido e mantido pela NASA. Esse projeto disponibiliza dados de reflectância da superfície terrestre de forma periódica e com resolução espacial de até 250 metros (para ser mais exato, 231.66 metros). Diversos produtos derivados de dados MODIS são disponibilizados (conjuntos padronizados de “datasets” de dados), dentre os quais está o produto MOD13Q1.005. Este produto disponibiliza dados dos índices de vegetação EVI (Enhanced Vegetation Index) e NDVI (Normalized Difference Ve-

getation Index), com resolução espacial nominal de 250 metros e disponibilizados a cada 16 dias. Os dados desse produto são originados do processamento de dados coletados diariamente, sendo que o melhor valor de EVI e NDVI desse período é tomado para compor o produto. Os dados são projetados com uma projeção Sinusoidal esférica[1]. Os dados são disponibilizados em arquivos do tipo HDF-EOS.

Neste estudo optou-se pelo uso de dados do produto MOD13Q1.005 por os dados já serem pré-processados para melhoria de sua qualidade. Os dados de EVI foram requeridos para a identificação de ciclos de culturas, e dados de EVI, NDVI e reflectâncias nas bandas do azul, vermelho, infravermelho próximo e infravermelho médio para a classificação das culturas. O "tile" [2] tomado para obtenção dos dados é o h13v11, o qual abrange todo o Estado do Paraná e outros estados da região Sul e Sudeste do Brasil.

Dados MODIS estão disponíveis gratuitamente para todo o território brasileiro com série histórica que inicia no ano 2000 e segue até os dias atuais (ano 2016). Os dados MOD13Q1.005 foram obtidos por meio do portal "Data Pool", do "NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota".

A série de dados MODIS obtida compreende 230 arquivos HDF-EOS, ocupando 46,7 GB (giga bytes) de armazenamento.

Este estudo também requereu imagens de melhor resolução espacial que as do produto MOD13Q1.005 a fim de coletar amostras de posições onde a cultura da soja foi cultivada durante o período de interesse. Essas amostras são necessárias no treinamento do classificador de culturas proposto. Para tanto, imagens de satélite do projeto LANDSAT 5 e 7 foram usadas. Apenas imagens das bandas do vermelho, infravermelho próximo e infravermelho médio foram usadas para composição de imagens com falsa cor das culturas. As imagens possuem resolução espacial de aproximadamente 30 metros, e são adquiridas a cada 16 dias.

A estrutura padrão dos dados (imagens) de satélite é raster, que consiste em uma matriz de dados com informações de projeção espacial associadas a fim de locali-

[1] Detalhes sobre essa projeção podem ser encontrados em `http://spatialreference.org/ref/sr-org/6842/`

[2] No contexto dos dados MODIS, um tile é um elemento da partição da superfície terrestre padronizada para disponibilização dos dados, com a mesma projeção sinusoidal das imagens. Na linha do equador, cada tile assemelha-se a um quadrado com 10 graus de lado (em coordenadas geográficas).

zação e relação entre diferentes imagens. A leitura e processamento dos dados foram feitos usando a linguagem de programação Python (versão 3.4), com os módulos h5py, gdal e numpy.

2.2.3 Filtragem dos dados

Filtragem pode ser considerada como a transformação de dados com o objetivo de separar componentes (informação) de interesse de outros considerados ruídos. Em caso de séries com dados faltantes, um filtro pode ser usado para substituir o dado faltante por uma alternativa viável.

Dados MODIS de NDVI e EVI podem apresentar valores faltantes, ausentes, decorrentes de excesso de cobertura de nuvens. Para substituir os dados faltantes das séries de dados um filtro simples foi usado. Esse filtro substitui um valor ausente pelo valor não ausente anterior mais próximo na série de dados. No caso em que um valor ausente é antecedido e sucedido por valores não ausentes, então o valor assumido é a média do antecessor e do sucessor. O algoritmo 1 descreve de forma detalhada o filtro. A figura 2.2 apresenta um esboço de possíveis resultados da aplicação do filtro.

```
if valor ausente é imediatamente sucedido por valor presente then
    if valor ausente possui agum antecessor presente then
        assuma a média aritmética entre o antecessor e o sucessor;
    else
        assuma valor do sucessor;
    end
else
    if valor ausente possui algum antecessor presente then
        assuma valor desse antecessor;
    else
        mantenha valor como ausente;
    end
end
```

Algoritmo 1: Algoritmo para preenchimento de série

A justificativa para uso desse filtro é que sua implementação é relativamente fácil para processamento de séries temporais de dados de imagens, permitindo que o processo seja executado em um tempo hábil.

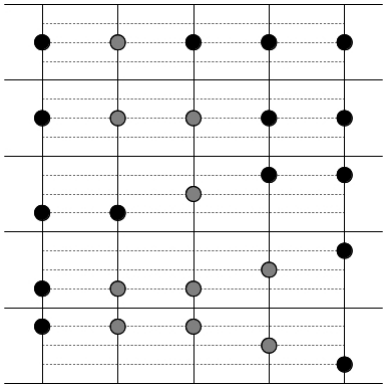


Figura 2.2 – Esboço de possíveis resultados da aplicação do filtro de dados proposto. Bolas pretas indicam dados presentes e bolas cinzas indicam dados ausentes substituídos por estimativas com o filtro (Fonte: autor)

2.2.4 Detecção de mudanças

Índices de vegetação possuem relação com o índice de área foliar (IAF) de culturas (Price, 1993; Amaral; Soares; Alves; Mello; Almeida; Silva; Silveira, 1996; Sugawara; Adami; Rudorff; Faria, 2009), como é o caso do EVI. Este índice é definido pela relação entre quantidade de superfície foliar (área de folha da cultura) por unidade de área de chão. Um valor de IAF=5 significa que sobre 1 m² de chão existem 5 m² de folhas. Por sua vez, o IAF possui relação com o desenvolvimento e crescimento de uma cultura. Genericamente, quanto mais uma cultura se desenvolve e cresce, mais folhas ela produz. O IAF pode chegar a um valor máximo para uma cultura, quando a taxa de nascimento (ou crescimento) de folhas se aproxima da taxa de perda de folhas velhas pela planta, mesmo que ela continue crescendo.

O aumento e diminuição de biomassa verde de culturas anuais interfere na variação dos valores de EVI. Após a emergência de uma cultura, o EVI aumenta chegando a um máximo. No momento da maturação seguida de colheita o valor do índice cai. Esse comportamento de aumento e diminuição do índice pode ser captado pela detecção de mudanças.

Neste estudo um método de detecção de mudanças baseado em diferença de valores com seus respectivos precedentes foi usado. A mudança é identificada quando a diferença defasada $diff$ ultrapassa um limite L, ou seja, quando $|diff| > L$. A grande vantagem desse método é sua simplicidade.

A fim de reduzir efeitos dos ruídos das séries de dados, a diferença defasada

$diff$ usada na detecção de mudanças é definida por:

$$diff_t = \frac{EVI_{t+2} + EVI_{t+1}}{2} - \frac{EVI_{t-1} + EVI_{t-2}}{2} \tag{2.1}$$

em que $diff_t$ representa a diferença defasada estimada para o momento t, e EVI_t representa o valor de EVI no momento t. Essa definição assume que para estimar a diferença para n momentos haja disponibilidade de $n + 4$ observações.

2.2.5 Áreas cultivadas

A identificação de áreas cultivadas é um processo que resulta em uma estimativa de mapa de áreas cultivadas com culturas anuais. O principal objetivo é selecionar as regiões que não correspondem a vegetação natural (florestas). É preciso ressaltar que o método não permite selecionar culturas perenes.

O mapa de áreas cultivadas com culturas anuais é usado no processo de classificação de culturas agrícolas e também na identificação de ciclo. A classificação de culturas é beneficiada pela redução de dados a serem processados e a redução de classes possíveis consideradas pelo classificador. A identificação de ciclo tem significado quando o dado processado se refere a alguma cultura agrícola, sendo evidente a importância do mapa na seleção de áreas de interesse.

De forma geral pode-se estabelecer que uma unidade do mapa de áreas cultivadas (posição no mapa) está associado a um índice que expressa a chance de tal unidade conter uma cultura anual. Não é requerido que esse índice represente uma probabilidade, mas apenas sirva para comparação entre unidades. Tal índice é então usado na seleção de áreas com culturas anuais.

É desejável que o índice seja sensível a mudanças no tipo de uso do solo, considerando o caso de áreas que não são cultivadas com culturas anuais e passam a ser, ou áreas cultivadas com culturas anuais que são posteriormente usadas para cultivos perenes. Esta sensibilidade pode ser incorporada na estimação do índice por uma função de ponderação, em que dados mais próximos do momento de estimação têm maior peso que dados mais distantes desse momento.

Embora haja possibilidade de usar um modelo complexo para estimação das áreas cultivadas com culturas anuais, optou-se neste trabalho pelo uso de um estimador simples. Estabeleceu-se um limiar L fixo, com o qual foi possível estimar a

frequência com que os valores da variável $diff$ passaram desse limite L durante o período de interesse. A justificativa para uso de tal método consiste no fato de que as amplitudes de variação de valores de EVI são menores ao longo dos anos para culturas perenes e florestas, comparando-as com as amplitudes para culturas anuais. Essa é uma forma indireta de contagem de ciclos de culturas anuais verificadas ao longo dos anos. Esta lógica fica mais clara ao observar a figura 2.3, a qual apresenta um esboço das séries de $diff$, com as quais estimou-se L. As séries representam uma cultura anual, cana-de-açúcar, cultura silvícola e floresta nativa [3].

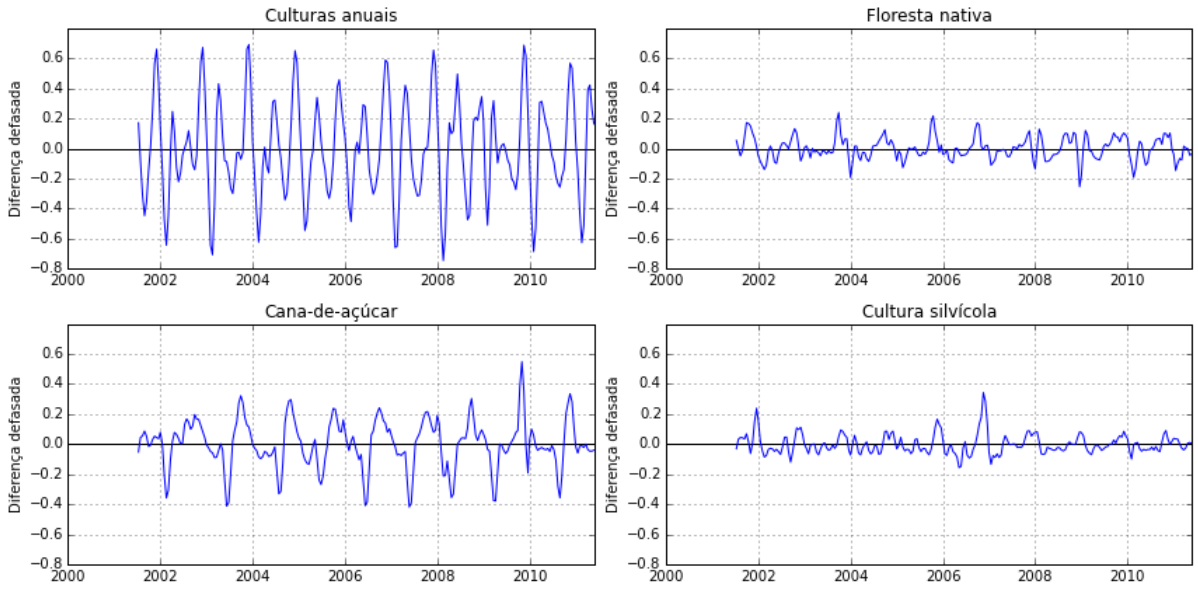


Figura 2.3 – Séries de dados estimados da variável $diff$ para culturas anuais e perene, e para floresta (verificação para data (Fonte: autor)

Em outras palavras, um objeto para o qual verifica-se pelo menos 5 ocasiões em que $diff$ ultrapassa o limite L é assumido como cultivo de culturas anuais. O valor negativo de L indica que a variação $diff$ de interesse refere-se a queda de valores de EVI, que ocorre próximo ao momento de colheita.

A definição do valor de L baseou-se na observação de gráficos de séries históricas de EVI para diferentes posições, para as quais identificou-se por interpretação visual de imagens de satélite o tipo de cobertura (cultura anual e cultura perene). O valor escolhido está acima da maior parte dos picos de valores de $diff$ para culturas perenes e para florestas (por volta de -0.1), e está abaixo da maior parte dos picos

[3] A identificação dessas classes foi realizada por interpretação visual de uma imagem Landsat 7 coletada em 11/02/2005, da órbita 223 e ponto 77. Os pontos usados para coleta são (-5342993, -2702103), (-5404605, -2800934), (-5390998, -2646560) e (-5311202, -2718110), com a mesma projeção das imagens MODIS.

para culturas anuais (por volta de -0.4). Quanto maior o valor de L em módulo maior a chance de diferenciar culturas anuais de perenes e florestas, já que a colheita de uma cultura anual leva a um decréscimo brusco de valores de EVI, que por sua vez gera valores de $diff$ grandes.

Estabeleceu-se o conjunto de valores {-0.1,-0.2,-0.3,-0.4} possíveis para L. Deste conjunto optou-se por um valor de L para ser usado na identificação de áreas cultivadas. O valor -0.1 não dá grande segurança para separação de culturas anuais de perenes e florestas. As variações anuais de EVI de florestas geram valores de $diff$ que ultrapassam L em muitos anos. O valor -0.2 permite separação de culturas anuais e semi-perenes das florestas. O valor -0.3 permite distinção de culturas anuas de semi-perenes, sendo que poucos valores de $diff$ ultrapassam L para a cana-de-açúcar (figura 2.3). O valor -0.4 é um valor para o qual a variável $diff$ de culturas anuais como a soja não atingem em diversos anos e, dessa forma, as culturas também podem não ser separadas. A fim de manter certo equilíbrio entre comissão e omissão nesse processo de classificação de culturas anuais, o valor tomado para L foi -0.3.

O mapa gerado por essa metodologia corresponde a um mapa categórico (booleano), em que valores 1 representam cultivo de culturas anuais e 0 caso contrário.

2.2.6 Identificação de ciclo

A identificação do ciclo de uma cultura consiste na identificação de dois momentos importantes: o início e o término do ciclo da cultura. A identificação desses momentos é possível devido a relação entre as alterações morfológicas das plantas que compõe a cultura e a variação das características espectrais (interação com a radiação solar) decorrentes dessas mudanças.

Na literatura, a identificação de ciclo é chamada de identificação ou mapeamento de fenologia. Em geral, os trabalhos nessa área baseiam-se nos valores máximos e mínimos de uma série de algum índice de vegetação para delimitar ciclos das culturas e, assim, sugerir a evolução da fenologia. Um exemplo de trabalho recente nessa área foi desenvolvido por Pan; Huang; Zhou; Wang; Cheng; Zhang; Blackburn; Yan; Liu, (2015), o qual usou série diária de NDVI obtida com dados dos satélites chineses HJ-1 A/B. Os autores aplicaram um software de identificação de fenologia (TI-

MESAT) para obter os parâmetros fenológicos data de início do ciclo, data de término do ciclo, data do maior valor de NDVI e comprimento do ciclo para alguns municípios da China. Diversos outros trabalhos tratam de identificação de fenologia (Sakamoto; Yokozawa; Toritani; Shibayama; Ishitsuka; Ohno, 2005;Zhao; Yang; Di; Li; Zhu, 2009;Beurs; Henebry, 2010).

O IAF está relacionado com o desenvolvimento de culturas anuais, e sua fenologia (Sakamoto; Yokozawa; Toritani; Shibayama; Ishitsuka; Ohno, 2005). O aumento de IAF sugere crescimento das plantas, e a redução do IAF sugere senescência das plantas, ataque de pragas e doenças, ou algum tipo de dano mecânico às plantas. Considerando as culturas anuais de milho e soja, a partir do início da fase vegetativa da cultura o índice de área foliar aumenta. Ele atinge um valor máximo durante o ciclo e decai no momento de senescência ou colheita da cultura.

Estudos estimam a relação existente entre índices de vegetação e o IAF (Amaral; Soares; Alves; Mello; Almeida; Silva; Silveira, 1996; Sugawara; Adami; Rudorff; Faria, 2009). Tipicamente, o aumento dos valores de EVI e NDVI indicam aumento de IAF. É importante considerar a existência do efeito de saturação do índice de vegetação, que consiste na redução expressiva do aumento o índice com o aumento de IAF (Jensen, 2009). Por essas considerações é possível compreender melhor o comportamento da relação entre EVI, IAF e o desenvolvimento da cultura (Sakamoto; Yokozawa; Toritani; Shibayama; Ishitsuka; Ohno, 2005). Uma lavoura sem cultura implantada (sem biomassa verde) apresenta valores de EVI e IAF baixos. Quando a cultura se desenvolve os valores dos índices aumentam. Próximo do momento da colheita há uma redução de biomassa da cultura, em especial de folhas, o que reduz o IAF e o índice de vegetação.

Tomando essas considerações, o momento em que valores de EVI começam a aumentar em uma série temporal pode ser usado como estimativa do momento de emergência da cultura. Vale ressaltar que a emergência a rigor ocorre alguns dias antes desse momento, pois o pequeno incremento área foliar logo após a emergência dificilmente pode ser identificado na série. A redução do valor do EVI pode ser considerado como indício de colheita. Para culturas que secam naturalmente antes da colheita, a redução de valores de EVI ocorre antes da colheita. Em culturas que são colhidas com biomassa verde o momento da colheita é muito próximo a redução do

índice. Como esboço, a figura 2.4 mostra a forma esperada de uma série de dados de EVI durante um ano safra de uma cultura anual. As duas retas verticais sugerem o momento de início e o de término do ciclo da cultura.

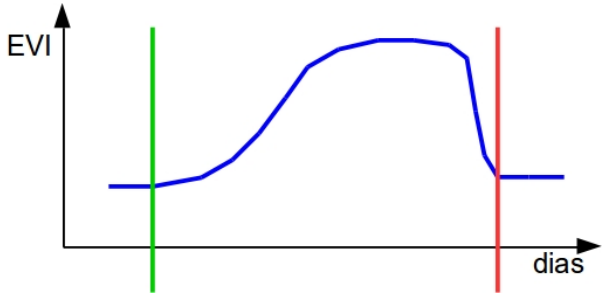


Figura 2.4 – Esboço da identificação de ciclo de culturas anuais com base em séries de dados de EVI (Fonte: autor)

Tomando por base a relação entre EVI e IAF, uma técnica de detecção de mudança pode ser usada para identificar o momento que marca o início do ciclo de uma cultura e o que marca o término do ciclo. A diferença defasada $diff$ é usada juntamente com um limite L para identificar uma mudança. Especificamente, identifica-se o momento de emergência quando $diff < L_e$ é seguido de $diff > L_e$, e o momento de colheita quando $diff > L_c$ é seguido de $diff < L_c$, sendo que L_e e L_c são os limites respectivos para emergência e colheita.

Como ressalva, em anos com quebras de safra ou com nebulosidade excessiva (interfere na qualidade dos dados de sensoriamento remoto) as estimativas de momento de emergência e de colheita podem ser afetadas, chegando ao ponto de não serem identificadas. Se apenas uma dessas datas é identificada, é possível estimar a outra tomando-se um comprimento de ciclo padrão para a cultura. Este procedimento não foi adotado neste estudo.

A figura 2.5 apresenta uma amostra de série de EVI e da variável $diff$ de uma lavoura cultivada com culturas anuais ao longo dos anos. É possível notar a relação entre aumento de EVI e valor de $diff$ positivo, e uma queda do valor de EVI e valor de $diff$ negativo.

O estabelecimento de algumas proposições lógicas que caracterizam o cultivo de culturas agrícolas é importante para dar coerência aos resultados. Uma dessas proposições é que *um ciclo de cultura não pode ser finalizado (terminado) sem que tenha sido iniciado*. Para a cultura da soja, a emergência pode representar o início do ciclo e a colheita (descaracterização total da cultura) representa o término do ciclo.

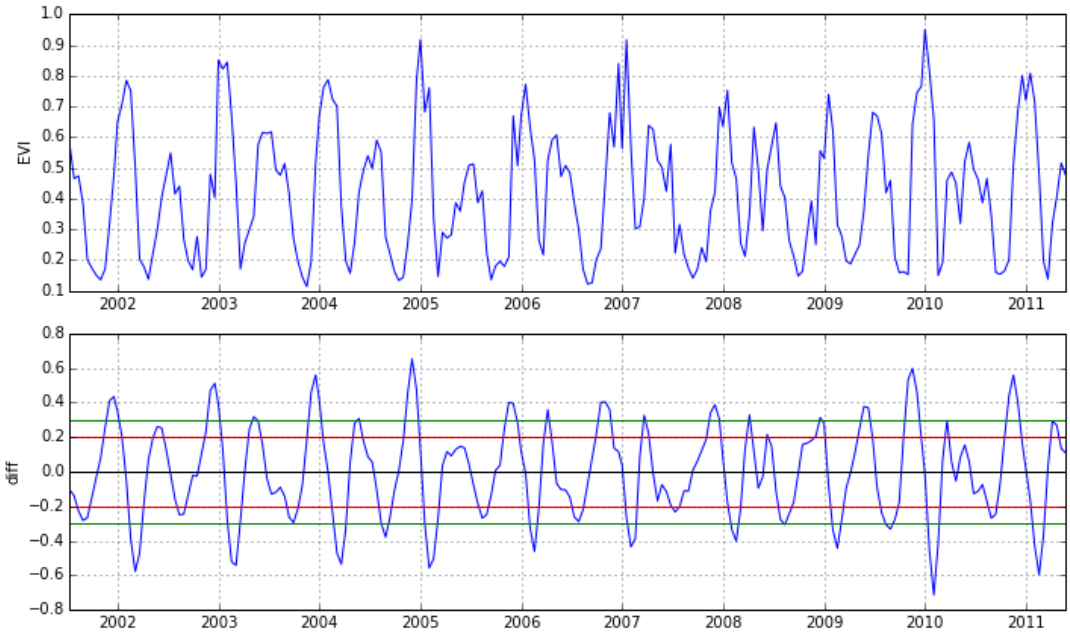


Figura 2.5 – Séries de dados de EVI e $diff$ (estimado pela equação 2.1) para uma lavoura com cultura anual (Fonte: autor)

Para a cultura da cana-de-açúcar, o plantio e rebrota representam o início de ciclo, enquanto que a colheita representa o término do ciclo.

Outras proposições que podem ser adotadas são:

- Entre dois inícios de ciclo sempre há um término de ciclo.
- Entre dois términos de ciclo sempre há um início de ciclo.

Essas duas proposições são válidas somente se assumir que *culturas implantadas em um mesma área não podem sobrepor seus ciclos*. A contestação óbvia para essa proposição refere-se aos cultivos consorciados, em que antes da colheita de uma cultura há a semeadura e início de desenvolvimento de outra. Este é o caso do consórcio de milho com pastagens.

Uma validação simples dos resultados pode ser feita pela observação do comprimento dos ciclos estimados. Se um ciclo é muito longo ou muito curto, comparando com o que se espera para uma certa cultura, uma análise mais detalhada da série de dados pode ser realizada a fim de identificar o fator que leva a anomalia. Essa validação não foi adotada no estudo.

O algoritmo que implementa o método fundamenta-se em encontrar o mínimo local da série de EVI imediatamente anterior a um ciclo de aumento de valores de EVI

e posterior redução, e o mínimo local imediatamente posterior ao ciclo, de modo que eles são assumidos como sendo respectivamente o início e o término do ciclo.

2.2.7 Classificação de culturas

O mapeamento de culturas no contexto agrícola se refere a associação de uma classe c que representa uma cultura pertencente a um conjunto de culturas possíveis $\mathbb{C}$ a um objeto espacial, sendo este um ponto ou uma região (polígono). Em geral, os métodos usados para classificação de objetos calculam um índice que define as chances de o objeto pertencer a uma certa classe. Este índice pode estar relacionado a uma medida de distância entre o vetor de atributos do objeto e o vetor de atributos da classe. O objeto é, então, associado a classe cuja chance é maior.

Os atributos usados para classificação de culturas são diversos. São exemplos de atributos as características morfológicas e fisiológicas das plantas que compõe a cultura, práticas adotadas no manejo da cultura, a dinâmica do desenvolvimento e crescimento, e forma de interação com a radiação solar. A dinâmica do desenvolvimento mencionada se refere à associação dos demais atributos ao componente tempo. Em outras palavras, se refere a evolução das características da cultura ao longo do ciclo.

Os métodos de classificação de culturas agrícolas podem ser separados em três tipos: automáticos, supervisionados e manuais. A classificação manual consiste na delimitação de regiões agricultadas por meio de um sistema de informação geográfica (SIG) e posteriormente associação das classes das culturas por interpretação visual das imagens. A classificação automática compreende os métodos de clustering de dados, os quais agrupam os objetos parecidos. Os métodos de classificação supervisionados são aqueles que requerem o fornecimento de atributos das classes a serem usadas na classificação para que os objetos espaciais sejam comparados com tais classes e então associados. Os métodos de classificação supervisionada de culturas são comumente usados por incorporarem conhecimentos técnicos do pesquisador e algoritmos eficientes de associação das classes aos objetos.

Diversos métodos estão disponíveis para classificação supervisionada de culturas. Dentre eles estão redes neurais (Shao; Lunetta, 2012), comparação de distribuição de probabilidades (por distância de Bhattacharyya, por exemplo), redes bayesianas (Pupin Mello; Rudorff; Adami; Rizzi; Aguiar; Gusso; Fonseca, 2010), e support

vector machine (SVM) (Mathur; Foody, 2008). Esses métodos requerem o fornecimento dos atributos que caracterizam cada classe de cultura, ou o fornecimento de objetos representativos de cada classe dos quais aqueles atributos podem ser estimados. O processo de ajuste do modelo com objetos representativos de cada classes é denominado treinamento do classificador.(Mello; Risso; Atzberger; Aplin; Pebesma; Vieira; Rudorff, 2013)

Algumas abordagens de classificação de culturas agrícolas transformam os dados (séries de imagens e bandas) por meio de técnicas como componentes principais para reduzir dimensionalidade dos dados e evidenciar expressões das culturas de forma mais nítida. Uma abordagem diferente para reduzir dimensionalidade de dados é apresentada em Mello, (2013), em que o autor utiliza um modelo de regressão com número de parâmetros menor que o número de variáveis (combinação de banda espectral e momento de observação) para representar essas variáveis. Essa transformação é importante se o método de classificação tem seu uso restringido quando o volume de dados é grande.

O SVM (“support vector machine”) é um classificador supervisionado (que requer treinamento com base em um conjunto de amostras) que mapeia os atributos de um objeto a ser classificado em um espaço com mais dimensões, e encontra um hiperplano que separa as classes. Esse hiperplano encontrado com base em um conjunto de amostras é então usado para classificação de objetos que não foram usados no treinamento do modelo. O método admite no momento da parametrização a possível ocorrência de classificação errada, que é interessante para reduzir o efeito de sobreajuste (“overfitting”). Além dos bons resultados de classificação, o SVM é relativamente leve para processamento de grande volume de dados, comportando paralelização da classificação por classificar cada objeto independentemente (considerando somente seus atributos).

Com o aumento da capacidade computacional utilizada na pesquisa, é natural que complexidade seja adicionada aos problemas de classificação de culturas. Segmentação e classificação passam a considerar não somente as dimensões espaciais e espectrais, mas também a temporal. Diversos trabalhos já abordam o tema.

Petitjean; Kurtz; Passat; Gançarski, (2012) apresenta uma metodologia de classificação de séries temporais de imagens multi-espectrais de sensoriamento remoto,

em que primeiramente a segmentação das imagens é feita e posteriormente clustering por K-means é aplicado para agrupar regiões semelhantes. O autor ainda destaca que a integração das dimensões temporal e espacial ainda é um problema a ser (melhor) resolvido. Um ponto mencionado pelo autor e importante no contexto de classificação, em especial na classificação de séries temporais, é a definição da medida de distância entre objetos. No trabalho o autor usa a distância euclidiana, mas sugere como alternativa medidas mais apropriadas para séries, como a apresentada em Petitjean; Inglada; Gancarski, (2012) que considera problemas de dados faltantes e séries com número de observações diferentes.

No trabalho de Mello; Vieira; Rudorff; Aplin; Santos; Aguiar, (2013), é possível considerar que a classificação de séries temporais foi utilizada. Na abordagem do autor um modelo é ajustado à série a fim de reduzir dimensionalidade dos dados. Após o ajuste, os parâmetros do modelo são utilizados na classificação supervisionada.

Neste estudo, a classificação de imagens foi executada com o objetivo de identificar áreas cultivadas com a cultura da soja nas safras de verão. Para tanto, o classificador SVM com kernel RBF ("radial basis function": $exp(-\gamma|x - x'|^2)$) foi usado. Este kernel foi escolhido seguindo recomendações de Kavzoglu; Colkesen, (2009) que verificou a classificação da superfície terrestre tem melhor performance com o kernel RBF que o polinomial. Apenas os pixels identificados como áreas cultivadas com culturas anuais foram classificados. As duas classes possíveis consideradas no estudo foram soja e milho, já que no Paraná elas são as culturas anuais majoritárias nas safras de verão. No entanto, o maior interesse está na cultura da soja.

A implementação do SVM para o processamento dos dados usou a biblioteca sklearn (escrita em Python) e baseada na biblioteca libsvm. Essa implementação considera dois parâmetros a serem fornecidos. O parâmetro C representa a penalidade dada aos erros de classificação no ajuste do modelo, sendo que quanto maior o valor de C maior é a chance de ocorrência de sobre-ajuste. Gamma é o parâmetro do kernel RBF usado.

O conjunto de atributos adotados para classificação é composto por todas as variáveis de sensoriamento remoto contidas no produto MOD13Q1.005, considerando todas as observações disponíveis ao longo do ciclo das culturas a serem classificadas:

- NDVI

- EVI
- Reflectância na banda do vermelho (ρ_r)
- Reflectância na banda do azul (ρ_b)
- Reflectância na banda do infravermelho próximo (ρ_{NIR})
- Reflectância na banda do infravermelho médio (ρ_{MIR})

Apresentando de outra forma, se τ é o conjunto de momentos t com observações dos dados do produto MOD13Q1.005 ao longo do ciclo da cultura, e $B = \{NDVI, EVI, \rho_r, \rho_b, \rho_{NIR}, \rho_{MIR}\}$ é o conjunto de variáveis e sensoriamento remoto , então o conjunto de atributos usados na classificação é $\{(x,t) : t \in \tau, x \in B\}$.

A adoção da seleção dos dados para o ciclo da cultura do objeto que se quer classificar tem um reflexo prático importante. Somente as regiões (objetos) para as quais ciclo de cultura anual foi identificado são passíveis de serem classificadas.

2.2.7.1 Sincronia das séries de dados

Para reduzir os possíveis erros decorrentes de assincronia de séries de dados, os resultados da identificação de ciclo das culturas (início e término de ciclo) foram usados para sincronizar as séries de dados. Desse processo obtém-se uma série de imagens em que um pixel de uma i-ésima imagem da série corresponde ao i-ésimo dado do ciclo da cultura a partir do momento identificado como início do ciclo.

O uso dos dados de sensoriamento remoto como atributos para classificação e a adoção da estimativa de ciclo da cultura para seleção desses dados gera um problema. Há variabilidade no comprimento de ciclo das culturas, e os algoritmos de classificação de culturas em geral requerem que o número de atributos de cada objeto a ser classificado seja o mesmo. Duas soluções podem ser usadas para contornar esse problema. Uma delas consiste em assumir valores padrões, que podem ser zeros, para observações não referentes ao objeto a ser classificado (assumir zeros após a colheita). Dessa forma todos os objetos teriam o mesmo comprimento de série de dados, mas um erro é inserido nos atributos. Outra alternativa é o uso de um artifício de escalonamento (remapeamento) dos dados. Para tanto, opta-se por um comprimento padrão de ciclo e escalona-se a série de dados para o período definido. Isso pode ser feito por

reamostragem da série com auxílio de regressão polinomial, spline, ou algum filtro de dados. Pontos específicos da reamostragem, ou parâmetros do modelo ajustado, são usados como atributos para a classificação. Além desses pontos, é fundamental que o comprimento estimado do ciclo seja usado como um atributo extra no classificador.

Neste estudo adotou-se a reamostragem por spline das séries. O valor do parâmetro que define a suavização da spline foi definido como 0,02. Este valor foi escolhido para que a spline tenha um efeito de suavização nos dados, removendo parte do ruído da série e mantendo a forma da mesma. A figura 2.6 apresenta um esboço do processo.

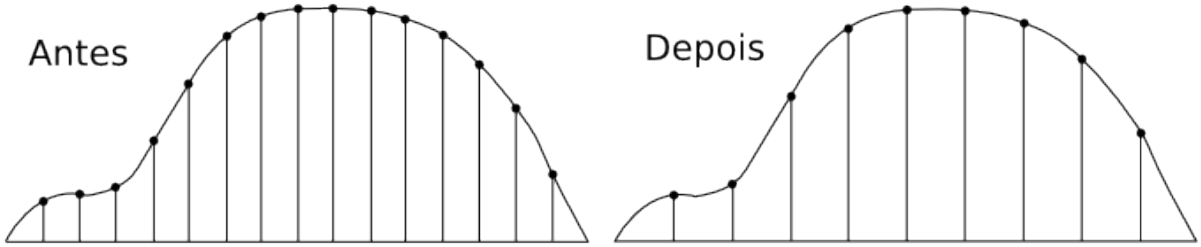


Figura 2.6 – Esboço do processo de reamostragem da série de EVI suavizada com spline (Fonte: autor)

2.2.7.2 Coleta de amostras

As amostras necessárias para o treinamento do classificador foram coletadas por meio de imagens LANDSAT 5 e 7. Essas amostras são localizações de lavouras cultivadas com as culturas de interesse identificadas por meio de técnica de interpretação visual de imagens.

Dois conjuntos de amostras foram coletados. O primeiro foi usado num primeiro ajuste do classificador, e o segundo foi usado na tentativa de melhoria da eficiência do mesmo. Os números de amostras coletadas por safra são apresentados na tabela 2.1 e 2.2, respectivamente para o primeiro e segundo conjunto de amostras.

2.2.7.3 Treinamento do modelo

Assumiu-se como procedimento padrão para o treinamento e para classificação a padronização dos dados, considerando a média aritmética e desvio padrão das variáveis das amostras.

Tabela 2.1 – Número de amostras coletadas e usadas no treinamento do SVM, para cada cultura e safra: primeiro conjunto

Safra	Soja	Milho
2001/2002	44	0
2002/2003	66	43
2008/2009	51	50
2009/2010	50	50
2010/2011	50	51

Tabela 2.2 – Número de amostras coletadas e usadas no treinamento do SVM, para cada cultura e safra: segundo conjunto

Safra	Soja	Milho
2001/2002	105	100
2002/2003	130	103
2003/2004	105	101
2004/2005	162	41
2006/2007	100	100
2007/2008	101	100
2008/2009	102	108
2009/2010	100	101
2010/2011	101	126

Para escolha dos valores dos parâmetros do classificador (C,gamma), as amostras foram separadas em dois conjuntos. Um com 60% das amostras foi usado para ajuste do classificador. As outras amostras foram usadas para calcular o erro absoluto médio do classificador. Os valores dos parâmetros foram escolhidos por uma busca em grid, de tal forma a minimizar o erro absoluto médio da classificação da amostra de teste.

O grid de valores de C e gamma usado para escolher os parâmetros corresponde ao resultado do produto cartesiano dos conjuntos $\{2^i : i \in \{-5, -4, ..., 15\}\}$ e $\{2^i : i \in \{-15, -14, ..., 3\}\}$.

Posteriormente o modelo foi ajustado com todas as amostras, e considerando os parâmetros escolhidos.

2.2.8 Recursos computacionais

Para o processamento dos dados, a linguagem python foi usada. Os módulos h5py e gdal foram usados para carregamento e manipulação de arquivos com dados espaciais. O módulo numpy foi usado para manipulação e operações matemáticas

sobre os dados carregados de arquivos. O pacote scikit-learn foi usado para a classificação de imagens por SVM. Os gráficos foram gerados pela biblioteca matplotlib do python. O algoritmo de identificação de ciclo de culturas foi implementado pelo autor, em python.

Para apoio, o SIG (sistema de informação geográfica) QGIS foi usado para visualização e produção de mapas dos resultados, o documento foi redigido em LaTeX, com auxílio do TexStudio, e compilado pelo pacote texlive.

2.3 Resultados

2.3.1 Filtragem

O método usado para filtragem dos dados teve como principal objetivo substituir valores faltantes nas séries de dados. Por partir disso, permitiu-se a adoção de um filtro simples, com implementação relativamente fácil.

A grande preocupação existente na proposição de um filtro foi seu efeito sobre os cálculos das diferenças defasadas usadas na identificação de áreas cultivadas com culturas anuais e no processo de identificação de ciclo das culturas. No entanto, o filtro não altera os valores presentes na série, nem a escala de tempo das séries de dados. A escala de tempo tem potencial de uso para identificar variações incoerentes com as séries de dados usuais das culturas pelo comprimento de ciclo esperado. A não alteração do valor presente é aceitável, neste caso por os dados de EVI já terem passado por pré-processamento e avaliação de qualidade antes de serem disponibilizados no produto MOD13Q1.005.

Uma possível melhora no método consiste em identificar outliers e reduzir o efeito de ruído indesejável. Para tanto, a utilização de wavelets aparenta potencial significativo por incluir dependência de tempo nas transformações e por manter características locais da série. Outra alternativa é o uso de regressão local como forma de suavizar a série. A regressão local permite ainda a estimação de estatísticas locais dos desvios, como o desvio padrão. Essa regressão local pode considerar as dimensões tempo, espaço e banda do espectro eletromagnético.

Chen; Jönsson; Tamura; Gu; Matsushita; Eklundh, (2004) mostram a aplicação do filtro Savitzky-Golay na filtragem de dados de NDVI de alta frequência. Tal filtro utiliza regressão polinomial local para estimar o dados a ser filtrado. Segundo os au-

tores, o filtro pode ser aplicado a séries com resoluções temporais variadas, incluindo séries diárias. Esse é um método com grande potencial, já que a implementação do filtro possui relativa facilidade.

A reamostragem para sincronia de séries de dados para a classificação das culturas pode ser considerada como sendo uma complementação a filtragem, já que o parâmetro usado na spline para reamostragem suavizou as séries removendo algum ruído.

2.3.2 Mapa de culturas anuais

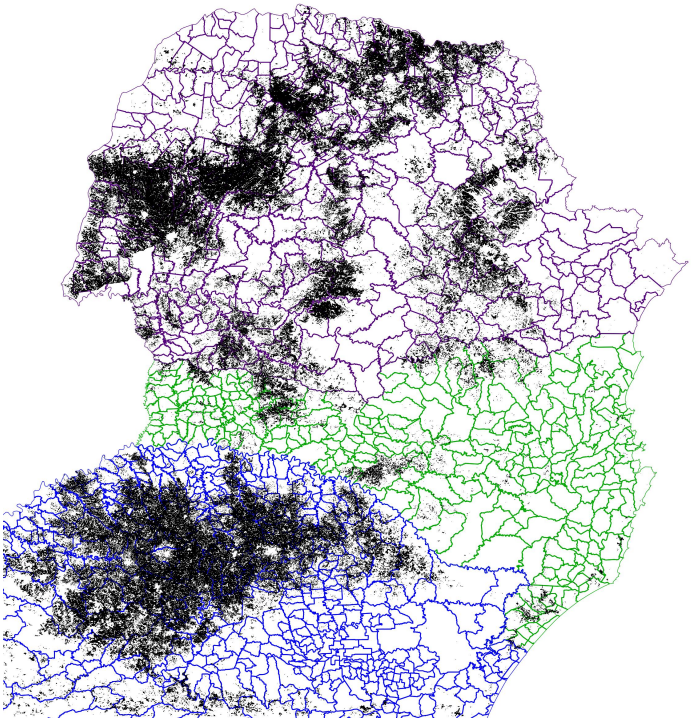


Figura 2.7 – Mapa de áreas cultivadas com culturas anuais no Estado do Paraná, Santa Catarina e parte do Rio Grande do Sul (Fonte: autor)

O mapa de cultivo de culturas anuais obtido na pesquisa e apresentado na figura 2.7 possibilita aumento da eficiência dos algoritmos que usam esse resultado, dentre eles o de identificação de ciclo e o de mapeamento automático de culturas.

O algoritmo de identificação de ciclo é beneficiado pelo mapa por reduzir os erros de estimação de ciclos de culturas anuais em vegetação nativa que apresenta alguma variação anormal na série de EVI e que assemelha-se ao padrão de uma cultura anual. Esse mapa ainda restringe as estimativas de ciclos de culturas a regiões agricultadas, tendo importância por regiões com variação de extensão de lâmina de

água (como em litorais) também apresentar uma variação de série de EVI que pode ser entendida pelo algoritmo como sendo de uma cultura anual.

A classificação automática de culturas é beneficiada pelo uso do mapa de culturas anuais por também restringir as áreas de classificação a regiões cultivadas, obviamente, e reduzir erros de classificação de objetos que não são culturas agrícolas anuais como sendo tais e por possuir características espectrais semelhantes a elas (vegetação natural em situação anômala).

Existe ainda uma outra (e grande) vantagem do uso de mapa de culturas anuais, e se refere a implementação do método. Essa vantagem é a restrição do conjunto de dados a ser processado, que reduz consideravelmente o volume de memória usada pelos programas que implementam os algoritmos de identificação de ciclo e de classificação de culturas. Para se ter uma noção dessa redução, estimou-se o número de pixels de uma imagem de EVI identificadas como cultivadas com culturas anuais e o total de pixels da imagem. Esses valores são apresentados na tabela 2.3.

Tabela 2.3 – Contagem de pixels classificados como contendo cultura anual no mapa de culturas anuais

	Quantidade de pixels	Proporção
Toda a imagem	23040000	100%
Culturas anuais	2058491	8,9%

Claramente a redução de dados a serem processados a quase um décimo é expressiva. Em especial, essa redução é importante para o processamento de séries históricas de imagens de várias bandas espectrais conjuntamente.

É preciso ressaltar que o método de estimação do mapa possui limitações. Uma delas é a incapacidade de detectar mudanças de uso do solo, como quando áreas cultivadas com culturas perenes passam a ser cultivadas com anuais, ou quando pastagens são substituídas por culturas anuais. Para tanto, o método poderia considerar padrões locais de comportamento da série.

2.3.3 Identificação de ciclo

Os resultados da identificação de ciclo são exemplificados pela figura 2.8. Claramente percebe-se a existência de ruídos na série de dados de EVI que podem influenciar a precisão das estimativas. Alguns ruídos podem até ser confundidos com

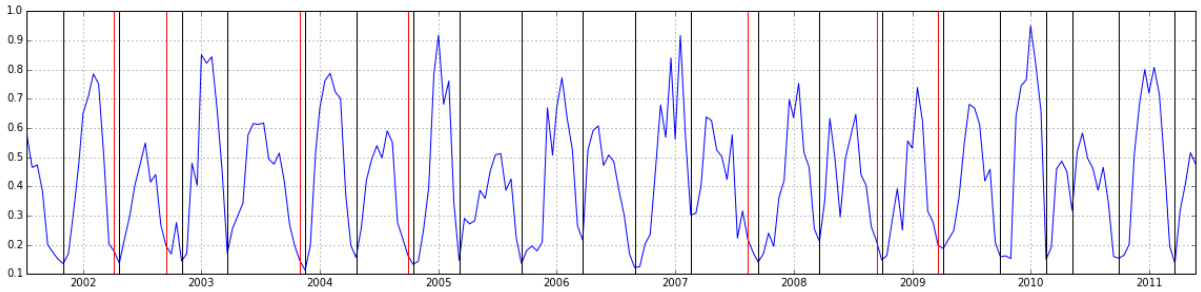


Figura 2.8 – Esboço do resultado obtido com a identificação de ciclo (Fonte: autor)

término seguido de início de ciclo em certas safras. Esse problema pode ser amenizado se considerar o uso de um filtro de série de dados mais robusto, ou ainda o uso de imagens de sensoriamento remoto de diferentes sensores como os do projeto LANDSAT 5, 7 e 8 para aumentar a disponibilidade de dados.

A exatidão da estimativa do ciclo da cultura depende da resolução temporal das imagens de sensoriamento remoto disponíveis. No caso do produto MOD13Q1.005, os dados são disponibilizados a cada 16 dias. Por isso os momentos de início e término de ciclo estão restritos a uma exatidão de 16 dias, havendo ainda o erro de estimação pelo método. Também nesse caso o uso de imagens de outras fontes pode beneficiar a estimação, ainda mais se os momentos de coleta dos dados são não coincidentes.

Ainda há de se considerar que a forma como o método é estruturado depende de as diferenças defasadas de EVI atingirem certos limites para verificar um início e um término de ciclo. Em anos de quebra de produção, como foi verificado na safra de verão de 2008/2009, o desenvolvimento das culturas é afetado de forma pronunciada causando uma anomalia no comportamento da série de EVI. Esse comportamento anômalo pode levar a perda de eficiência do método e a não identificação de um ciclo de uma cultura cultivada.

Embora o método possa ser melhorado, já percebe-se grande potencial de uso da informação de ciclo das culturas no monitoramento de culturas. Tratando-se de seguro agrícola, essa informação poderia ser usada cautelosamente para inferir se uma cultura foi implantada no período recomendado pelo zoneamento agroclimático para uma certa região.

Certamente não se deixa de lado um dos objetivos para utilização do método nesta pesquisa. O objetivo de seleção de dados na dimensão temporal para uso em

monitoramentos. Se o pesquisador considera a precisão e exatidão dos resultados como aceitáveis, então é possível usá-los para seleção de dados para a modelagem de culturas em larga escala e de forma automática, mas de preferência supervisionada.

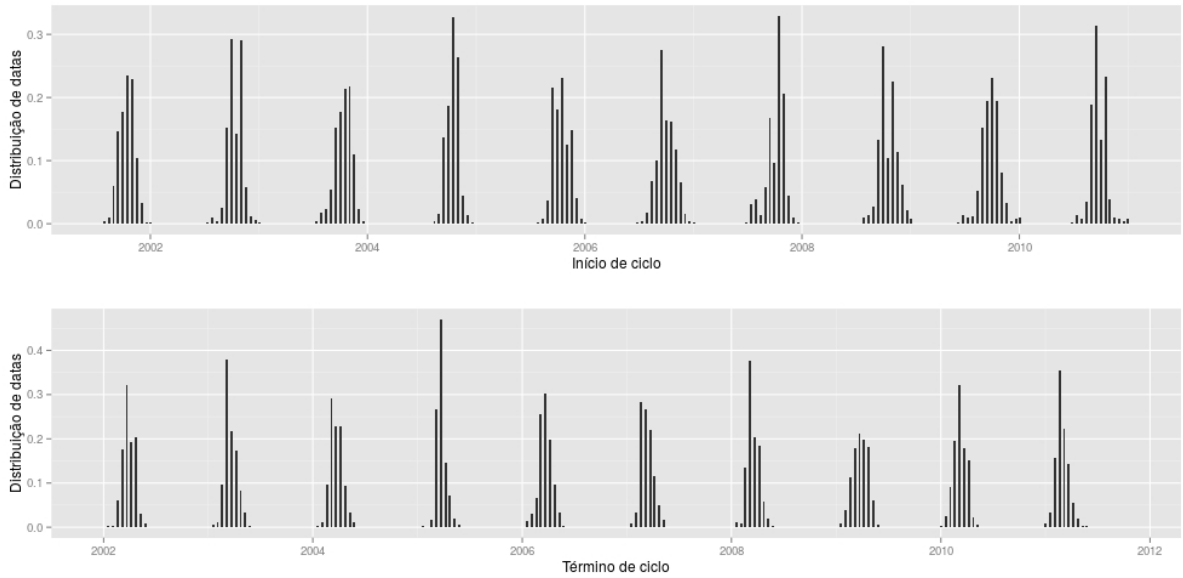


Figura 2.9 – Distribuição dos momentos de início e término de ciclo para a cultura da soja no Estado do Paraná (Fonte: autor)

A título de exposição, a figura 2.9 apresenta a distribuição de datas de início e datas de término dos ciclos identificados para cada uma das safras abordadas pela pesquisa para as regiões classificadas como contendo soja no Paraná. Uma impressão (já esperada) é que a dispersão dos inícios de ciclo é diferente entre os anos. A safra de 2008/2009 apresenta maior dispersão de inícios e términos de ciclo que a safra de 2004/2005. Essa dispersão pode ter relação direta com o regime de chuvas no período em que as culturas foram implantadas, ou com a homogeneidade das datas de colheita da cultura de inverno.

2.3.4 Classificação de culturas

Dentre os valores testados para os parâmetros do SVM, optou-se pelos valores C=8,0 e gamma=0,0078125.

Os resultados da parametrização do SVM são apresentados na tabela 2.4 e 2.5, considerando um primeiro conjunto de amostras e um segundo conjunto de amostras (ampliado), respectivamente. A tabela 2.4a e 2.5a apresenta os resultados de classificação das amostras usadas no treinamento do SVM. A tabela 2.4b e 2.5b apresenta

os resultados de classificação das amostras de teste. Por fim, a tabela 2.4c e 2.5c apresenta os resultados de classificação das amostras quando todas são usadas no treinamento do classificador.

Tabela 2.4 – Dados de pixels classificados como contendo cultura anual no mapa de culturas anuais

	Estimado	
	Milho	Soja
Milho	82	0
Soja	0	161

(a) Amostras para treinamento

	Estimado	
	Milho	Soja
Milho	65	5
Soja	4	89

(b) Amostras para teste

	Estimado	
	Milho	Soja
Milho	152	0
Soja	0	254

(c) Todas as amostras

Tabela 2.5 – Resultados da classificação das amostras com os modelos do SVM ajustados

	Estimado	
	Milho	Soja
Milho	525	3
Soja	11	592

(a) Amostras para treinamento

	Estimado	
	Milho	Soja
Milho	334	18
Soja	16	387

(b) Amostras para teste

	Estimado	
	Milho	Soja
Milho	872	8
Soja	22	984

(c) Todas as amostras

Como mencionado, primeiramente um conjunto de amostras coletadas das safras 2001/2002, 2002/2003, 2008/2009, 2009/2010 e 2010/2011 foram usadas na pa-

rametrização. Uma primeira impressão que esses resultados dão, principalmente os das tabelas 2.4a e 2.4c, é que ocorre sobre-ajuste do SVM. Nenhuma amostra usada no treinamento é classificada incorretamente. Ao observar os valores na tabela 2.4b percebe-se que realmente há um certo nível de sobre-ajuste. No teste do classificador verifica-se que 4 das 93 amostras de lavouras de soja são classificadas como milho, e 5 das 70 amostras de milho são classificadas como soja. Assim, houve 5,6% de erro de comissão para a classe soja e 6,1% para a classe milho.

Posteriormente, aumentou-se o número de amostras usadas no ajuste do classificador, sendo que elas foram obtidas das safras de 2001/2002 a 2010/2011. O resultado disso foi a melhoria da classificação para as safras anteriormente não abrangidas pelas amostras (análise visual). Verifica-se pelas tabelas 2.5a, 2.5b e 2.5c que o sobre-ajuste foi reduzido.

Embora não seja possível perceber pelas tabelas 2.4 e 2.5, a visualização dos resultados da classificação (imagens) evidencia uma limitação do método. Muitos pixels de borda de talhões de soja foram classificados como sendo não soja (como milho). Assim houve sobre-estimação de área cultivada com milho, e subestimação da área cultivada com soja. Por outro lado, as áreas classificadas como soja representam pixels “puros”, ou seja, os dados espectrais não se referem a misturas com outras culturas. Este fato é de certa forma vantajoso para seleção de dados para modelagem da cultura da soja (redução da interferência de culturas vizinhas nos dados de sensoriamento remoto).

Considerando os resultados apresentados na tabela 2.5b, o erro resultante da classificação é relativamente alto: 4,6% de comissão e 4,1% de omissão para a soja. Isto restringe o uso da técnica para a finalidade de obtenção de áreas cultivadas de forma automática, ainda mais se considerar o erro com a classificação de pixels de bordas de talhões. No entanto, o resultado ainda mostra-se promissor para a seleção de dados de cultivos de soja para uso em modelagens a nível regional.

É importante mencionar que a classificação de culturas é afetada pela condição de quebra de produção de algumas safras. A interação da cultura com a radiação solar que gera a resposta captada pelo sensor orbital (dado observado) depende da condição em que a cultura está. Na situação ideal o classificador deve ser robusto para considerar essas variações para uma mesma cultura. Como a metodologia usada

nesta pesquisa não considerou esse ponto, há a possibilidade de que em anos de quebra de safra ocorreu perdas de eficiência do classificador por esse motivo.

Uma possível verificação de consistência e qualidade das estimativas pode ser efetuada com a comparação de estimativas oficiais de áreas cultivadas com as culturas agrícolas de interesse. Ainda é possível comparar os resultados da classificação automática com um mapeamento manual de culturas por meio de interpretação de imagens e usando imagens com melhor resolução espacial (LANDSAT, por exemplo).

Os resultados da classificação são apresentados na figura 2.10.

2.4 Considerações finais

A seleção de dados espaço temporais se demonstra um processo complexo e fundamental. A cadeia de processos que compõe essa seleção requer conhecimentos em diversas áreas como classificação de imagens, estruturas de dados, e implementação computacional. Mesmo assim sua adoção é de fundamental importância para aumentar a representatividade de dados espaço temporais, como os de sensoriamento remoto e meteorológicos.

Decisões a certa de diversos pontos críticos em sua metodologia são difíceis pelo “tradeoff” entre complexidade metodológica e implementação. Desses pontos destacam-se resolução temporal e espacial das imagens de índice de vegetação, fontes de dados, método de classificação, amostra de dados para treinamento de classificadores, algoritmo de identificação de ciclo e seus parâmetros.

Além da importancia no aumento da representatividade de dados em pesquisas, o processo seleção de dados espaço temporais também é base para entendimento da dinâmica de ocução do solo. As áreas cultivadas são obtidas no processo de classificação das culturas implantadas em um certo momento, e a distribuição de datas de plantio é obtida a partir da identificação dos ciclos das culturas.

Muita pesquisa existe nesta área, mas ela ainda requer avanços no sentido de facilitação da aplicação da seleção de dados em pesquisas. Um grande avanço seria a disponibilização de produtos resultantes de classificação e identificação de ciclo em formato parecido com os produtos MODIS, e que fossem acessíveis pela comunidade científica para aplicação direta em suas pesquisas.

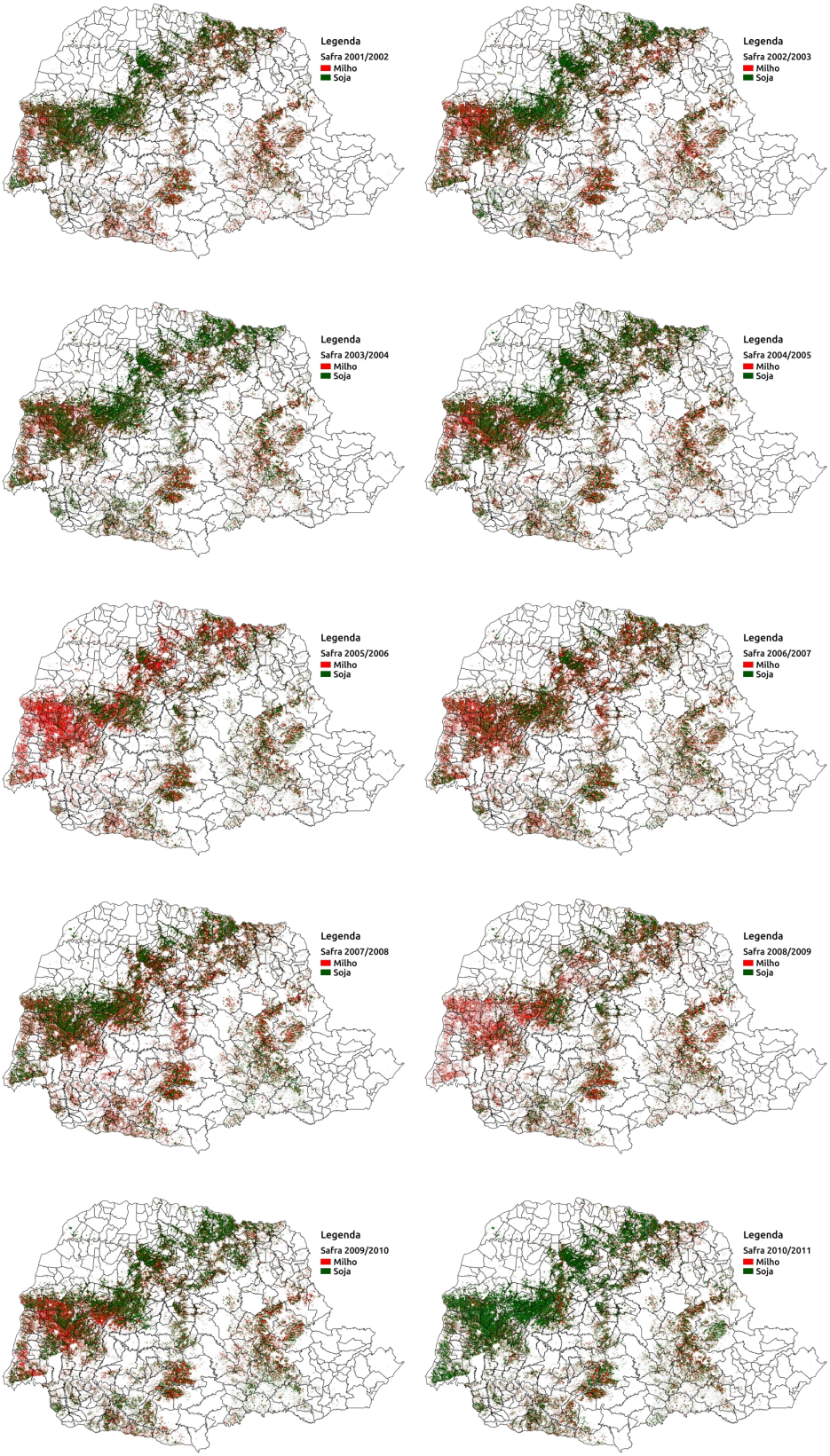


Figura 2.10 – Mapas de cultivo de soja e milho do Estado do Paraná para as safras de 2001/2002 a 2010/2011 (Fonte: autor)

Referências

Amaral, S.; Soares, J. V.; Alves, D. S.; Mello, E. M. K. de; Almeida, S. A.; Silva, O. F. da; Silveira, A. M. Relações entre Índice de Área Foliar (LAI), Área Basal e Índice de Vegetação (NDVI) em relação a diferentes estágios de crescimento secundário na Floresta Amazônica em Rondônia. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 8, 1996, Salvador. **Anais ...** São José dos Campos: INPE, 1996. p. 485–489.

Beurs, K. M. de; Henebry, G. M. **Spatio-Temporal Statistical Methods for Modelling Land Surface Phenology**. New York, US: Springer, 2010. p. 177–208.

Chen, J.; Jönsson, P.; Tamura, M.; Gu, Z.; Matsushita, B.; Eklundh, L. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. **Remote Sensing of Environment**, New York, US, v. 91, n. 3-4, p. 332–344, 2004.

Jensen, J. R. **Sensoriamento remoto do ambiente:** uma perspectiva em recursos terrestres. São José dos Campos: Parêntese Editora, 2009. p. 598.

Kavzoglu, T.; Colkesen, I. A kernel functions analysis for support vector machines for land cover classification. **International Journal of Applied Earth Observation and Geoinformation**, Amsterdam, NL, v. 11, n. 5, p. 352–359, 2009.

Mathur, A.; Foody, G. M. Crop classification by support vector machine with intelligently selected training data for an operational application. **International Journal of Remote Sensing**, Abingdon, UK, v. 29, n. 8, p. 2227–2240, 2008.

Mello, M. P. **Spectral-temporal and Bayesian methods for agricultural remote sensing data analysis**. 2013. 122 p. Tese (Doutorado em Sensoriamento Remoto) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2013.

Mello, M. P.; Risso, J.; Atzberger, C.; Aplin, P.; Pebesma, E.; Vieira, C. A. O.; Rudorff, B. F. T. Bayesian Networks for Raster Data (BayNeRD): Plausible Reasoning from Observations. **Remote Sensing**, Postfach, CH, v. 5, n. 11, p. 5999–6025, 2013.

Mello, M. P.; Vieira, C. a. O.; Rudorff, B. F. T.; Aplin, P.; Santos, R. D. C.; Aguiar, D. a. STARS: A new method for multitemporal remote sensing. **IEEE Transactions on Geoscience and Remote Sensing**, Piscataway, US, v. 51, n. 4, p. 1897–1913, 2013.

Pan, Z.; Huang, J.; Zhou, Q.; Wang, L.; Cheng, Y.; Zhang, H.; Blackburn, G. A.; Yan, J.; Liu, J. Mapping crop phenology using NDVI time-series derived from HJ-1 A/B data. **International Journal of Applied Earth Observations and Geoinformation**, Amsterdam, NL, v. 34, p. 188–197, 2015.

Petitjean, F.; Inglada, J.; Gancarski, P. Satellite Image Time Series Analysis Under Time Warping. **IEEE Transactions on Geoscience and Remote Sensing**, Piscataway, US, v. 50, n. 8, p. 3081–3095, 2012.

Petitjean, F.; Kurtz, C.; Passat, N.; Gançarski, P. Spatio-temporal reasoning for the classification of satellite image time series. **Pattern Recognition Letters**, Amsterdam, NL, v. 33, n. 13, p. 1805–1815, 2012.

Price, J. Estimating leaf area index from satellite data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 31, n. 3, p. 727–734, 1993.

Pupin Mello, M.; Rudorff, B. F. T.; Adami, M.; Rizzi, R.; Aguiar, D. a.; Gusso, A.; Fonseca, L. M. G. A simplified Bayesian Network to map soybean plantations. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYPOSIUM, 2010, Honolulu, US. **Anais ...** New York, US: IEEE, 2010. p. 351–354.

Sakamoto, T.; Yokozawa, M.; Toritani, H.; Shibayama, M.; Ishitsuka, N.; Ohno, H. A crop phenology detection method using time-series MODIS data. **Remote Sensing of Environment**, New York, US, v. 96, n. 3, p. 366–374, 2005.

Shao, Y.; Lunetta, R. S. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. **ISPRS Journal of Photogrammetry and Remote Sensing**, Amsterdam, NL, v. 70, p. 78–87, 2012.

Sugawara, L. M.; Adami, M.; Rudorff, B.; Faria, V. Avaliação de três métodos de estimativa de índice de área foliar aplicados à cana-de-açúcar. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14, 2009, Natal. **Anais ...** São José dos Campos: INPE, 2009. p. 499–506.

Zhao, H.; Yang, Z.; Di, L.; Li, L.; Zhu, H. Crop phenology date estimation based on NDVI derived from the reconstructed MODIS daily surface reflectance data. In: INTERNATIONAL CONFERENCE ON GEOINFORMATICS, 17, Geoinformatics, 2009, Fairfax, US. **Anais ...** New York, US: IEEE, 2009. p. 1–6.

3 DISTRIBUIÇÃO DE PROBABILIDADE CONDICIONAL DE RENDIMENTO AGRÍCOLA APLICADA AO SEGURO AGRÍCOLA

Resumo

Esta pesquisa tem como objetivo aplicar redes bayesianas para estimação da distribuição de probabilidade de rendimento da cultura da soja em alguns municípios do Estado do Paraná. O modelo considera que o ambiente de produção juntamente com a tecnologia do sistema de produção condicionam o rendimento da cultura. Dados meteorológicos (ANA e INMET) e de sensoriamento remoto (NDVI e EVI do produto MOD13Q1.005 do MODIS) para as safras de 2001/2002 a 2010/2011 foram usados na estimação da distribuição de probabilidade. Os dados meteorológicos foram interpolados e agregados por média para obter séries municipais. Seleção espaço-temporal dos dados de índices de vegetação foi aplicada para cálculo da amplitude do índice para cada município. Diferentes distribuições de probabilidade foram ajustadas para as variáveis de ambiente e regressão beta foi usada para descrever o rendimento da cultura e amplitude dos índices de vegetação de forma condicional (considerando dependência da média e precisão com as variáveis explicativas). A distribuição log-normal mostrou-se mais adequada para caracterizar variáveis do ambiente de produção. O modelo probabilístico se demonstrou pouco representativo subestimando as taxas de prêmio de seguro em relação a taxas praticadas no mercado, mas ainda assim apresenta contribui para o entendimento comparativo de situações de risco de quebra de produção da cultura da soja.

Palavras-chave: Seguro agrícola; Redes Bayesianas; Seleção de dados espaço-temporais; Regressão beta; Sensoriamento remoto; Meteorologia

Abstract

Bayesian networks were applied to estimate the probability distribution of soybean yield for counties in Paraná State. The model assumes that production environment and technology conditionate crop yield. Meteorological (ANA and INMET) and remote sending data (NDVI and EVI from MODIS MOD13Q1.005 product) for crop seasons of 2001/2002 to 2010/2011 were used to estimate the probability distribution of soybean yield. The meteorological data was interpolated and aggregated by mean for each county. Spatio-temporal data selection is used to select the vegetation indexes and calculate the vegetation index amplitude for the counties. Different probability distributions were fitted for the environment variables, and beta regression was used to describe the crop yield and vegetation indexes amplitude in a conditional way (modeling both mean and precision of beta as a function of the explaining variables). The log-normal distribution showed to be the better to describe the production environment variables. The final probabilistic model did not represented well the magnitude of the risk of yield losses underestimating insurance premium rates when taking market as a

reference, but even though it contributes to understand risk scenarios of soybean yield losses relativelly.

Keywords: Crop insurance; Bayesian networks; Spatio-temporal data selection; Beta regression; Remote sensing; Meteorology

3.1 Introdução

A expansão do mercado de seguro agrícola no Brasil ainda barra em problemas de disponibilidade e qualidade de informações que permitam caracterizar o risco da produção agrícola. Essas informações são importantíssimas para elaboração de novos produtos de seguro e para planejar a expansão do segmento.

A disponibilidade e qualidade de informações sobre risco estão relacionadas a problemas recorrentes no seguro agrícola. A assimetria de informação é um problema caracterizado por uma das partes do seguro (segurado ou seguradora) deter mais informações, ou informações de melhor qualidade, que a outra parte levando a uma vantagem para decisão sobre o estabelecimento do contrato de seguro. O risco moral é outro problema associado ao seguro agrícola que consiste em o segurado não conduzir o cultivo de maneira a minimizar os riscos de perdas de acordo com o que foi acertado na contratação do seguro (agricultor assume descaso com a produção pelo fato de estar coberto por seguro). Um outro problema que pode ser citado é o de antecipar a verificação de sinistros, que impacta o planejamento de mobilização de auditores para averiguação de sinistros, e preparação de fundos para pagamento de indenizações.

Esses três problemas citados podem ser separados em dois grupos. A assimetria de informação é um problema que tem seu objeto de geração antes da contratação do seguro, pois influencia nas decisões sobre contratação do seguro e precificação do prêmio a ser cobrado por ele. O risco moral e a antecipação de verificação de sinistros são problemas que têm o objeto gerador após a contratação do seguro, apesar de que o risco moral já pode existir antes da contratação pela forma de agir do segurado.

A modelagem de culturas agrícolas insere-se nesse contexto por fornecer subsídio informacional que permite reduzir parte dos problemas mencionados. A modelagem dos fatores de risco cobertos por um seguro e da relação deles com o rendimento das lavouras ou da renda do segurado permite a estimação da indenização a ser paga

para um contrato de seguro a ser firmado. Essas informações ainda permitem que a seguradora estime o prêmio a ser cobrado dadas as características do contrato e da produção do produtor agrícola. Nesse sentido, a seguradora se mune de informações que permitem a redução da assimetria de informação quando supõe-se que o segurado tem mais conhecimentos sobre os riscos de interesse.

Essa modelagem, quando aplicada para estimar a condição de um cultivo, pode ser usada para solucionar parte do problema do risco moral e antecipação da verificação de sinistros. O monitoramento de culturas permite estimar características do desenvolvimento de culturas de forma remota, com ou sem apoio de visitas a campo. Dentre as características, a identificação de momento de semeadura, estimação de fases de desenvolvimento e identificação das áreas cultivadas podem contribuir para verificação da ocorrência de descaso por parte do produtor quanto a condução da cultura. É preciso enfatizar que o uso desse tipo de tecnologia como único subsídio para definição do problema moral não é recomendado, já que limitações metodológicos são conhecidas e podem levar a prejuízos para a segurador. O monitoramento das lavouras de forma remota também pode ser usado para estimação de rendimento da cultura ao final do seu ciclo, permitindo a estimação de indenizações a serem pagas a um produtor. Essa estimação permite o planejamento da alocação de auditores de sinistros antecipadamente em regiões com maior incidência de quebra de produção e possibilidade de sinistros.

Com base no exposto, verifica-se que ainda há interesse em metodologias que contribuam para geração e melhoria das informações disponíveis para o mercado de seguro agrícola a fim de ajudar as seguradoras e segurados a tomarem decisões. O presente trabalho objetiva beneficiar a geração de informações úteis para o mercado de seguros por apresentar uma metodologia que integra o processamento de dados de sensoriamento remoto e meteorológicos a identificação de ciclo e mapeamento de culturas de forma automática para melhorar estimativas da distribuição de probabilidade da variável rendimento da cultura da soja por meio de redes bayesianas (modelos hierárquicos bayesianos).

3.2 Metodologia

A metodologia utilizada nesta pesquisa é estruturada de forma a possibilitar a estimação da distribuição de probabilidade conjunta de variáveis meteorológicas, de sensoriamento remoto, e rendimento da cultura da soja, na forma de uma rede bayesiana (modelo hierárquico bayesiano). Dentre os riscos cobertos pelo seguro agrícola, o risco de seca é o foco.

A distribuição de probabilidade estimada nesta pesquisa serve como subsídio para estimação de indenização e de prêmio puro para um seguro agrícola, com enfoque no risco de seca, e para o monitoramento de culturas seguradas.

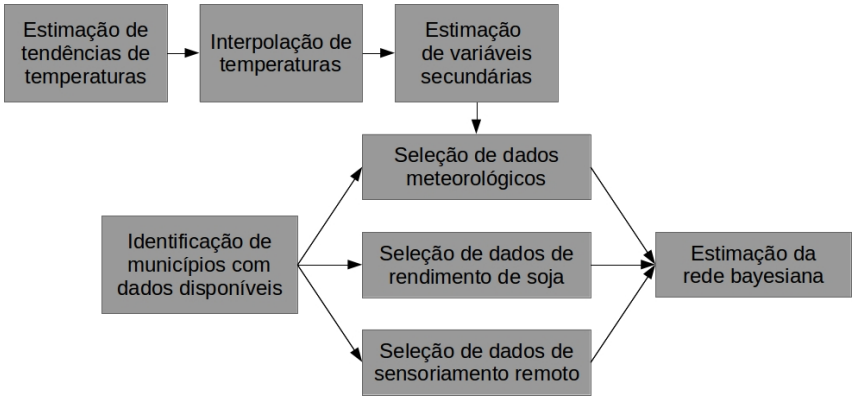


Figura 3.1 – Fluxograma da metodologia (Fonte: autor)

A figura 3.1 apresenta um fluxograma que resume a metodologia usada no trabalho e descrita a diante.

3.2.1 Definição do modelo

O modelo utilizado neste estudo leva em consideração que existem índices que caracterizam o sistema de produção de uma cultura e outros índices que caracterizam o estado de uma cultura implantada. Ambos os índices possuem uma relação com o rendimento no final de seu ciclo. índices meteorológicos podem indicar se o ambiente de produção fornece as condições necessárias para o adequado desenvolvimento e crescimento da cultura. Índices de estado da cultura, como alguns de sensoriamento remoto, indicam o estado de desenvolvimento da cultura, por exemplo a quantidade de biomassa produzida pela cultura.

Em outras palavras, o rendimento de uma cultura é condicionado aos fatores de produção. Esses fatores são constituídos pelo sistema de produção e pelas caracte-

rísticas meteorológicas do ambiente de produção. O sistema de produção no contexto considerado aqui inclui todas as técnicas e tecnologias de nutrição, de manejo e de mecanização, além das características de ocorrência de doenças, de ocorrência de pragas, e edáficas (de solo) do ambiente de produção. O sistema de produção (SP) ainda inclui as características genéticas da cultura. As características meteorológicas do ambiente de produção são caracterizadas de forma indireta por um conjunto de índices meteorológicos (IM). De forma geral, a relação entre o rendimento da cultura (Y) e os fatores de produção pode ser expressa por:

$$Y = f(IM(t,s), SP(t,s), u)$$

em que t representa um momento no tempo e s uma posição no espaço. Nessa representação ambos os índices meteorológicos e o sistema de produção possuem caracterização espaço-temporal. Há um componente u da relação que expressa a variabilidade não explicada.

O estado da cultura, representado indiretamente pelo rendimento, condiciona os índices de estado (IE). Para um certo estado da cultura, os índices de estado assumem valores com diferentes probabilidades de ocorrência. De forma genérica essa relação é expressa por:

$$IE = g(Y, v)$$

considerando um componente v de incerteza.

As variáveis índices meteorológicos, índices de estado da cultura e a variável rendimento foram utilizadas na construção do modelo.

3.2.1.1 Índices meteorológicos

No contexto desta pesquisa, índices meteorológicos podem ser variáveis meteorológicas primárias (como temperatura, precipitação, velocidade e direção do vento, umidade do ar, dentre outras), e variáveis derivadas (soma térmica, precipitação acumulada, evapotranspiração potencial, dentre outras) calculadas a partir das primárias. Os índices meteorológicos caracterizam o ambiente de produção e sugerem se ele está fornecendo as condições necessárias para o desenvolvimento e crescimento da cultura.

Os seguintes índices meteorológicos fazem parte do modelo:

- Precipitação acumulada
- Número de dias do maior período sem chuva
- Soma térmica

A precipitação acumulada reflete a disponibilidade de água para o desenvolvimento da cultura. A disponibilidade de água é importante para manutenção do funcionamento das células e tecidos da planta e fundamental para a assimilação de carbono e transpiração. De forma simplificada, a assimilação de carbono gera biomassa e leva ao ganho de rendimento pela cultura (Pereira; Angelocci; Sentelhas, 2002).

O número de dias do maior período sem chuva reflete a ocorrência de secas durante o ciclo da cultura. Seca se refere a um período longo sem chuva que reduz a disponibilidade de água no solo a um nível que prejudica a planta e reduz seu rendimento.

A soma térmica é um índice derivado das temperaturas diárias a que a cultura é exposta. A soma térmica pode ser calculada por meio das diferenças entre temperatura média diária e uma temperatura referencial chamada temperatura base. A temperatura reflete a velocidade potencial de desenvolvimento e crescimento da cultura, enquanto que a soma térmica indica quão adiantado o desenvolvimento da cultura está. Em alguns modelos a soma térmica é considerada como uma escala de tempo para o desenvolvimento da cultura (Steduto; Raes; Hsiao; Fereres, 2008).

3.2.1.2 Índices de estado da cultura

Como já mencionado, os índices de estado da cultura sugerem a condição de desenvolvimento da cultura, podendo indicar quantidade de biomassa produzida. Os índices provenientes de sensoriamento remoto refletem o comportamento da cultura quanto à interação com a radiação luminosa e à captação de energia luminosa.

O conceito de reflectância será importante nos seguintes parágrafos.

Reflectância de um objeto é a característica intrínseca do mesmo que consiste na proporção de radiação eletromagnética incidente sobre ele que é refletida, para um certo comprimento de onda ou faixa de comprimento de onda (banda). O conjunto que define a relação entre reflectância e comprimento de onda para um certo

objeto denomina-se assinatura espectral do objeto. A reflectância de um objeto ainda depende da direção de incidência da radiação e da direção da reflexão da mesma, sendo assim denominada de reflectância bidirecional. A integração da reflectância bidirecional em todo o domínio das direções denomina-se simplesmente de reflectância. A reflectância é representada por ρ e o comprimento de onda é subscrito (ρ_a corresponde a reflectância no comprimento de onda a).

O NDVI ("Normalized Difference Vegetation Index") é um índice de vegetação derivado das reflectâncias de um objeto nos comprimentos de onda do vermelho (ρ_R) e do infra-vermelho próximo (ρ_{NIR}). Ele é obtido pela relação

$$NDVI = \frac{(\rho_{NIR} - \rho_R)}{(\rho_{NIR} + \rho_R)}$$

O EVI ("Enhanced Vegetation Index") é outro índice de vegetação, mas é derivado das reflectâncias nos comprimentos de onda do vermelho (ρ_R), azul (ρ_B) e infra-vermelho próximo (ρ_{NIR}). O EVI usado nesta pesquisa é obtido pela relação

$$EVI = 2,5\frac{(\rho_{NIR} - \rho_R)}{(\rho_{NIR} + 6\rho_R - 7,5\rho_B)}$$

Tanto o EVI como o NDVI são usados para inferir sobre a biomassa de culturas. Uma aplicação recorrente de NDVI é para a estimação do índice de área foliar IAF (total de áreas de folhas por unidade de área na superfície terrestre), uma medida que pode ser usada para indicar a capacidade de absorção da energia solar pela cultura. Esses índices servem como proxy para biomassa.

O EVI possui certa vantagem em relação ao NDVI quando se trata de culturas com alta biomassa. Segundo Jensen, (2009) e Huete; Didan; Miura; Rodriguez; Gao; Ferreira, (2002) o EVI é mais sensível que o NDVI quando se trata de culturas com alta biomassa.

É importante destacar que esses índices têm capacidade limitada para expressar rendimento agrícola, principalmente por o IAF não definir por si só o rendimento. Outro índice que envolva dados de sensoriamento remoto e até mesmo dados meteorológicos pode ser usado como um índice de estado no modelo, como resultados de modelos agrometeorológicos-espectrais.

Tanto o NDVI como o EVI podem ser obtidos para uma posição no espaço e no

tempo. Para caracterizar o ciclo de uma cultura é possível utilizar a amplitude desses índices ao longo do ciclo. Pupin Mello; Rudorff; Adami; Rizzi; Aguiar; Gusso; Fonseca, (2010) usam a amplitude de EVI em uma rede bayesiana para mapear soja no Estado do Rio Grande do Sul.

Os seguintes índices de estado da cultura fazem parte do modelo:

- Amplitude do NDVI ao longo do ciclo
- Amplitude do EVI ao longo do ciclo

No texto que segue as amplitudes dos índices NDVI e EVI serão tratados como ANDVI e AEVI, respectivamente.

3.2.2 Rede bayesiana

Redes bayesianas são modelos gráficos direcionados acíclicos. Também podem ser denominados modelos hierárquicos bayesianos. Uma explicação sucinta sobre redes bayesianas é data por Mello; Risso; Atzberger; Aplin; Pebesma; Vieira; Rudorff, (2013).

Redes bayesianas podem representar dimensões espaciais e temporais entre variáveis. Nesses casos a rede bayesiana também representa a distribuição de probabilidade conjunta das variáveis de interesse com defasagens temporais e espaciais. Redes bayesianas dinâmicas são usadas em casos que incluem a componente temporal. A possibilidade de modelagem espaço-temporal torna as redes bayesianas uma forma de modelagem adequada ao contexto agrícola.

Redes bayesianas podem ser compostas por variáveis categóricas, numéricas discretas e numéricas contínuas, incluindo ainda combinações desses tipos de variáveis.

Um aspecto muito interessante da modelagem com redes bayesianas é a possibilidade de separação do modelo em módulos. Considerando uma rede bayesiana G=(V,A), composta por um conjunto de vértices V e um conjunto de arcos A, um módulo de G é composto por um subconjunto de vértices e um subconjunto de arcos de G. O modulo por si só é uma rede bayesiana. Essa separação de uma rede bayesiana permite sua parametrização de forma separada (Rasmussen; Madsen; Lund, 2013). Para exemplificação, seja V={a,b,c,d,e}, A={(e,d),(e,c),(d,b),(c,b),(b,a)}, e a dis-

tribuição de probabilidade conjunta desta rede bayesiana pode ser representada da seguinte forma:

$$P(a,b,c,d,e) = P(a,b|c,d,e)P(c,d|e) = P(a,b|c,d)P(c,d|e) \quad (3.1)$$

A distribuição $P(c,d|e)$ pode ser parametrizada com base em um banco de dados de observações (c,d,e) das variáveis. A distribuição $P(a,b|c,d)$ pode ser parametrizada com base em outro banco de dados que contenha observações (a,b,c,d). Posteriormente a distribuição $P(a,b,c,d,e)$ é obtida pela relação 3.1. Isso só é possível pela suposição de independência que G estabelece entre as variáveis (“a”,“b”) e “e” quando estão condicionadas a valores das variáveis “c” e “d”, as quais atuam como interface entre os dois módulos. Esse resultado permite que dois módulos do modelo sejam parametrizados de forma independente e com bancos de dados diferentes, desde que as variáveis aleatórias comuns sejam as mesmas.

No contexto da modelagem de rendimento agrícola, séries históricas de índices meteorológicos podem ser usados na parametrização de um módulo meteorológico (possibilitando previsão para anos seguintes), e bancos de dados de rendimento agrícola condicionado aos fatores de produção que o geraram para parametrizar um módulo de rendimento. Cada um desses módulos pode se parametrizado por um especialista que dispõe dos dados para tanto. De um ponto de vista prático, em uma equipe de especialistas encarregada de modelar um fenômeno complexo cada especialista parametriza a parte do modelo relacionada a sua especialidade, e usando seu banco de dados para isso.

3.2.2.1 Estrutura da rede bayesiana

Neste estudo, a estrutura da rede bayesiana é imposta. O diagrama que representa o modelo e apresentado na figura 3.2. As flechas representam relação de dependência, podendo ainda ser entendida como relação de causalidade. As variáveis com fundo branco são variáveis primárias usadas na derivação dos índices meteorológicos e foram apresentadas apenas para exposição, não sendo usadas diretamente na parametrização do modelo.

Em decorrência da indisponibilidade de dados georreferenciados de rendimentos as análises consideraram as variáveis de interesse agregadas para o nível de mu-

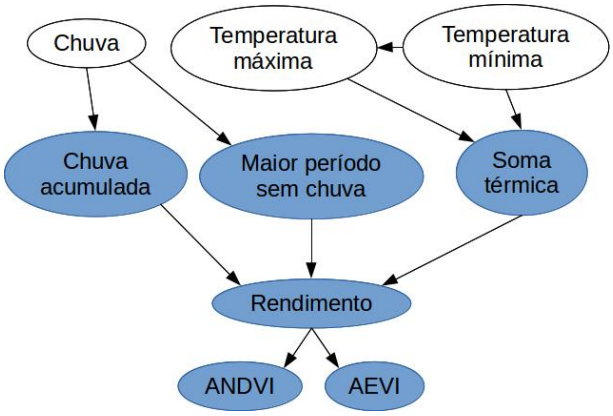


Figura 3.2 – Rede bayesiana representando as relações de independência entre as variáveis do modelo (Fonte: autor)

nicípio.

Assumiram-se formas de distribuição de probabilidade a priori para as variáveis de interesse. Para as variáveis soma térmica, precipitação acumulada, e maior período sem chuva adota-se uma distribuição com suporte inferior zero (variável aleatória positiva). Para as variáveis rendimento e índices de vegetação adotam-se distribuições de probabilidade com suporte inferior e superior, e que podem apresentar assimetria. A interface entre as distribuições foi feita por meio de modelo linear, com componente autorregressivo recomendado por Ozaki, (2009).

O conjunto de dados disponíveis para ajuste do modelo foram separados em duas partições. Uma delas foi usada para ajuste do modelo e a outra para avalização do modelo.

Dois módulos da rede bayesiana foram identificados para a parametrização do modelo. Um desses módulos contém os nós correspondentes apenas a índices meteorológicos (módulo meteorológico). O outro módulo (da cultura) abrange os demais nós, juntamente com as variáveis do módulo meteorológico para interface.

A separação da rede bayesiana nesta pesquisa tem por objetivo obter melhor descrição da distribuição conjunta de probabilidade de cada módulo pelo uso de diferentes bancos de dados para sua parametrização. O módulo meteorológico foi parametrizado pelo banco de dados meteorológicos BDMEP e ANA. O módulo da cultura foi parametrizado pelo banco de dados de rendimentos, sensoriamento remoto, e meteorológico selecionados para o período de interesse.

3.2.2.2 Parametrização

Existem diferentes possibilidades para ajuste de parâmetros de uma rede bayesiana. O ajuste pode ser feito por maximização de verossimilhança, por inferência bayesiana, ou por técnicas de ajuste aproximado como o algoritmo EM ("Expectation Maximization", em inglês, ou Maximização da Esperança, tradução literal para o português). Certos métodos de ajuste são robustos para amostras com dados faltantes, como é o exemplo do algoritmo EM. A possibilidade de ajuste por inferência bayesiana também possibilita ajuste do modelo usando amostra pequena de dados, porém necessita o fornecimento de uma distribuição de probabilidade a priori, e esta pode ser não informativa.

Os parâmetros de cada um dos módulos pode ser ajustado por máxima verossimilhança, de forma independente. Isso não garante que os parâmetros não serão viesados.

3.2.3 Dados

Séries de dados meteorológicos, de rendimento agrícola e de sensoriamento remoto foram utilizados nas análises. Considerando a disponibilidade de dados, a série de rendimento agrícola é a mais curta e, por isso, definiu o período de interesse: de 01/06/2001 a 01/06/2011. Os períodos de dados considerados são apresentados no quadro 3.1.

Quadro 3.1 – Períodos das séries de dados usadas nas análises

Série de dados	Início da série	Término da série
Sensoriamento remoto	01/01/2001	31/12/2011
Meteorologia	01/01/1970	31/12/2011
Rendimento agrícola	01/06/2001	01/06/2011

A seguir os dados utilizados nas análises desta pesquisa são descritos.

3.2.3.1 Dados de rendimento

Os dados de rendimento de culturas usados nesta pesquisa são provenientes da COAMO (Cooperativa Agropecuária Mourãoense), uma cooperativa originada na região de Campo Mourão no Paraná. Eles se referem a rendimentos médios de soja,

por cooperado e para as safras de 2001/2002 a 2010/2011. As lavouras dos cooperados cujos dados foram usados na pesquisa situam-se em municípios do Paraná e Santa catarina. A única referência espacial disponível é o município em que as lavouras se situam. Tais municípios são apresentados no mapa da figura 3.3.

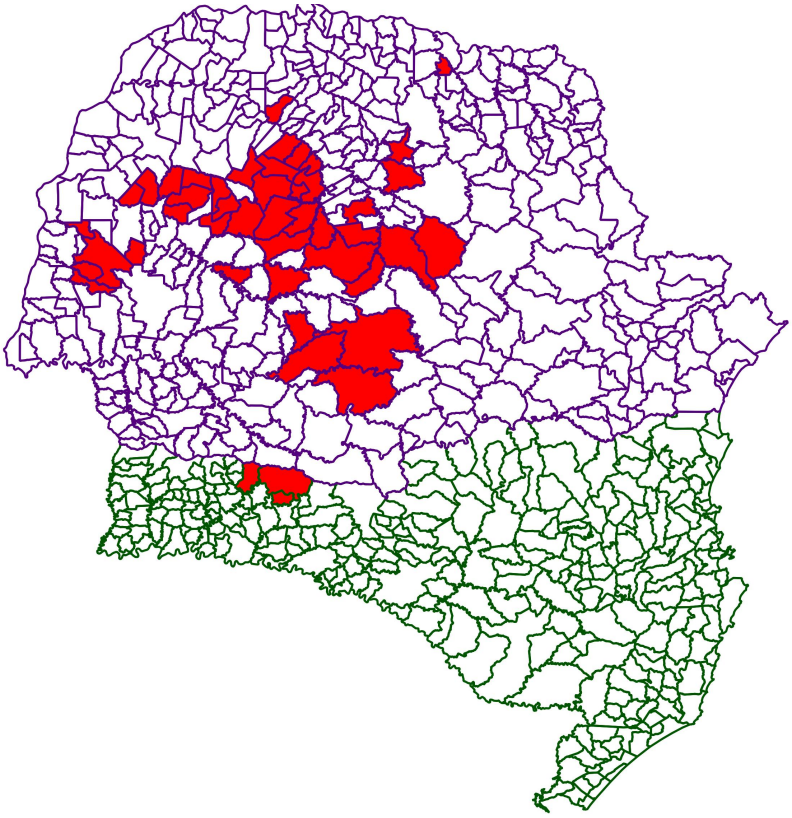


Figura 3.3 – Mapa apresentando municípios de interesse para o estudo (Fonte: autor)

3.2.3.2 Dados de sensoriamento remoto

Os dados de sensoriamento remoto são oriundos do sensor MODIS (“Moderate-Resolution Imaging Spectroradiometer”), um projeto de aquisição e processamento de dados de sensoriamento remoto orbital, que disponibiliza imagens de satélite com frequência de até 15 minutos para alguns produtos. Os produtos de dados gerados pelo projeto são disponibilizados gratuitamente e regularmente desde o ano 2000, e abrangem o Brasil em sua totalidade.

Um dos produtos de dados usados na pesquisa é o MOD13Q1.005, com o qual são disponibilizados dados de reflectância das superfície nas bandas do vermelho, azul, infravermelho próximo e infravermelho médio, além dos índices de vegetação

NDVI e EVI. A resolução temporal associada a esse produto é de 16 dias, e a resolução espacial é de 250 metros.

O outro produto de dados usado é o MOD11A2.005, pelo qual são disponibilizadas estimativas de temperatura superficial. A resolução temporal desses dados é de 8 dias, e a espacial é de 1 km.

O intervalo de datas usado para selecionar os dados MODIS corresponde ao período de disponibilidade de dados de rendimento da cultura da soja, de 01/06/2001 a 01/06/2011.

Os dados MODIS foram obtidos por meio do portal “Data Pool”, do “NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota”.

3.2.3.3 Dados meteorológicos

Os dados meteorológicos usados no projeto são oriundos do BDMEP (Banco de Dados Meteorológicos para Ensino e Pesquisa), um banco de dados com séries históricas de dados meteorológicos disponibilizados pelo INMET (Instituto Nacional de Meteorologia), e da ANA (Agência nacional de águas). O banco de dados da ANA possui registros de precipitação para um grande número de estações. O BDMEP disponibiliza dados de temperaturas máxima e mínima diárias, mas com número de estações inferior ao da ANA.

A distribuição espacial das estações meteorológicas usadas neste estudo é apresentada na figura 3.4.

3.2.4 Estimação de dados meteorológicos

A indisponibilidade de estações meteorológicas para todo município contemplado por essa pesquisa torna a interpolação espacial dos dados meteorológicos necessária. No caso da variável temperatura, a interpolação por cokrigagem foi adotada. A variável precipitação foi extrapolada por associação entre o município e a estação meteorológica mais próxima do mesmo.

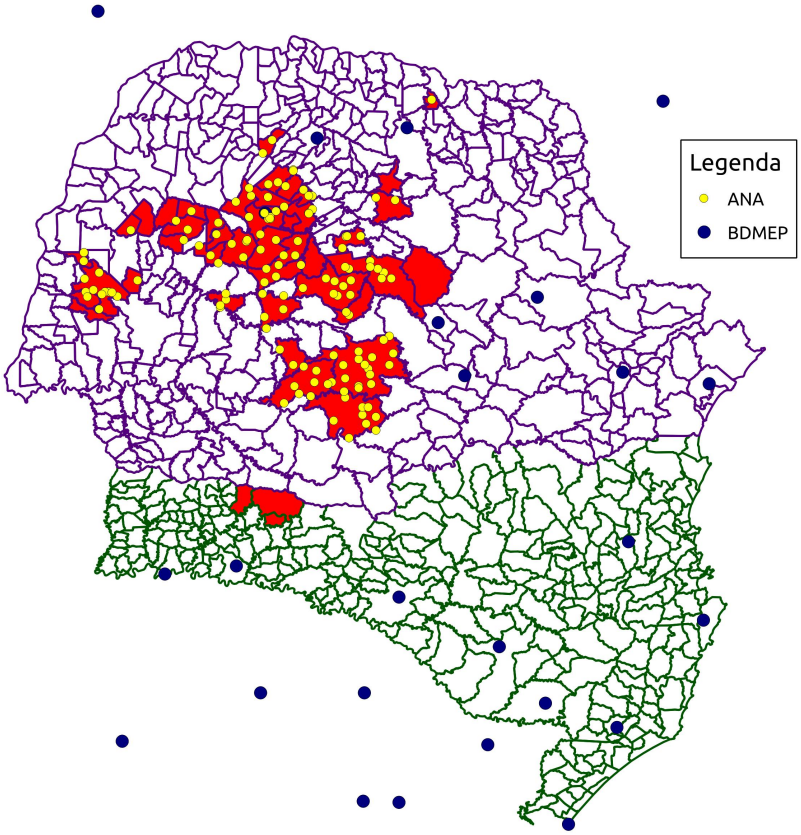


Figura 3.4 – Mapa de estações meteorológicas usadas no estudo (Fonte: autor)

3.2.4.1 Temperatura

A interpolação espacial de variáveis meteorológicas faz sentido quando é possível assumir que há dependência espacial entre as observações disponíveis dos dados e assumir uma relação entre a variável a ser estimada e covariáveis disponíveis. É sabido que a variável temperatura possui essas características.

A krigagem é um método de interpolação que considera que a variável de interesse possui dependência espacial, e que objetos próximos tendem a ser mais parecidos que objetos distantes uns dos outros (Kilibarda; Hengl; Heuvelink; Gräler; Pebesma; Tadi'c; Bajat, 2014; Wu; Li, 2013).

Assume-se o modelo (Diggle; Ribeiro, 2007, p. 37):

$$\hat{T} - \bar{T} = \sum_{i=1}^{n} \omega_i * (T_i - \bar{T})$$

em que $\hat{T}$,$\bar{T}$, T_i correspondem a temperatura estimada, a tendência da temperatura e a observação de temperatura no ponto i, respectivamente. w_i é a ponderação

dada a cada observação na estimação e expressa a dependência espacial entre as variáveis.

A tendência da variável a ser interpolada pode ser estimada por meio de co-variáveis. O modelo usado para tanto pode ser determinístico, com fundamentação física, ou empírico. É comum na literatura citar-se o uso de coordenadas geográficas, distância do litoral e elevação em relação ao nível do mar no processo de interpolação de temperatura. No entanto essas variáveis não captam características de microclima relacionadas a temperatura.

Proxies da temperatura do ar podem ser usadas no processo de interpolação. Estimativas de temperatura da superfície terrestre por meio de sensoriamento remoto constituem um caso viável, como é verificado em Hengl; Heuvelink; Tadie; Pebesma, (2012). A estimativa de uma tendência sazonal por meio dessas proxies é interessante para que não haja necessidade da compatibilidade de comprimentos das séries de dados da variável proxy e da variável a ser interpolada.

Seja T a variável que representa a temperatura do ar, e Z a variável que representa a temperatura da superfície terrestre, então assume-se o modelo para a relação entre as duas:

$$T = a + bZ + e$$

em que e representa o erro do modelo e a e b representam os coeficiente do modelo linear. A tendência da variável T pode, então, ser definida por $E(T|s,t) = a + bE(Z|s,t)$. Considerando que os dados de temperatura estimados por sensoriamento remoto (produtos MODIS MOD11A2.005 e MY11A2.005) são disponibilizados de 8 em 8 dias, e de forma padronizada entre os anos, é possível estimar a tendência para cada data no ano. Esta é uma forma de captar a sazonalidade da variável temperatura.

É preciso destacar que os dados de temperatura dos produtos MODIS MOD11A2.005 e MY11A2.005 são disponibilizados para dois momentos do dia: LST_Day_1km e LST_Night_1km (diurno e noturno, respectivamente). Como os momentos de coleta dos dois produtos não coincidem, 4 momentos diferentes de coleta de dados estão disponíveis para as análises.

Este estudo considerou as estimativas de temperatura por sensoriamento remoto para a geração das tendências a serem usadas na interpolação por cokrigagem

da temperatura agregada para 8 dias. A agregação consistiu no cálculo das médias das variáveis para o período de 8 dias coincidente com o do produto MOD11A2.005 (e MYD11A2.005), para cada estação meteorológica considerada.

A função “krige” do pacote gstat do R foi usado para as interpolações. A tendência foi estimada com apenas uma das variáveis de temperatura do MODIS: aquela cujo R^2 com a temperatura média foi maior e ainda possui relação linear significativa pela estatística t ao nível de significância de 5%.

Tabela 3.2 – Resultados dos ajustes de regressões lineares das variáveis temperatura máxima

Estimate	Std. Error	t value	Pr(>|t|)
17,1739	0,0617	278,18	0,0000
0,6083	0,0039	155,01	0,0000

Tabela 3.3 – Resultados dos ajustes de regressões lineares das variáveis temperatura mínima

Estimate	Std. Error	t value	Pr(>|t|)
7,2668	0,0655	110,86	0,0000
0,5231	0,0041	126,20	0,0000

Tabela 3.4 – Resultados dos ajustes de regressões lineares das variáveis temperatura média

Estimate	Std. Error	t value	Pr(>|t|)
12,3180	0,0579	212,80	0,0000
0,5598	0,0037	152,64	0,0000

3.2.4.2 Precipitação

A variável meteorológica precipitação possui características diferentes da temperatura. Embora haja dependência espacial entre a precipitação que ocorre um um lugar e a precipitação que ocorre em outro, há também o fenômeno de chuva localizada. Diz-se que ocorre chuva localizada quando há uma variação brusca de quantidade de chuva que ocorre em lugares próximos, podendo em um lugar chover bastante e em outro (a 2km de distância, por exemplo) não chover nada. Se a precipitação é interpolada, cria-se um superfície com variação gradual dos volumes de chuva e a estimação de variáveis como o comprimento do maior período sem chuva é prejudicado para regiões com dados interpolados.

Quando a disponibilidade de estações meteorológicas é grande, pode-se assumir que os dados a serem estimados de precipitação para um ponto no espaço corresponde às observações da variável na estação meteorológica mais próxima. Essa abordagem foi adotada neste trabalho. A estimativa de precipitação para um determinado município é assumida como sendo a observação da variável na estação meteorológica mais próxima do centroide do município.

3.2.5 Seleção dos dados espaço-temporais

Os dados de sensoriamento remoto e meteorológicos são usados na parametrização da rede bayesiana que representa a distribuição de probabilidade de rendimento (da produção) da cultura da soja ao nível de agregação de município. Entretanto, para que o modelo represente a cultura é necessário que as amostras de dados sejam, da mesma forma, representativas. Para tornar esses dados representativos, há necessidade de seleção dos dados considerando as dimensões espacial (onde a cultura é cultivada) e temporal (quando a cultura é cultivada). A dimensão temporal requer que a amostragem considere o ciclo da cultura, ou seja, o início e término do ciclo. Isto é possível por meio de um algoritmo de identificação de ciclo da cultura de verão. A dimensão espacial requer a consideração da localização das lavouras e, para tanto, um mapa de culturas é necessário (máscara de culturas).

A classificação e identificação de ciclo da cultura são usados na seleção dos dados de sensoriamento remoto e dados meteorológicos. O porquê que justifica o uso dessa metodologia é que a parametrização de modelos que estimam rendimento com base em variáveis meteorológicas e posteriormente estabelecem relação com variáveis de sensoriamento remoto deve ser condizente com as condições do ambiente em que a cultura se desenvolveu. Essa importância é ainda maior quando as estimativas se referem a grupos de indivíduos (regiões).

A série histórica considerada na parametrização do módulo meteorológico é mais longa que a série histórica de dados de sensoriamento remoto. Assim não foi possível estimar os períodos de cultivo para todos as safras compreendidas por aquela série histórica. A fim de manter coerência das análises, um momento padrão para início e término de ciclo estimado por meio de média simples dos resultados de identificação de ciclo e o mapa de culturas anuais foram adotados para aquelas safras sem

disponibilidade de dados de sensoriamento remoto.

3.2.6 Transformação de variáveis explicativas

Até o momento apresentou-se o método de obtenção de variáveis primárias, mas o modelo é composto por índices derivados dessas variáveis. A obtenção de cada índice segue métodos específicos que são apresentados a seguir.

Algumas características dos índices são desejáveis. Uma delas é a de representação do ciclo da cultura como um todo. É sabido que uma mesma cultura pode ter comportamento de desenvolvimento diferente em função do sistema e ambiente de produção. A representatividade adequada do índice é necessária para que a comparação entre diferentes cultivos da mesma cultura (soja) seja possível.

Para os índices que necessitam o estabelecimento de um período de interesse para sua estimação, os resultados da identificação de ciclo da cultura foram usados para obtenção desses períodos.

3.2.6.1 Soma térmica

Retomando o conceito de soma térmica (GD, medida em graus dia) seu cálculo parte da suposição da existência de uma temperatura base (Tb), abaixo da qual o desenvolvimento da cultura é considerado nulo. Uma estimativa de temperatura base para a cultura da soja é $Tb = 14°C$ (Camargo; Brunini; Miranda, 1987).

A soma térmica pode ser estimada de diversas formas (Caicedo; Torres; Cure, 2012), sendo que cada forma assume um comportamento para a variável temperatura e é adequada para determinada resolução de tempo das observações. Nesta pesquisa adotou-se uma das formas mais simples de estimação. Assumindo n períodos para os quais deseja-se estimar a soma térmica, e que $t \in \{1, 2, ..., n\}$ é um desses períodos. A soma térmica é estimada por:

$$GD = \sum_{t=1}^{n} max(0, Tmed_t - Tb) \tag{3.2}$$

Em que $Tmed_t$ representa a temperatura média do período t, e pode ser estimada pela média aritmética entre a temperatura máxima e a temperatura mínima ($Tmed = (Tmax + Tmin)/2$). Neste estudo, $Tmed$ corresponde ao resultado da inter-

polação da temperatura descrita na seção 3.2.4.1.

A soma térmica foi agregada ao nível municipal pelo cálculo da média das estimativas da variável referentes a regiões cultivadas com soja dentro de cada município de interesse. A “Malha Municipal Digital 2007” disponibilizado pelo IBGE (`http://www.ibge.gov.br/home/geociencias/default/_prod.shtm`\#TERRIT, acesso em 10/02/2015) foi usado para tanto.

No processo de seleção dos dados para estimação das estatísticas de referência, poucas observações foram encontradas para o município de Altamira do Paraná (geocodigo, 4100459). Por esse motivo esse município foi desconsiderado das análises.

3.2.6.2 Precipitação acumulada

A precipitação acumulada á calculada pela soma dos valores de precipitação ocorridos em um certo período. Se t é um dos n subperíodos do período de interesse, então a precipitação acumulada Pa é obtida por:

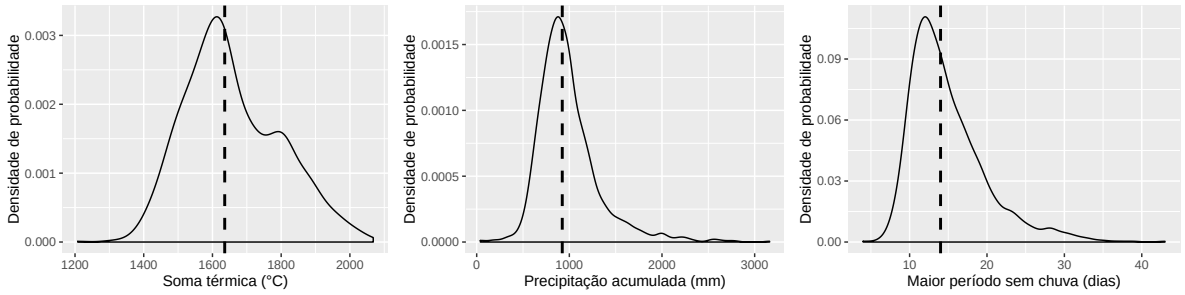
$$Pa = \sum_{t=1}^{n} P_t$$

Sendo P_t a precipitação verificada no subperíodo t. O período de interesse foi definido por meio da média aritmética dos inícios e términos de ciclo das áreas cultivadas com soja em cada município.

Para os anos sem disponibilidade de dados de sensoriamento remoto para estimação do mapa de culturas e ciclo das culturas, o mapa de áreas cultivadas e os inícios e términos médios históricos de ciclo da soja (estimados com base nos anos com disponibilidade desses dados) foram usados. As médias de início e término de ciclo foram calculadas considerando as áreas cultivadas com soja.

3.2.6.3 Maior período sem chuva

O maior período sem chuva em um ciclo de cultura é obtido estimando os comprimentos dos intervalos entre as chuvas ocorridas, e tomando-se o maior deles. A definição dos períodos de interesse é a mesma usada para estimar a precipitação acumulada.

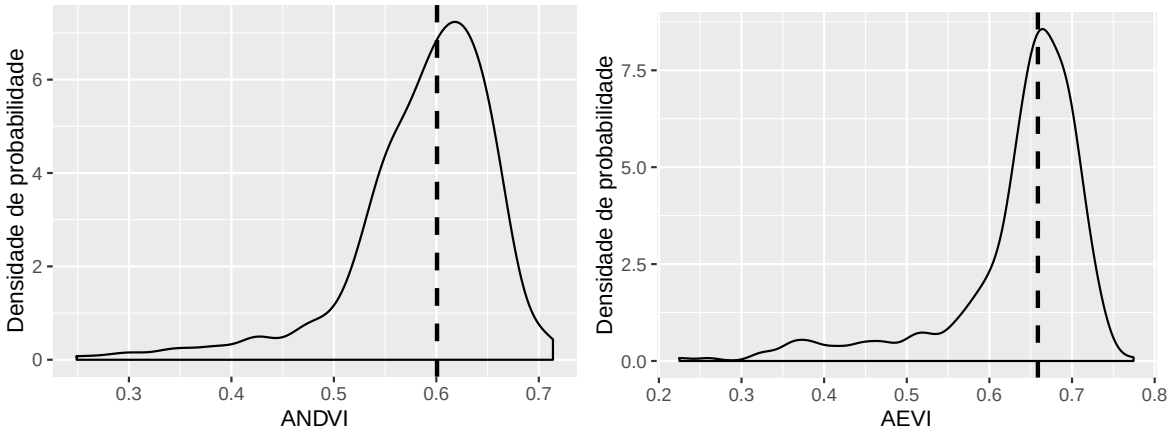


(a) Soma térmica (b) Precipitação acumulada (c) Maior período sem chuva

Figura 3.5 – Densidades de probabilidades estimadas por kernel das variáveis do ambiente de produção (Fonte: autor)

3.2.6.4 ANDVI e AEVI

Como já mencionado anteriormente, as amplitudes dos índices NDVI e EVI (ANDVI e AEVI, respectivamente), são obtidos pela diferença entre o máximo e o mínimo valor de cada índice no ciclo da cultura.



(a) ANDVI (b) AEVI

Figura 3.6 – Densidades de probabilidades estimadas por kernel dos índices de estado da cultura (Fonte: autor)

3.2.7 Rendimento relativo da cultura

Goodwin; Ker, (1998) sugere a correção da variável rendimento para remoção de tendência no processo de ajuste de uma distribuição de probabilidade para a variável. Segundo o autor, esse procedimento é especialmente importante para remover o problema de heterocedasticia encontrado em seu estudo. Nesse caso, pressupõe-se que cada produtor agrícola adota um sistema de produção que associa-se a uma tendência de aumento de rendimento de produção ao longo dos anos. Para tornar os

dados de diferentes produtores compatíveis, em lugar dos dados de rendimento os desvios dessa variável com relação a tendência são usados nas análises. A tendência do rendimento pode ser obtida por uma regressão linear simples, em que a variável resposta é rendimento e a variável explicativa é o ano da safra. Considerando Y o rendimento, e $E(Y|t)$ a tendência condicionada à safra de interesse t, os desvios relativos DR são obtidos da seguinte forma:

$$DR_t = \frac{Y_t - E(Y|t)}{E(Y|t)}$$

Essa medida, adimensional, indica qual a variação relativa (em %) a esperança da variável.

Alternativamente a este procedimento, é possível adotar um modelo para o ajuste da distribuição de probabilidade do rendimento para o qual se estime a esperança e a variância, ou algum parâmetro de precisão, condicionais conjuntamente. Este é o método adotado neste estudo. A regressão beta (Grün; Kosmidis; Zeileis, 2012) possibilita estimar a esperança e a precisão condicionais da distribuição beta para uma variável com domínio (0,1). Considerando isso, neste estudo o rendimento de produtividade relativo a uma produtividade de referência 4800 kg.ton^{-1} (80 sacas.ha^{-1}) agrupado ao nível municipal foi adotada como variável resposta na regressão beta. Os dados foram agrupados ao nível municipal por meio de média ponderada pela área cultivada de cada produtor:

$$YR_{mt} = \frac{1}{\sum_{i \in m} A_i} \sum_{i \in M} \frac{Y_{it}}{80} A_i$$

em que M é o conjunto de produtores com lavouras em um certo município m, A_i é a área cultivada por um produtor i, Y_{it} é a produtividade média do produtor i no momento t, e YR_{mt} é o rendimento relativo para o município m no momento t.

É importante destacar que adotando esse procedimento a variável resposta é o rendimento relativo municipal.

3.2.8 Ajuste dos modelos

Os parâmetros das distribuições de probabilidade que compõe a rede bayesiana foram ajustados por meio de máxima verossimilhança. Cada módulo foi tratado de

forma independente, para enfatizar a modularidade da rede.

Os parâmetros dos índices meteorológicos foram ajustados usando o pacote **fitdistrplus** do R, que usa o método de máxima verossimilhança.

As variáveis rendimento relativo e índices de estado foram modeladas por regressão beta, cujos parâmetros foram ajustados por meio do pacote **betareg** do R.

Tomando a forma reparametrizada da distribuição beta apresentada pela equação 3.3, a regressão beta é usada para ajustar os parâmetros β e θ das funções link apresentadas nas equações 3.4 e 3.5 pelo método da máxima verossimilhança.

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{3.3}$$

$$\mu = \frac{e^{\beta X}}{1+e^{\beta X}} \tag{3.4}$$

$$\phi = e^{\theta X} \tag{3.5}$$

μ corresponde a esperança da distribuição, e ϕ a um parâmetro de precisão. A função de densidade de probabilidade apresentada em 3.3 é facilmente obtida a partir da forma tradicional da função de densidade tradicional da distribuição beta, expressão 3.6.

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{3.6}$$

Para isso, assumem-se as seguintes relações:

$$\begin{aligned} \mu &= \tfrac{p}{(p+q)} \\ \phi &= p+q \\ q &= \phi(1-\mu) \\ p &= \phi\mu \end{aligned} \tag{3.7}$$

O ajuste de parâmetros da regressão beta requer que a variável explicada seja definida apenas no domínio (0,1). O rendimento relativo satisfaz essa suposição se considerar que o rendimento da cultura nunca é zero e não atinge seu potencial máximo. No caso dos índices de estado não é possível garantir que eles permaneçam no domínio (0,1), mas observando suas densidades de probabilidade aproximadas (figura

3.6) ainda é possível assumir essa suposição.

Distribuições gama são ajustadas pelo método da máxima verossimilhança, usando a função *fitdist* do pacote *fitdristrplus* do R.

3.2.9 Simulação de aplicação dos resultados

A fim de demostrar o uso da rede bayesiana no contexto de precificação de prêmio e avaliação de risco, alguns cenários para valores das variáveis que compõe o modelo são estabelecidos e uma avaliação de risco comparativa é feita.

Alguns cenários representam diferentes condições de desenvolvimento da cultura simulando diferentes valores para as variáveis que descrevem o ambiente de produção (variáveis meteorológicas). Esta simulação representa a situação em que a cultura ainda não foi implantada e dispõe-se de expectativas para os valores de tal variáveis, tipica do momento em que avalia-se o risco e taxas de prêmio são definidas para contratos de seguro.

Em outros cenários supõe-se que as variáveis de ambiente de produção foram observadas, a cultura se desenvolveu, e dados de ANDVI e AEVI estão disponíveis. Estes cenários representam a situação em que um contrato de seguro foi firmado e há interesse em monitoramento desse contrato para estimação da ocorrência de sinistros.

Definição 1. Um método *Markov chain Monte Carlo (MCMC)* para simulação de uma distribuição f é qualquer método que produz uma cadeia de Markov ($X^{(t)}$) ergótica cuja distribuição estacionária é f. [1] (Robert; Casella, 2004).

3.2.9.1 Simulação com MCMC

O processo de simulação por MCMC pode variar em função do algoritmo de aceitação e rejeição adotado. Neste estudo o algoritmo de Metropolis-Hastings foi adotado. No processo, parte-se de uma distribuição de probabilidade f da qual se deseja obter uma amostra. Em seguida define-se uma função de densidade de probabilidade condicional (distribuição instrumental, de proposta, ou de salto) $q(y|x)$. É requerido que seja possível gerar uma amostra de q facilmente, e que q seja explici-

[1] Tradução do original: "A *Markov chain Monte Carlo (MCMC) method* for the simulation of a distribution f is any method producing an ergodic Markov chain ($X^{(t)}$) whose stationary distribution is f ".

tamente disponível (até uma constante multiplicativa independente de x) ou simétrica ($q(y|x) = q(x|y)$), e que $f(y)/q(y|x)$ seja menor que uma constante independente de x.

O algoritmo Metropolis-Hastings, associado a f e q, produz uma cadeia de Markov ($X(i)$) por meio do seguinte processo (Robert; Casella, 2004):

Dado x_i,
1. Gerar $Y_i \sim q(y|x_i)$. 2. Tomar

$$X_{i+1} = \begin{cases} Y_i & \text{com probabilidade } \rho(x_i, Y_i), \\ x_i & \text{com probabiliade } 1 - \rho(x_i, Y_i), \end{cases}$$

em que

$$\rho(x, y) = \min \left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}$$

Algoritmo 2: Metropolis-Hastings

O valor $x^{(0)}$ inicial deve ser tal que $f(x_0) > 0$.

Citando diretamente o que Gelman; Carlin; Stern; Dunson; Vehtari; Rubin, (2013) expõe, uma boa distribuição de salto tem as seguintes propriedades:

- É fácil de se amostrar de q
- O cálculo de $\rho(x, y)$ é fácil
- Cada salto percorre uma distância razoável no espaço de Y (senão o passeio aleatório move muito lentamente)
- Os saltos não são rejeitados muito frequentemente (senão o passeio aleatório perde muito tempo sem se mover)

Diferentemente do algoritmo Metropolis, o Metropolis-Hastings não requer que a função de densidade da distribuição de salto seja simétrica.

3.3 Resultados

A seguir resultados finais são apresentados. Foco é dado as dificuldades encontradas e possíveis problemas relacionados a metodologia usada.

3.3.1 Interpolação dos dados meteorológicos

O ajuste do modelo para estimação das tendências requereu associação dos dados de temperatura de sensoriamento remoto aos dados das estações meteorológicas. Um erro associado aos dados das estações meteorológicas é decorrente da

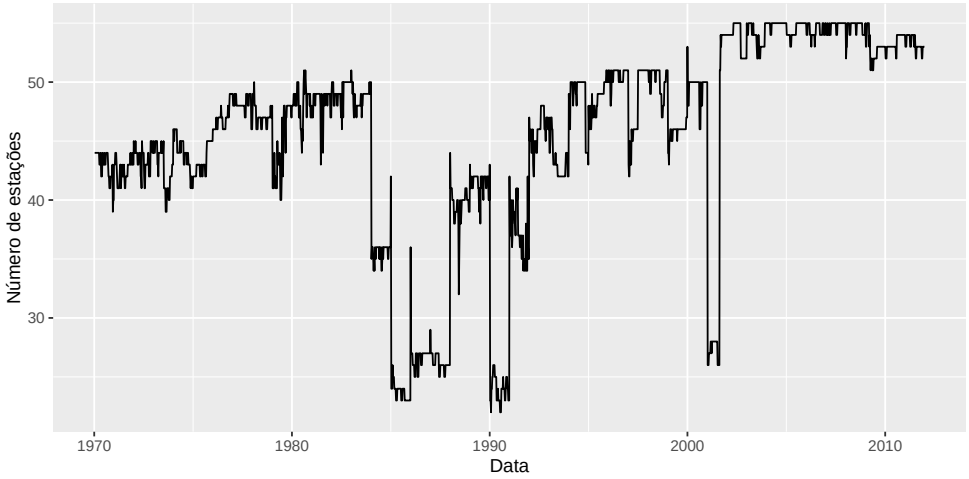


Figura 3.7 – Número de estações com dados disponíveis para o período de 01/01/1970 a 31/12/2011 na região de interesse (Fonte: autor)

precisão das coordenadas geográficas das estações disponíveis ao público. No caso do banco de dados BDMEP as coordenadas geográficas (em graus) das estações possuem precisão de duas casas decimais. Isso representa uma possibilidade de variação da posição de mais de 1 km. Levando em consideração a resolução espacial das imagens com os dados de temperatura de sensoriamento remoto, que é de 1 km, há possibilidade de associação dos dados de forma errada entre as fontes. Verifica-se ainda problema comum com a referência espacial para dados meteorológicos, que é a indisponibilidade da informação de projeção e do datum usados no registro da coordenada geográfica da estação meteorológica.

Outra fonte de erro de associação também está relacionado a resolução espacial das imagens MODIS. Um pixel com resolução de 1km trata todos os objetos que estão na área que ele representa como sendo homogêneos. No entanto é sabido que isso não ocorre, e que o dado obtido não corresponde necessariamente à média das respostas desses objetos.

A figura 3.7 retrata o número de estações com dados disponíveis ao longo do tempo. Verifica-se que por volta do ano 1980 ocorre pequena variabilidade de disponibilidade de dados, mas no período entre 1985 e 1988 há baixa disponibilidade de dados de estações meteorológicas. Há uma recuperação entre 1988 e 1990, mas nos dois anos seguintes também verifica-se baixa disponibilidade de dados. A partir de 1993 verifica-se uma tendência de maior disponibilidade de dados. A observação dessa informação é interessante para ter noção das limitações das estimativas de dados ao longo do tempo decorrentes de disponibilidade de dados. Nesse sentido, espera-se

que em anos com mais dados as estimativas estejam mais próximas dos valores reais.

Figura 3.8 – Gráfico de dispersão dos dados de temperatura máxima e mínima, e de temperatura dos produtos MOD11A2.005 e MYD11A2.005, referentes a posição da estação com código 83692 (Fonte: autor)

A figura 3.8 retrata a relação existente entre os dados de temperatura máxima e mínima com os dados de temperatura da superfície terrestre dos produtos MOD11A2.005 e MYD11A2.005, para uma das estações meteorológicas considerada (código 83692 no BDMEP). Mesmo com diferente dispersão de dados nos gráficos é possível perceber relação positiva entre a temperatura de estações meteorológicas e

a temperatura da superfície terrestre. Tanto para temperatura máxima como para a mínima a dispersão é maior quando o dado de temperatura de sensoriamento remoto é coletado durante o dia (LST_Day_1km). Visualmente não há diferença entre as dispersões ao comparar os dados do produto MOD11A2.005 com os do MYD11A2.005. Com o objetivo de avaliar de forma mais objetiva da relação entre as variáveis, os resultados das regressões lineares entre elas são apresentados na tabela 3.5.

Tabela 3.5 – Resultados dos ajustes de regressões lineares das variáveis temperatura máxima, temperatura mínima e temperatura média em função das temperaturas estimadas por sensoriamento remoto.

Y	X	Estimativa	Erro padrão	Valor t	Pr($>\|t\|$)
Temperatura máxima	(Intercept)	12,67	0,0917	138	0,0e+00
Temperatura máxima	mod11a2_temp_day	0,51	0,0034	151	0,0e+00
Temperatura máxima	(Intercept)	14,88	0,0637	233	0,0e+00
Temperatura máxima	mod11a2_temp_night	0,74	0,0040	185	0,0e+00
Temperatura máxima	(Intercept)	12,52	0,1169	107	0,0e+00
Temperatura máxima	myd11a2_temp_day	0,45	0,0038	118	0,0e+00
Temperatura máxima	(Intercept)	15,71	0,0684	230	0,0e+00
Temperatura máxima	myd11a2_temp_night	0,71	0,0045	159	0,0e+00
Temperatura mínima	(Intercept)	3,33	0,0976	34	2,7e-249
Temperatura mínima	mod11a2_temp_day	0,43	0,0036	121	0,0e+00
Temperatura mínima	(Intercept)	3,86	0,0624	62	0,0e+00
Temperatura mínima	mod11a2_temp_night	0,71	0,0039	183	0,0e+00
Temperatura mínima	(Intercept)	4,34	0,1258	34	1,7e-253
Temperatura mínima	myd11a2_temp_day	0,34	0,0041	84	0,0e+00
Temperatura mínima	(Intercept)	4,58	0,0648	71	0,0e+00
Temperatura mínima	myd11a2_temp_night	0,69	0,0042	164	0,0e+00
Temperatura média	(Intercept)	8,03	0,0862	93	0,0e+00
Temperatura média	mod11a2_temp_day	0,47	0,0031	148	0,0e+00
Temperatura média	(Intercept)	9,42	0,0550	171	0,0e+00
Temperatura média	mod11a2_temp_night	0,72	0,0034	210	0,0e+00
Temperatura média	(Intercept)	8,41	0,1122	75	0,0e+00
Temperatura média	myd11a2_temp_day	0,40	0,0037	108	0,0e+00
Temperatura média	(Intercept)	10,19	0,0583	175	0,0e+00
Temperatura média	myd11a2_temp_night	0,70	0,0038	184	0,0e+00

Tabela 3.6 – Valores de R^2 obtidos nas regressões das variáveis meteorológicas de temperatura e as variáveis estimadas por sensoriamento remoto.

	Temperatura máxima	Temperatura mínima	Temperatura média
mod11a2_temp_day	0,47	0,36	0,46
mod11a2_temp_night	0,57	0,56	0,64
myd11a2_temp_day	0,39	0,25	0,36
myd11a2_temp_night	0,55	0,56	0,62

Tomando os resultados apresentados na tabela 3.5, verifica-se que todos os modelos são significativos pela estatística t. Optou-se pelos dados do produto MOD11A2.005 e os dados coletados a noite para a estimativa da tendência a ser usada

na interpolação, pelo motivo de esta variável proporcionar os maiores valores de R^2 dentre as estimativas. Esses valores de R^2 são apresentados na tabela 3.6.

3.3.2 Seleção de dados

O uso de estimativas de mapas e de ciclos de culturas têm grande apelo na seleção de dados. No entanto, seu potencial é pleno quando há disponibilidade de dados de sensoriamento remoto.

A seleção de dados em anos que não dispõe de dados de mapas e ciclos de culturas exige a estimativa dessas informações a partir dos anos com esses dados disponíveis. Um erro gerado pela estimativa pode reduzir os benefícios dessa seleção. No entanto a comparação de dois casos é necessária para verificar os ganhos com o uso da metodologia. Um desses casos é a suposição de início e término padrões usados na literatura. O outro caso é o uso das estimativas de início e término por meio da metodologia proposta. A melhoria de estimativas do modelo de rendimento pode ser usada como indicador da contribuição da seleção de dados por mapa e ciclo das culturas.

3.3.3 Variáveis explicativas

A partir desta parte da pesquisa os resultados preliminares ainda são escassos. Isso se deve ao fato de a seleção de dados meteorológicos e de sensoriamento remoto por mapa e ciclo das culturas ter requerido melhorias, em especial no mapa de culturas, o que levou atraso na execução do trabalho.

As figuras 3.9b, 3.9a e 3.9c apresentam gráficos das densidades de probabilidades estimadas por kernel para as variáveis que compõe a rede bayesiana, com exceção dos índices ANDVI e AEVI.

As densidades de probabilidade sugerem que as distribuições de probabilidade das variáveis meteorológicas apresentam assimetria a direita. Sabendo que elas possuem suporte inferior, sugere-se o uso de prioris de variáveis positivas e que permitam assimetria a direita para uso na rede bayesiana.

No caso da variável rendimento relativo, cuja estimativa da densidade de probabilidade é apresentada na figura 3.9d, a aparente assimetria é a esquerda. Tomando a suposição de que a variável rendimento (absoluto) possui suporte inferior (zero) e

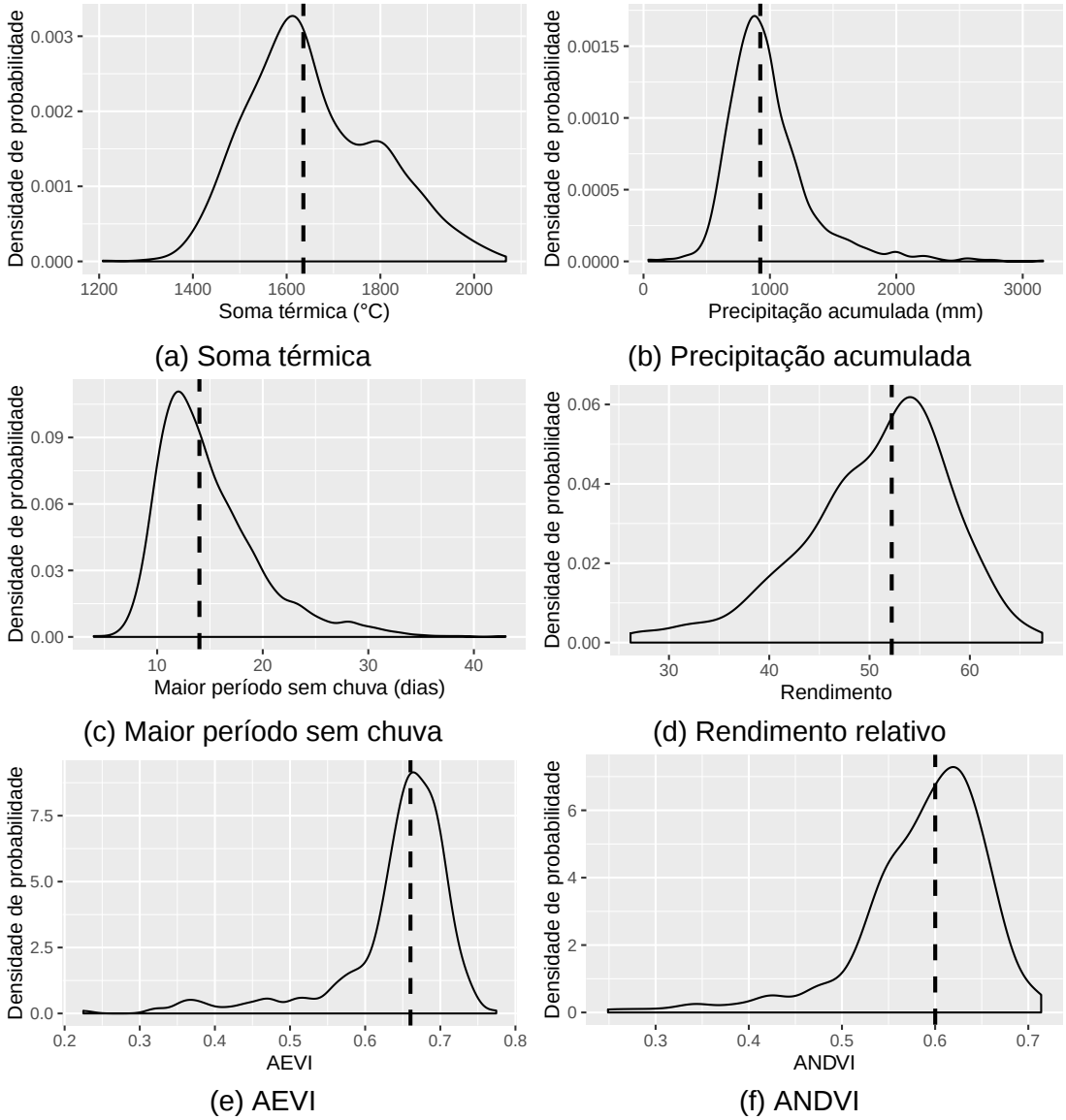


Figura 3.9 – Densidades de probabilidade estimadas por kernel das variáveis do modelo (Fonte: autor)

um suporte superior (Y_{max}), então o rendimento relativo também possui suporte inferior e superior (0,1). Assim, sugere-se que a distribuição a priori a ser usada na rede bayesiana tenha ambos os suportes e permita assimetria a esquerda.

As densidades de probabilidades dos índices ANDVI e AEVI aparentemente possuem assimetria a esquerda. Considerando que o índice NDVI possui suportes inferior e superior, então ANDVI também possui. Sabe-se que o índice EVI não possui tais suportes, mas observando o gráfico 3.9e ainda é plausível assumir que AEVI possui suporte inferior e superior. Assim também sugere-se assumir prioris com os suportes inferior e superior, e que permitam assimetria.

3.3.4 Rede bayesiana

A discretização de variáveis é comumente usadas no contexto de aplicação de redes bayesianas. Neste trabalho tentou-se discretizar as variáveis explicativas e resposta, porém mesmo com a opção de número pequeno de intervalos verificou-se grande número de intervalos sem observações representantes. A consequência disso é a presença de muitos intervalos com probabilidade estimada próximo a zero.

A discretização das variáveis daria grande facilidade para a estimação de parâmetros da rede bayesiana. A estimação por máxima verossimilhança ocorreria pela contagem de ocorrências dos valores das variáveis em cada intervalo. Por fim, optou-se por trabalhar com distribuições de probabilidade de variáveis contínuas.

Um dos motivos para as variáveis AEVI e ANDVI não explicarem bem o rendimento pode dever-se a desconsideração da retirada do efeito da cobertura do solo. A cobertura do solo influencia o valor mínimo de EVI e NDVI.

3.3.4.1 $P(AP)$

A seguir os resultados dos ajustes das funções de densidade de probabilidade para as variáveis que compõe o módulo denominado ambiente de produção (variáveis meteorológicas) são apresentados. Gráficos de densidade de probabilidade estimados pelo método não paramétrico e para as três outras distribuições sugeridas (log-normal, gama e weibull). Apresenta-se também gráficos de Cullen e Frey (Delignette-muller; Dutang, 2015) que auxiliam a escolha da distribuição de probabilidade para a variável, por meio da assimetria e curtose. Além disso, o teste Kolmogorov-Smirnov é realizado e os índices AIC e BIC são calculados para dar mais suporte a escolha do melhor ajuste de distribuição de probabilidade para as variáveis.

Primeiramente o ajuste para a variável soma térmica será discutido. Por meio do gráfico de densidade de probabilidade apresentado na figura 3.10 nota-se a distribuição log-normal e gama apresentam ajustes muito parecidos. Inclusive, essas duas distribuições se aproximam bastante da densidade não paramétrica em suas caudas. Embora a densidade não paramétrica sugira a existência de bimodalidade na distribuição de probabilidade da soma térmica, que pode ser causada por algum evento cíclico como El Niño e La Niña, as distribuições log-normal e gama aparentemente se ajustam

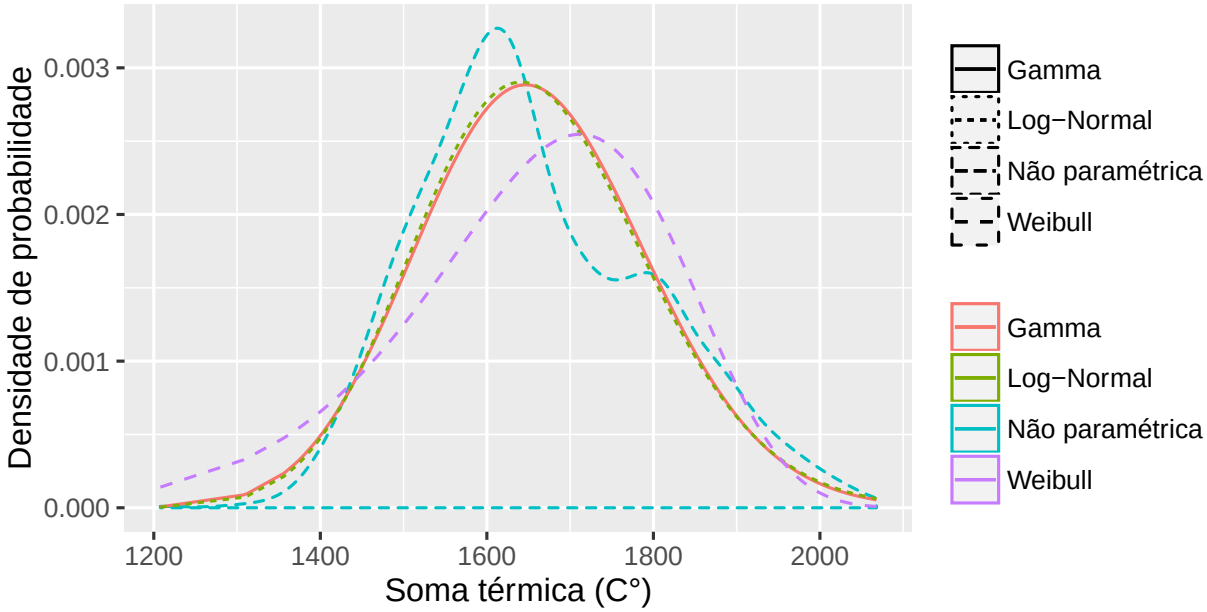


Figura 3.10 – Densidades de probabilidade estimadas por kernel da variável soma térmica e as densidades de probabilidades estimadas pela função fitdist do pacote ”fitdistrplus”(Fonte: autor)

bem aos dados de soma térmica.

O gráfico de Cullen e Frey apresentado na figura 3.11, por outro lado, sugere que a variável soma térmica possua distribuição beta. Alternativamente é possível supor que ela possua distribuição normal. Embora o gráfico de Cullen e Frey seja útil para escolha de distribuições de probabilidade, outras informações precisam ser levadas em consideração para tal escolha.

A tabelas 3.7 apresentam os resultados dos testes de Kolmogorov–Smirnov para os ajustes das distribuições da variável soma térmica. O teste usado é de uma amostra, com hipótese alternativa do tipo “two-sided” segundo a função “ks.test” disponível no R. Verifica-se que o teste é significativo para todas as distribuições testadas, o que significa que a hipótese de que a variável possua essas distribuições é rejeitada. Mesmo com esse resultado, ainda é possível usar a estatística do teste como uma medida para comparar os ajustes entre eles. Neste caso quanto menor o valor da estatística melhor é o ajuste e, assim, verifica-se que a distribuição log-normal apresenta melhor ajuste.

O resultado dessa comparação é confirmado ao observar os valores dos critérios AIC e BIC apresentados na tabela 3.8 os quais também sugerem que a distribuição log-normal é a mais adequada para essa variável, dentre as três distribuições testadas.

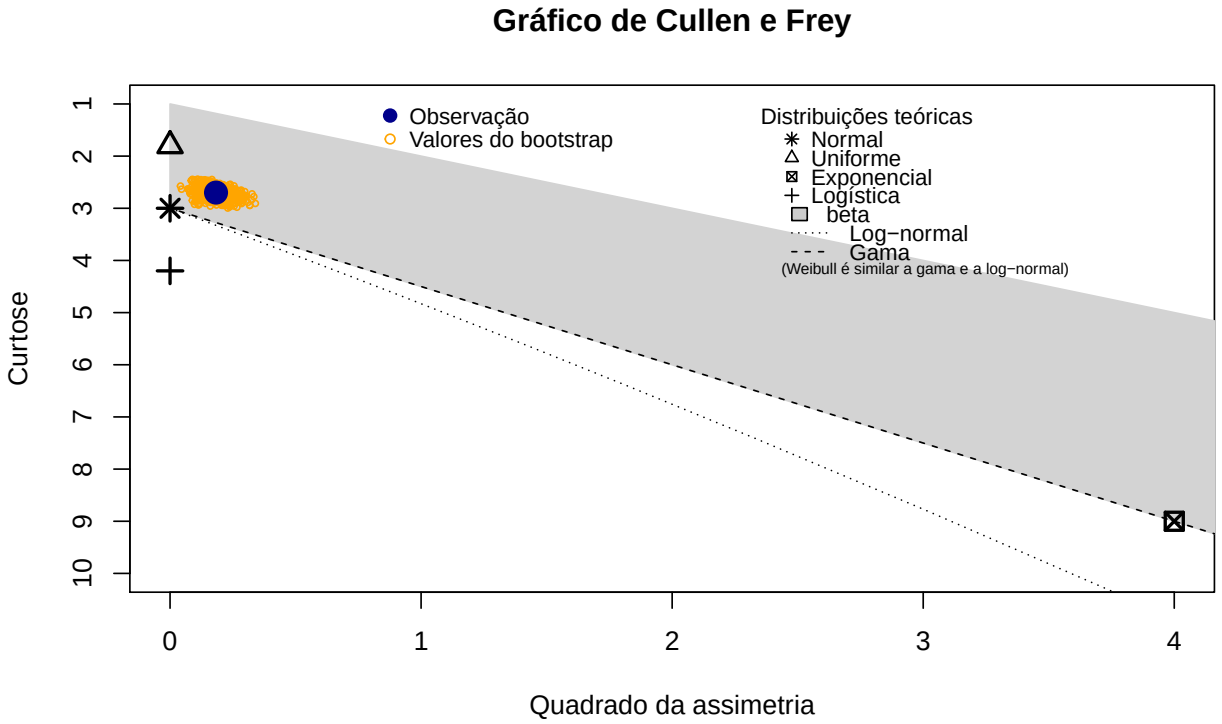


Figura 3.11 – Gráfico de Cullen e Frey para a amostra de dados de soma térmica (fonte: autor)

Tabela 3.7 – Resultados dos testes de Kolmogorov-Smirnov de uma amostra para avaliação da adequabilidade das distribuições ajustadas para a variável soma térmica

	Estatística
Gamma	0,07289765
Log-normal	0,06733437
Weibull	0,11537673

Após essa análise de ajustes, a tabela 3.9 apresenta os valores dos parâmetros ajustados para as distribuições gama, log-normal e weibull para a variável soma térmica.

A figura 3.12 mostra gráficos de diferentes funções de densidade de probabilidade ajustadas para a variável precipitação acumulada. Assim como para a variável soma térmica, a primeira impressão é que as distribuições log-normal e gama são as que melhor explicam a distribuição da variável precipitação acumulada ao comparar com a distribuição estimada por kernel. A distribuição weibull apresenta menor capacidade de explicar a variável na cauda a esquerda, região de grande importância ao se considerar o risco de quebra de rendimento de culturas agrícolas gerado por volumes baixos de chuva ao longo de seus ciclos.

O gráfico de Cullen e Frey apresentado na figura 3.13 dá suporte ao que foi

Tabela 3.8 – Valores dos critérios AIC e BIC de escolha de modelos para ajuste de distribuição para soma térmica

	AIC	BIC
Gamma	18188,47	18199,01
Log-normal	18177,81	18188,34
Weibull	18450,98	18461,51

Tabela 3.9 – Resultados dos ajustes de parâmetros das distribuições de probabilidades para a variável precipitação acumulada

	Estimativa	Desvio padrão
Shape	142,87	5,20
Rate	0,09	0,00

(a) Gamma

	Estimativa	Desvio padrão
Meanlog	7,41	0,00
Sdlog	0,08	0,00

(b) Log-normal

	Estimativa	Desvio padrão
Shape	11,89	0,23
Scale	1724,35	4,07

(c) Weibull

constatado no gráfico de densidades de probabilidades da figura 3.12. A curtose e a assimetria estimados para os dados sugere que a distribuição log-normal é a mais adequada para os dados, embora ele não permita distinguir a weibull da gama e da log-normal.

Novamente, o teste de Kolmogorov-Smirnov mostrado na tabela 3.10 não favorece a escolha das distribuições de probabilidade propostas. O teste é significativo para todas as distribuições, sugerindo que a real distribuição de probabilidade da variável precipitação acumulada é diferente das distribuições ajustadas. No entanto, a estatística do teste também sugere que a distribuição log-normal é a que apresentou melhor ajuste das três.

Tabela 3.10 – Resultados dos testes de Kolmogorov-Smirnov de uma amostra para avaliação da adequabilidade das distribuições ajustadas para a variável precipitação acumulada

	Estatística
Gamma	0,077072
Log-normal	0,059799
Weibull	0,112911

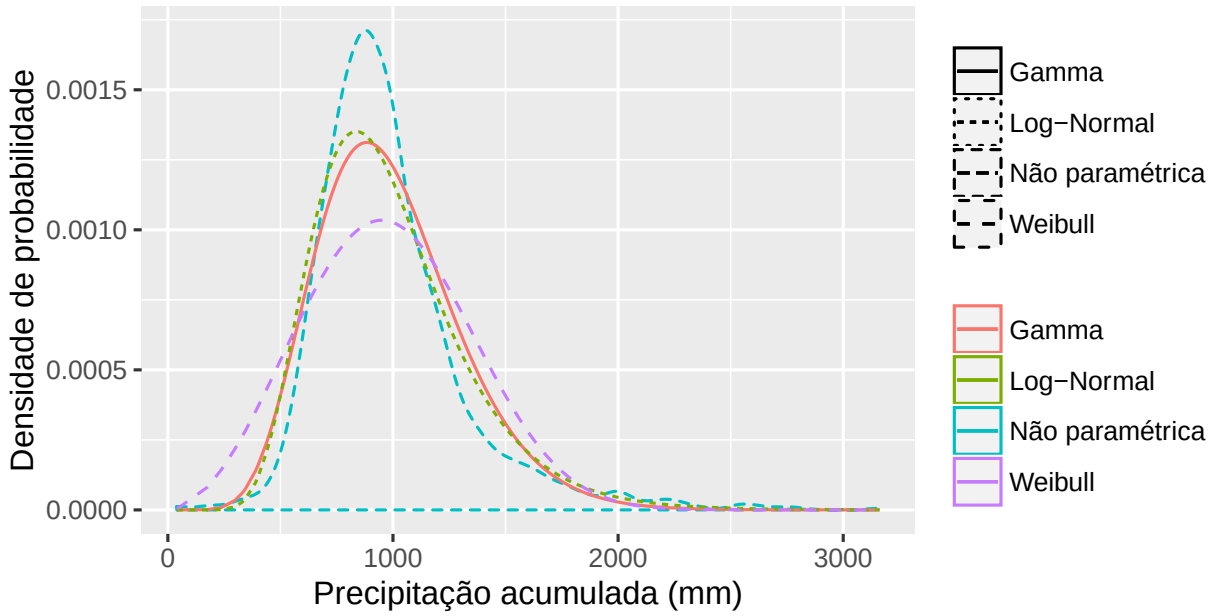


Figura 3.12 – Densidades de probabilidade estimadas por kernel da variável precipitação acumulada e as densidades de probabilidades estimadas pela função fitdist do pacote ”fitdistrplus”(Fonte: autor)

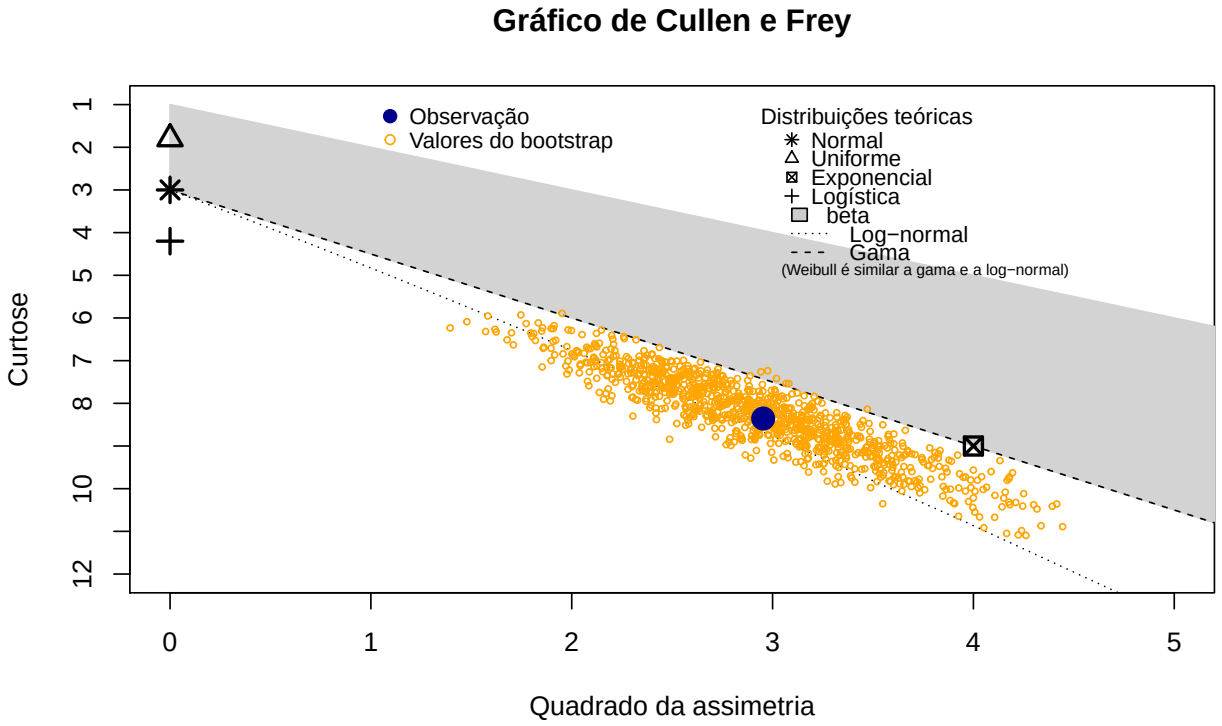


Figura 3.13 – Gráfico de Cullen e Frey para a amostra de dados de precipitação acumulada (fonte: autor)

Ao se avaliar os valores dos critérios AIC e BIC, apresentados na tabela 3.11, ambos sugerem que a gama seja escolhida como distribuição de probabilidade da variável precipitação acumulada. Como a diferença dos valores dos critérios entre as distribuições log-normal e gama é pequena, o conjunto de resultados composto pelo gráfico de densidade de probabilidade, o gráfico de Cullen e Frey, o teste de Kolmogorov-Smirnov e os próprios critérios AIC e BIC levam a escolha da distribuição log-normal como aquela que melhor se ajusta a variável precipitação acumulada, dado o conjunto de dados disponíveis para o ajuste.

Tabela 3.11 – Valores dos critérios AIC e BIC de escolha de modelos para ajuste de distribuição para precipitação acumulada

	AIC	BIC
Gamma	19027	19037
Log-normal	19071	19081
Weibull	19300	19310

Com isso, a tabela 3.12 apresenta os valores dos parâmetros ajustados para as distribuições gama, log-normal e weibull para a variável precipitação acumulada.

Tabela 3.12 – Resultados dos ajustes de parâmetros das distribuições de probabilidades para a variável precipitação acumulada

	Estimativa	Desvio padrão
Shape	9,61	0,33
Rate	0,01	0,00

(a) Gamma

	Estimativa	Desvio padrão
Meanlog	6,84	0,01
Sdlog	0,33	0,01

(b) Log-normal

	Estimativa	Desvio padrão
Shape	2,89	0,05
Scale	1100,16	11,09

(c) Weibull

Para completar a análise das variáveis meteorológicas, agora o ajuste das distribuições de probabilidade para a variável maior período sem chuva será apresentado e discutido. O gráfico de densidades de probabilidade apresentado na figura 3.14 indica que as funções de densidade de probabilidade das distribuições log-normal e gama mais se aproximam da distribuição de probabilidade estimada por kernel, em especial

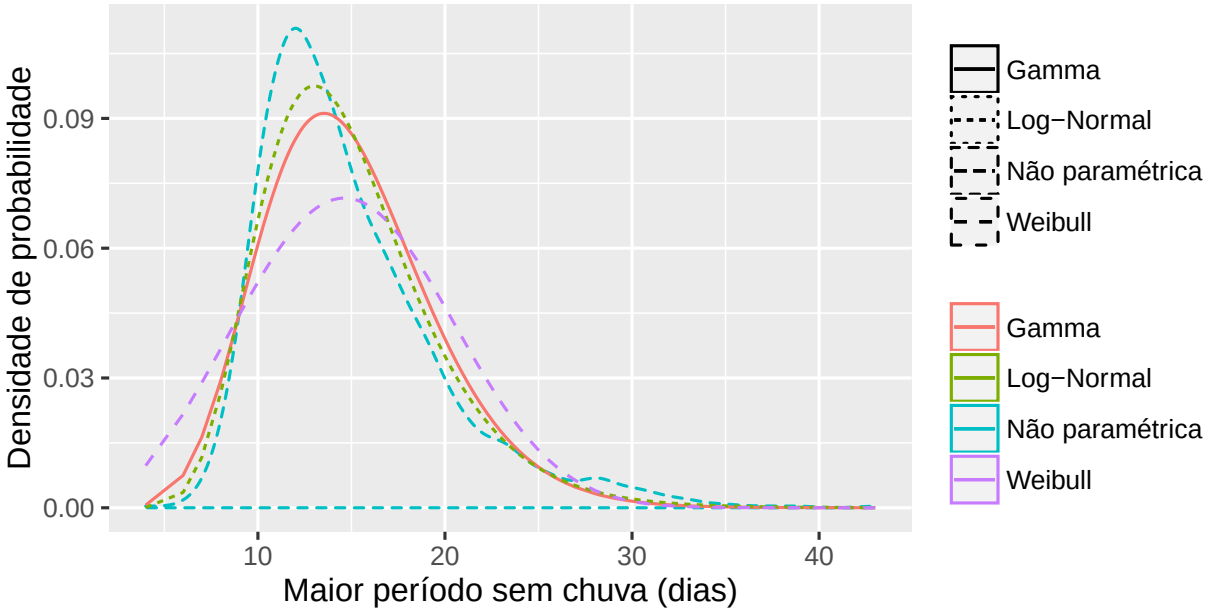


Figura 3.14 – Densidades de probabilidade estimadas por kernel da variável maior período sem chuva e as densidades de probabilidades estimadas pela função fitdist do pacote ”fitdistrplus”(Fonte: autor)

nas caudas da distribuição. Assim como para as outras variáveis meteorológicas, a weibull pior representa a distribuição na cauda a esquerda. Agronomicamente e se tratando de risco de quebra de rendimento de culturas agrícolas, a cauda a direita é mais importante já que ela descreve o risco de ocorrência se veranicos longos, ou secas.

A figura 3.15 mostra o gráfico de Cullen e Frey para a amostra de dados de maior período sem chuva. A curtose e assimetria sugerem que a distribuição que melhor representa a variável é a gama, embora o gráfico também sugira a possibilidade de considerar a distribuição beta como boa alternativa.

A tabela 3.13 apresenta os resultados dos testes de Kolmogorov–Smirnov para os ajustes das distribuições da variável maior período sem chuva. Da mesma forma que para as outras variáveis meteorológicas analisadas anteriormente, o teste é significativo para as três distribuições ajustadas, sugerindo que todas são significativamente diferentes da distribuição verdadeira da variável. E novamente, considerando a estatística do teste como um critério de qualidade de ajuste, a distribuição log-normal é a que melhor explica a distribuição dos dados da variável maior período sem chuva.

Seguindo a mesma linha do teste de Kolmogorov-Smirnov, os critérios AIC e BIC, apresentados na tabela 3.14, sugerem que a distribuição log-normal é a que me-

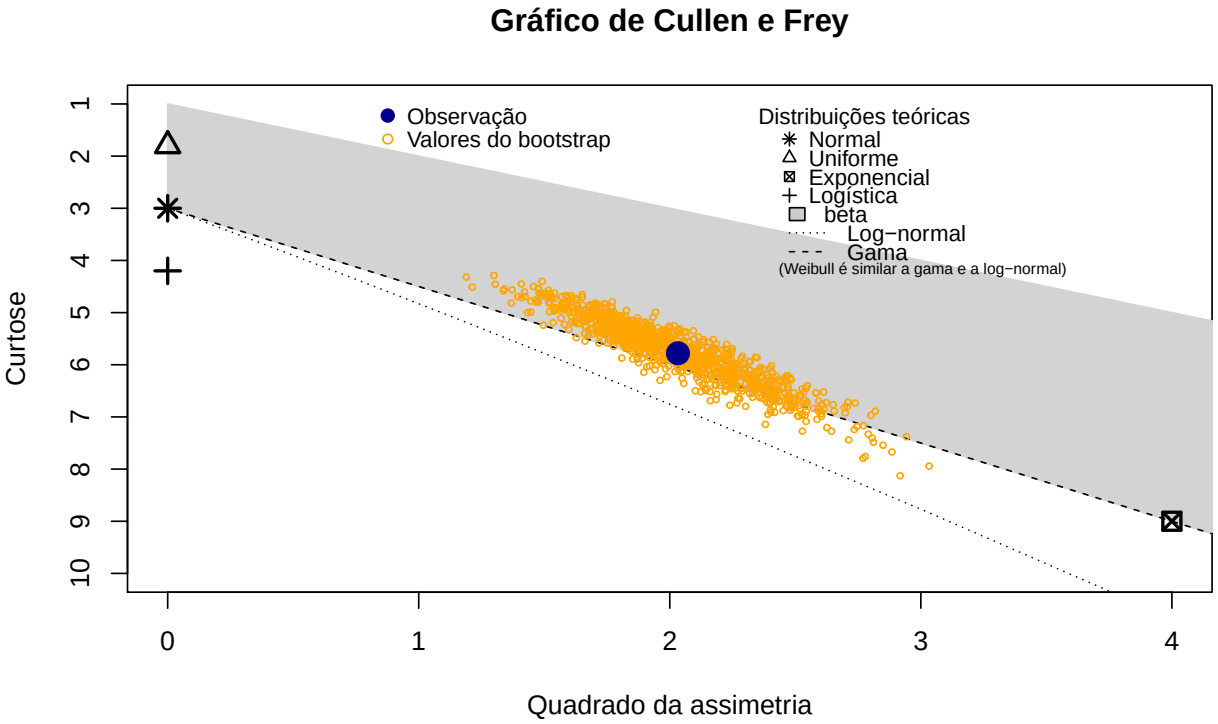


Figura 3.15 – Gráfico de Cullen e Frey para a amostra de dados de maior período sem chuva (fonte: autor)

Tabela 3.13 – Resultados dos testes de Kolmogorov-Smirnov de uma amostra para avaliação da adequabilidade das distribuições ajustadas para a variável maior período sem chuva

	Estatística
Gamma	0,095098
Log-normal	0,075510
Weibull	0,123663

lhor explica a variável maior período sem chuva. Tomando todas as informações sobre o ajustes para essa variável, adota-se a distribuição log-normal como sendo a que melhor se ajustou.

A tabela 3.15 apresenta os valores dos parâmetros ajustados para as distribuições gama, log-normal e weibull para a variável maior período sem chuva.

3.3.4.2 $P(Y|AP)$ e $P(IE|Y)$

Como mencionado na seção 3.2.8, as distribuições do rendimento relativo e dos índices de estado são ajustadas por meio de regressão beta.

Em comparação as discussões de ajustes anteriores, esta se baseia em menos informações para as avaliações de ajuste. Por a distribuição de interesse ser definida

Tabela 3.14 – Valores dos critérios AIC e BIC de escolha de modelos para ajuste de distribuição para maior período sem chuva

	AIC	BIC
Gamma	7733,4	7743,7
Log-normal	7651,9	7662,3
Weibull	8038,9	8049,2

Tabela 3.15 – Resultados dos ajustes de parâmetros das distribuições de probabilidades para a variável maior período sem chuva

	Estimativa	Desvio padrão
Shape	10,74	0,41
Rate	0,72	0,03

(a) Gamma

	Estimativa	Desvio padrão
Meanlog	2,66	0,01
Sdlog	0,30	0,01

(b) Log-normal

	Estimativa	Desvio padrão
Shape	3,05	0,06
Scale	16,65	0,16

(c) Weibull

como sendo a beta, não há necessidade de analisar gráficos de Cullen e Frey. Os principais critérios de avaliação são, então, compostos pelos testes de significância gerados pelo pacote betareg, e por uma análise gráfica das densidades de probabilidades aproximada a partir de simulações dos valores das variáveis explicadas após o ajuste dos modelos.

Assumindo o rendimento máximo de grãos de 4800 kg.ha^{-1} (80 sc.ha^{-1}), a variável rendimento relativo é obtida pela divisão do rendimento observado pelo rendimento máximo. Assim, para o conjunto de dados disponível, as observações da variável rendimento relativo estão no intervalo entre zero e um.

Uma comparação é realizada aqui para avaliar dois modelos propostos para o rendimento relativo. Em um a esperança da distribuição beta é função das variáveis de ambiente de produção e o parâmetro de precisão é uma constante. Em outro a esperança e a precisão são funções das variáveis de ambiente. As tabelas 3.16a e 3.16b apresentam os resultados dos ajustes para esses modelos, respectivamente.

Ao se considerar o parâmetro de precisão como uma constante, a função de

Tabela 3.16 – Resultado dos ajustes de parâmetros da regressão beta na modelagem da distribuição de probabilidade de rendimento relativo

Parâmetro	Estimativa	Erro padrão	Estatística t	Pr(>|t|)
Esper. (Intercepto)	5,7e-01	0,32704	1,73	8,4e-02
Esper. Soma térm.	-8,8e-05	0,00020	-0,44	6,6e-01
Esper. Chuva acum.	5,8e-04	0,00011	5,38	7,3e-08
Esper. Maior peri.	-2,5e-02	0,00353	-7,21	5,5e-13
Phi (Intercepto)	3,6e+01	2,87542	12,57	3,1e-36

(a) Rendimento relativo (μ)

Parâmetro	Estimativa	Erro padrão	Estatística t	Pr(>|t|)
Esper. (Intercepto)	0,65017	0,30818	2,11	3,5e-02
Esper. Soma térm.	-0,00018	0,00019	-0,96	3,4e-01
Esper. Chuva acum.	0,00062	0,00010	6,19	6,2e-10
Esper. Maior peri.	-0,02307	0,00395	-5,83	5,4e-09
Phi (Intercepto)	0,52602	1,33102	0,40	6,9e-01
Phi Soma térm.	0,00181	0,00082	2,22	2,6e-02
Phi Chuva acum.	0,00086	0,00044	1,97	4,9e-02
Phi Maior peri.	-0,03980	0,01462	-2,72	6,5e-03

(b) Rendimento relativo (μ, ϕ)

densidade de probabilidade da distribuição beta usada no método de máxima verossimilhança é apresentada na equação 3.8.

$$f(y;\beta,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{3.8}$$

com $\mu = (e^{\beta X})/(1+e^{\beta X})$, Y representando a variável rendimento relativo, e X representando as variáveis de ambiente.

Observando os resultados apresentados na tabela 3.16a, verifica-se que apenas parâmetro associado a variável soma térmica é não significativo ao nível de 5%, com hipótese nula de ser igual a zero. Os demais são significativos ao nível de 1%. Os sinais que acompanham os parâmetros indicam o sinal da derivada primeira da relação entre as variáveis de ambiente e os parâmetros da distribuição beta. Dado isso, verifica-se que a chuva acumulada tem relação positiva com o rendimento relativo, e o maior período de chuva tem relação negativa com o rendimento relativo. Embora o parâmetro da soma térmica não seja significativo, o sinal negativo do parâmetro pode estar associado a sua relação com o comprimento de ciclo e exposição ao risco de seca.

Ao se considerar a esperança e a precisão como dependentes das variáveis

de ambiente, a função de densidade de probabilidade da distribuição beta usada no método de máxima verossimilhança é a apresentada na equação 3.9.

$$f(y;\beta,\theta) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{3.9}$$

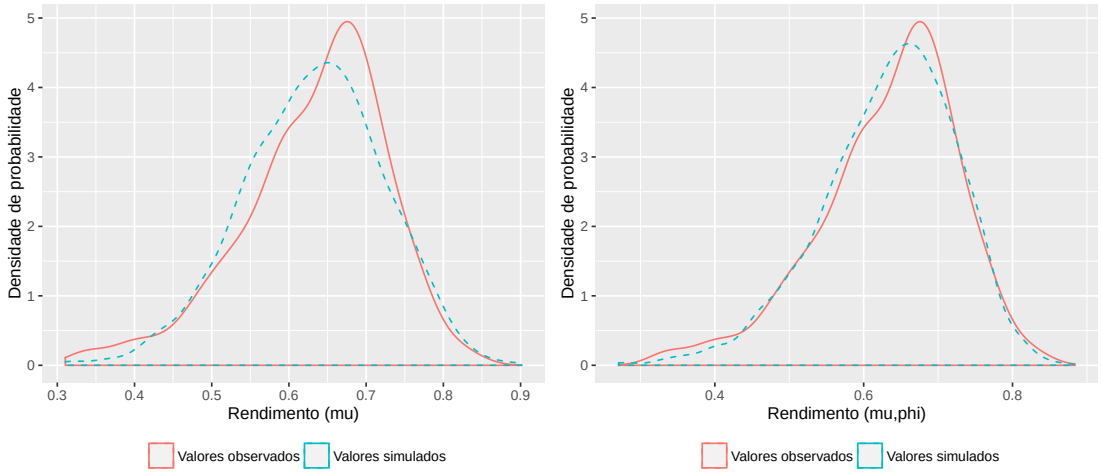
com $\mu = (e^{\beta X})/(1+e^{\beta X})$, $\phi = e^{\theta X}$, Y representando a variável rendimento relativo, e X representando as variáveis de ambiente.

Repetindo a análise anterior, apenas o intercepto associado a precisão da distribuição é não significativo ao nível de 5%. Ao nível de 1% o parâmetro associado a soma térmica que explica a esperança também é não significativo. Novamente os sinais dos parâmetros que explicam a esperança são os mesmos. O da soma térmica possui sinal negativo, da precipitação acumulada possui sinal positivo, e o do maior período sem chuva possui sinal negativo.

Como novidade, tem-se os parâmetros associados a precisão da distribuição beta que está associado a incerteza da variável rendimento relativo para uma dada esperança, e que possui relação inversa com a variância para uma dada esperança (a variância da distribuição beta é obtida por $\mu(1-\mu)/(\phi+1)$). Os sinais dos parâmetros associados as variáveis soma térmica e precipitação acumulada possuem sinal positivo, indicando que o aumento dos valores dessas variáveis leva a uma redução na incerteza do rendimento relativo. Esse efeito da soma térmica pode ser devido a que o aumento do comprimento do ciclo reduz o impacto de fatores de risco pontuais no ciclo da cultura, e que a cultura tem mais tempo para se recuperar de fatores de risco não explicados pelo modelo, como a desfolha causada por pragas. No caso do efeito da precipitação acumulada, a redução da incerteza na variável rendimento relativo pode se dever ao aumento da resiliência da cultura a fatores de risco não explicados pelo modelo quando ela está mais bem suprida com água. O efeito do maior período sem chuva (“veranico”) vai contra o efeito da precipitação acumulada e da soma térmica. Quanto maior o período sem chuva, menor é o tempo com condições não ruins para a cultura se recuperar e, também devido a isso, mais debilitada e menos resiliente a cultura fica para se recuperar de fatores de risco não explicados pelo modelo.

A fim de escolher entre o modelo com parâmetro de precisão fixo, ou com precisão dependente das variáveis de ambiente, uma análise gráfica de qualidade de ajuste foi realizada. A partir dos dados observados das variáveis de ambiente, para cada ob-

servação delas 10 simulações da variável rendimento relativo foram obtidas por meio da distribuição beta ajustada para cada um dos modelos. O número 10 é considerado apenas para aumentar o tamanho da amostra de dados e melhorar a consistência dos resultados entre simulações diferentes. Um gráfico de densidade de probabilidade dessas amostras simuladas foi plotado para cada um dos modelos, e apresentado na figura 3.16.



(a) (b)

Figura 3.16 – Gráfico de dispersão de rendimento da soja (IBGE) (Fonte: autor)

Mesmo com a simplicidade desse critério de avaliação, é possível perceber diferença no ajuste dos modelos. Aparentemente, o modelo que assume dependência entre a precisão e as variáveis de ambiente (figura 3.16b) tem melhor ajuste ao se considerar este critério, especialmente nas proximidades da moda e na cauda a esquerda. É possível perceber pela figura 3.16a que a precisão da distribuição beta é levemente subestimada ao considerá-la constante, o que leva a um achatamento do gráfico de densidade de probabilidade.

Seguindo a mesma linha de análise, a seguir os ajustes dos modelos que descrevem a relação entre rendimento relativo e índices de estado (AEVI e ANDVI) são apresentados na tabela 3.17.

Avaliando primeiramente os ajustes dos modelos que consideram precisão constante, apresentados nas tabelas 3.17a e 3.17c, percebe-se que com exceção do intercepto da esperança todos os outros são significativamente diferentes de zero para o modelo do índice de AEVI como para o de ANDVI.

Ao se considerar os modelos em que a precisão é dependente do rendimento

relativo (tabelas 3.17b e 3.17d), o resultado é similar. O ajuste do modelo que explica o índice ANDVI revela que dois parâmetros são não significativos ao nível de 5%, o intercepto da esperança e o coeficiente da variável rendimento relativo. No caso do modelo que explica o índice AEVI o único parâmetro não significativo é o intercepto da esperança.

Tabela 3.17 – Resultado dos ajustes de parâmetros da regressão beta na modelagem das distribuições de probabilidade de ANDVI e de AEVI

Parâmetro	Estimativa	Erro padrão	Estatística t	Pr(>|t|)
Esper. (Intercept)	0,12	0,10	1,2	2,5e-01
Esper. rend	0,39	0,16	2,5	1,4e-02
Phi (phi)	57,94	4,58	12,7	1,0e-36

(a) ANDVI (μ)

Parâmetro	Estimativa	Erro padrão	Estatística t	Pr(>|t|)
Esper. (Intercept)	0,14	0,11	1,3	1,9e-01
Esper. rend	0,36	0,16	2,2	3,0e-02
Phi (Intercept)	3,37	0,54	6,2	4,8e-10
Phi rend	1,10	0,85	1,3	1,9e-01

(b) ANDVI (μ, ϕ)

Parâmetro	Estimativa	Erro padrão	Estatística t	Pr(>|t|)
Esper. (Intercept)	-0,15	0,12	-1,3	2,0e-01
Esper. rend	1,17	0,18	6,3	2,3e-10
Phi (phi)	44,80	3,53	12,7	7,7e-37

(c) AEVI (μ)

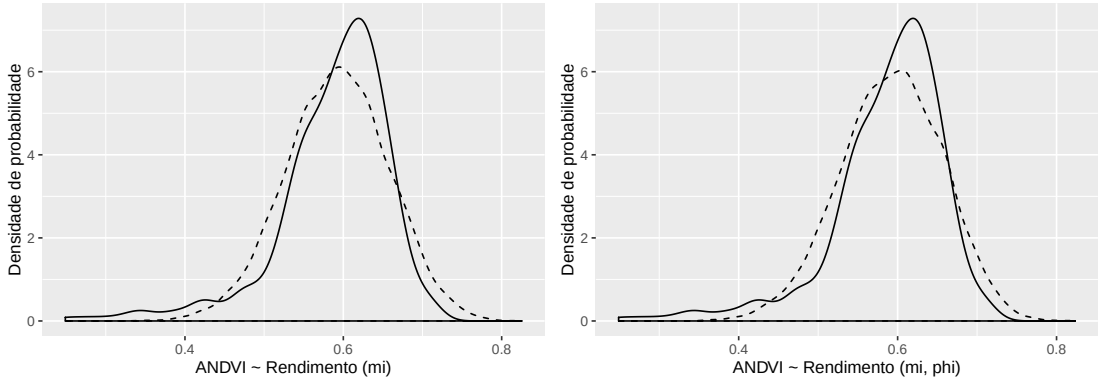
Parâmetro	Estimativa	Erro padrão	Estatística t	Pr(>|t|)
Esper. (Intercept)	0,093	0,13	0,71	4,8e-01
Esper. rend	0,797	0,20	4,06	4,8e-05
Phi (Intercept)	1,414	0,54	2,64	8,4e-03
Phi rend	3,891	0,84	4,63	3,6e-06

(d) AEVI (μ, ϕ)

Os gráficos de densidades estimadas por kernel a partir dos valores observados dos índices e de valores simulados usando o modelo e observações dos rendimentos relativos são apresentados nas figuras 3.17 e 3.18, para os índices ANDVI e AEVI respectivamente. Comparando essas duas figuras, a primeira impressão é de que a regressão beta não capta adequadamente a dispersão do índice AEVI quando a variável explicativa é o rendimento relativo. Por outro lado, o gráfico de densidade estimado a partir de simulação é mais parecido com o gráfico de densidade estimado a partir de dados observados em comparação com o ANDVI.

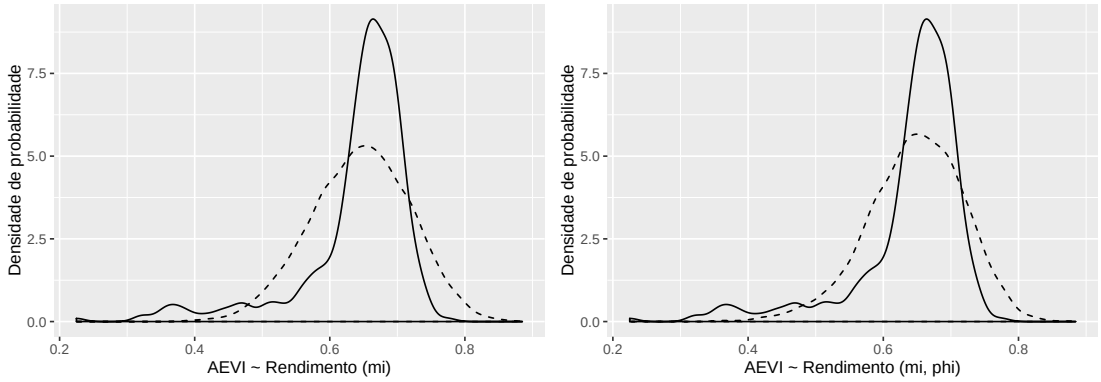
Outro ponto que se nota é a coerência entre os resultados apresentados nas tabelas 3.17a e 3.17b com os gráficos 3.17a e 3.17b. A incorporação de um componente dependente do rendimento relativo para explicar a precisão da distribuição não altera significativamente, nem visualmente o ajuste da distribuição.

Em se tratando do índice AEVI, o efeito da modelagem da precisão também é coerente. Novamente tomando as tabelas 3.17c e 3.17d e os gráficos 3.18a e 3.18b, percebe-se que o rendimento relativo contribui para explicar a precisão do modelo. Esse efeito é verificado pela redução do achatamento da densidade de probabilidade estimada pela simulação.



(a) (b)

Figura 3.17 – Gráfico de dispersão de rendimento da soja (IBGE) e de soma de NDVI (Fonte: autor)



(a) (b)

Figura 3.18 – Gráfico de dispersão de rendimento da soja (IBGE) e de soma de NDVI (Fonte: autor)

3.3.5 $P(Y|IE, AP)$

Com $P(Y|AP)$ e $P(IE|Y)$ estimados é possível simular valores da distribuição $P(Y|IE, AP)$. Esta distribuição é, no entanto, desconhecida. Para a simulação, a técnica de simulação MCMC pode ser usada. Primeiramente é necessário especificar os modelos que representam $P(Y|AP)$ e $P(IE|Y)$. Também é necessário especificar a distribuição conjunta e condicionada $P(Y, IE|AP)$.

Pelo teorema de Bayes, pelo axioma das probabilidades totais, e pelas suposições de independência entre variáveis:

$$P(Y|IE, AP) = \frac{P(Y, IE|AP)}{\int_Y P(Y, IE|AP)} \tag{3.10}$$

Com isso é possível evitar a derivação de uma forma analítica para $P(Y|IE, AP)$ e MCMC é usado para a simular uma amostra dessa distribuição.

O processo de simulação por MCMC requer a especificação de uma priori para $P(Y|IE, AP)$, que pode ser não informativa, e uma função a posteriori. Os logaritmos dessas funções podem ser usados para reduzir erros computacionais de cálculo.

O processo também necessita do estabelecimento de uma função de “salto”, a qual é usada para gerar candidatos para a distribuição a posteriori. Neste caso a variável Y é transformada do domínio $(0, 1)$ para o domínio $(-\text{inf}, \text{inf})$ por uma função logit, resultando em Y'. O candidato é gerado a partir de uma distribuição normal $N(Y', 0.5)$. Este candidato é transformado novamente para o domínio $(0, 1)$ pela função inversa da logit. Também é necessário o estabelecimento da função densidade de probabilidade da função salto.

Com essas especificações, o algoritmo de simulação Metropolis-Hastings, ou amostrador de Gibbs, pode ser usado. Neste estudo o algoritmo Metropolis-Hastings é adotado. Mais informações sobre esses algoritmos podem ser encontradas em Gelman; Carlin; Stern; Dunson; Vehtari; Rubin, 2013.

3.3.6 Aplicação do modelo para casos hipotéticos

Com todos esse ferramental de modelos em mãos, é possível utilizá-lo em avaliações de riscos em cenários hipotéticos. Num primeiro cenário, supõe-se que não se dispõe de observação das variáveis dos modelos. Este é um caso típico antes do

início da safra, quando os prêmios de contratos de seguros são definidos. Neste caso a amostra de rendimento relativo é obtida em dois passos. O primeiro consiste em gerar uma amostra das variáveis de ambiente de produção por meio de $P(AP)$. O segundo passo é usar os dados simulados no passo anterior para gerar uma amostra de rendimento relativo por meio de $P(Y|AP)$. O gráfico de densidade aproximada é apresentado na figura 3.19.

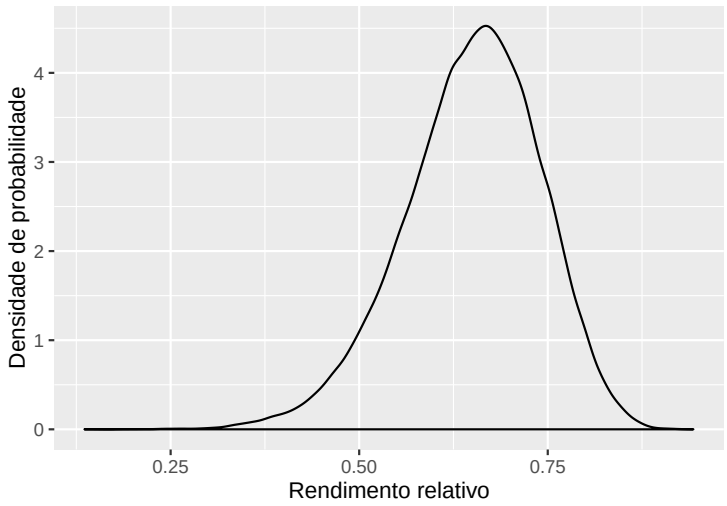


Figura 3.19 – Densidade de probabilidade aproximada da variável rendimento relativo para o cenário 1

Neste cenário, a média do rendimento relativo simulado é de 65% da produtividade potencial (52 sacas por hectare), com desvio padrão 0.09. Verifica-se que a amostra simulada possui desvio padrão pequeno, evidenciado pelo coeficiente de variação estimado de 14%. Isto tem algumas implicações importantes na aplicação do modelo ajustado.

Mesmo sem incluir informação de observação das variáveis de ambiente de produção ou do índice de estado da cultura, a variabilidade da variável rendimento relativo é relativamente pequena. Isso implica em caudas achatadas, que significa pequena probabilidade de ocorrência de valores pequenos ou grandes de rendimento relativo. Do ponto de vista técnico, essa variabilidade pode significar que a cultura da soja é resistente ou resiliente a condições adversas durante o ciclo de desenvolvimento. Do ponto de vista de análise de risco, essa variabilidade pode ser interpretada como baixa capacidade do modelo representar os riscos que ocorrem com a cultura da soja. No entanto, ao verificar o processo de ajuste, os dados coletados, e a origem da variável rendimento relativo o entendimento desse fenômeno fica mais claro. A variável rendi-

mento relativo representa rendimentos médios de rendimentos agrícolas de produtores de soja para cada safra e município disponíveis no banco de dados usado neste estudo, o que pode levar a uma redução da variabilidade da variável. Esses dados de rendimento podem compor uma amostra não representativa para rendimentos pequenos ocorridos nos anos de adversidades climáticas. Além disso, a coleta de dados das outras variáveis que compõe o modelo, ao nível municipal, também pode levar a redução da variabilidade do rendimento relativo no processo de simulação.

A grande implicação da variabilidade pequena no processo de precificação de contratos de seguro agrícola de produção é que os prêmios puros são subestimados e inferiores aos valores praticados no mercado, mesmo removendo taxas administrativas e outros custos que elevam o prêmio cobrado no mercado. Mesmo com este grande problema, é possível usar o prêmio estimado com os resultados do ajuste do modelo de forma relativa, mas com cautela.

Para essa comparação de prêmios, define-se um contrato de seguro de produtividade da soja cuja produção segurada é de 3300 kg.ha^{-1} (cerca de 55 sc.ha^{-1}), nível de cobertura é 70%, e o limíte máximo de indenização é 80%. Neste caso o rendimento garantido por hectare é de 2310 kg.

Tomando agora um segundo cenário para iniciar as comparações. Neste caso supõe-se que se tem ideia de quais serão os valores das variáveis soma térmica, precipitação acumulada e maior período sem chuva seja verificado no final do ciclo da cultura, e que esses valores são os apresentados na tabela 3.18. Este cenário representa uma situação ótima de produção, com volume grande de precipitação acumulada, ausência de veranicos, e soma térmica média.

Tabela 3.18 – Valores de variáveis observadas no cenário 2

Variáveis	Valores
Soma térmica	1700,00
Precipitação acumulada	1100,00
Maior período sem chuva	7,00

Utilizando a distribuição de probabilidade do rendimento obtida anteriormente ,$P(Y|AP)$, é possível simular uma amostra de rendimento obtido ao final do ciclo para o segundo cenário. Uma aproximação da densidade de probabilidade do rendimento relativo é apresentado na figura 3.20.

Neste cenário, a média do rendimento relativo simulado é de 70% da produti-

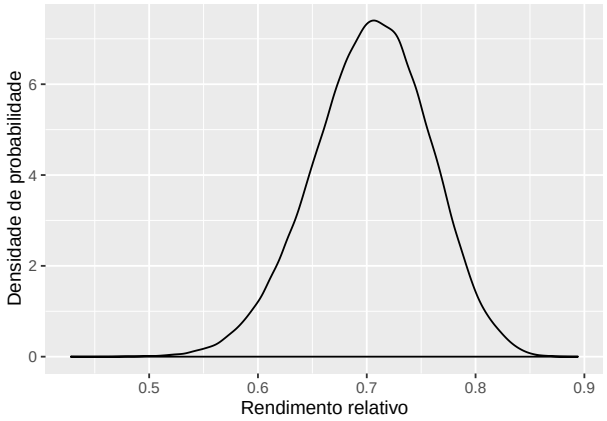


Figura 3.20 – Densidade de probabilidade aproximada da variável rendimento relativo para o cenário 2

vidade potencial (56 sacas por hectare), com desvio padrão 0.053, e coeficiente de variação 8%. Com base nessas estatísticas descritivas, já é possível perceber o impacto da inclusão da informação sobre a distribuição da variável rendimento relativo. Um primeiro ponto que se nota é que a média do rendimento aumentou e passou de 65% para 70%, o que era esperado já que o primeiro cenário representa uma condição de produção típica de ano regula, e o segundo cenário representa uma condição de produção muito boa. A variabilidade da variável no segundo cenário, verificada pelo desvio padrão e pelo coeficiente de variação, é reduzida com a inclusão da informação sobre o ambiente de produção. A explicação para isto é que a inclusão de informação aumentou o nível de certeza (precisão) sobre a estimativa de rendimento relativo.

O prêmio puro, medido em unidades de produção, do seguro de produtividade pode ser estimado pela seguinte expressão:

$$premio = E(min(max(prod_{gar} - prod_{obt}, 0), LMI)) \tag{3.11}$$

em que $prod_{gar}$ é o resultado da multiplicação do rendimento segurado pelo nível de cobertura, $prod_{obt}$ é o rendimento obtido pelo produtor (rendimento simulado), e LMI é o resultado da multiplicação do limite máximo de indenização pelo rendimento segurado.

Tomando os prêmios puros calculados a partir das amostras simuladas de rendimento relativo para o primeiro e segundo cenários, 0.163 e 0.000063 , verifica-se grande diferença entre a magnitude dos valores. É preciso enfatizar, que por motivos relacionados aos dados e métodos usados para o ajuste dos modelos esses prêmios

são extremamente baixos ao compará-los com prêmios praticados no mercado. Claramente, seguindo o movimento da média e da variabilidade do rendimento relativo, o segundo cenário é de muito menos risco que o primeiro.

Tomando um terceiro cenário, em que as variáveis de ambiente de produção possuem os valores especificados na tabela 3.19. Ele representa uma situação com alto risco de quebra de produção da soja. O volume de chuva acumulada é pequeno, um veranico de 20 dias ocorre durante o desenvolvimento da cultura, e a soma térmica alta representa alta demanda hídrica da cultura pela elevação da transpiração.

Tabela 3.19 – Valores de variáveis observadas no cenário 3

Variáveis	Valores
Soma térmica	1800,00
Precipitação acumulada	600,00
Maior período sem chuva	20,00

Como se espera em uma situação característica de seca (déficit hídrico), a média de rendimento relativo é menor que a dos cenários anteriores, com valor de 56%. Neste cenário, o desvio padrão de 0.084 é menor que o verificado no primeiro cenário, indicando aumento de certeza sobre os valores de rendimento relativo a serem verificados neste cenário. Comparando os prêmios puros estimados do primeiro e do terceiro cenários, 0.163 e 0.684 respectivamente, verifica-se um aumento considerável em seus valores coerente com o aumento do risco de quebra de rendimento suposto no cenário 3. O prêmio do terceiro cenário é cerca de 4 vezes maior que o prêmio do primeiro cenário. Novamente, enfatiza-se que esses valores de prêmio não são coerentes em magnitude com valores praticados no mercado brasileiro. Dessa forma eles são apenas úteis em uma abordagem comparativa de cenários.

O quarto e o quinto cenários consideram situações em que a cultura completa seu ciclo e dispõe-se de observações das variáveis de ambiente de produção e de índices de estado da cultura, AEVI e ANDVI. Os valores das variáveis que definem o quarto cenário são apresentados na tabela 3.20, e do quinto na tabela 3.21. As variáveis de ambiente de produção apresentam os mesmos valores nestes cenários, enquanto que os índices de estado no quarto cenário representam uma cultura com menor biomassa que no quinto cenário[2].

[2]O AEVI e o ANDVI estão associados ao índice de área foliar, que por sua vez está relacionado a quantidade de biomassa da cultura acima do solo.

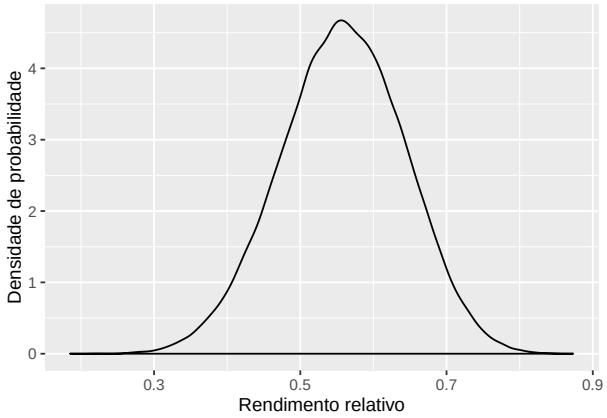


Figura 3.21 – Densidade de probabilidade aproximada da variável rendimento relativo para o cenário 3

Não há sentido em estimar o prêmio do seguro para estes cenários, mas sim a indenização a ser paga, por a disponibilidade dos índices de estado requerer que a cultura já esteja terminando seu desenvolvimento. A estimativa da esperança da indenização a ser paga é, por definição, o prêmio puro do seguro sendo cada um considerado em situações diferentes (um no momento de contratação do seguro, e o outro durante o ciclo da cultura). A distribuição de probabilidade ajustada $P(Y|AP, IE)$ pode ser usada para simular uma amostra de rendimentos relativos para este cenário.

Tabela 3.20 – Valores de variáveis observadas no cenário 4

Variáveis	Valores
Soma térmica	1700,00
Precipitação acumulada	600,00
Maior período sem chuva	15,00
AEVI	0,40
ANDVI	0,50

Tabela 3.21 – Valores de variáveis observadas no cenário 5

Variáveis	Valores
Soma térmica	1700,00
Precipitação acumulada	600,00
Maior período sem chuva	15,00
AEVI	0,70
ANDVI	0,80

O processo de simulação por MCMC apresentou uma taxa de aceitação de 39% para o quarto cenário, e de 47% para o segundo cenário. Em ambos os casos o processo de simulação convergiu rapidamente para a distribuição de interesse, para

50000 iterações. Mesmo assim, o primeiro terço de valores simulados foi descartado. Uma representação gráfica da série de valores simulados é apresentada na figura 3.22.

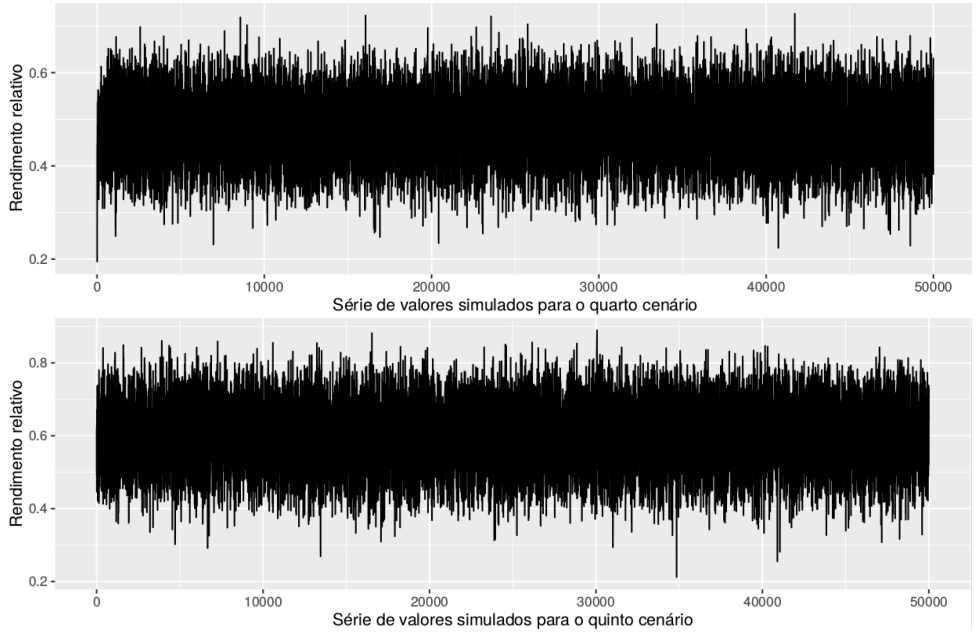


Figura 3.22 – Séries de valores simulados de rendimento relativo por MCMC, a partir da distribuição condicional P(Y|AP,IE).

A densidade de probabilidade aproximada da amostra simulada para o quarto cenário é apresentada na figura 3.23. A densidade aproximada para o quinto cenário é apresentada na figura 3.24. Verifica-se que para mesmos valores para as variáveis de ambiente de produção a distribuição de rendimento relativo é bastante influenciada pelos valores dos índices de estado. A média de rendimento relativo no quarto cenário passa de 48% para 60% no quinto cenário. Essa variação tem um significado importante: embora a disponibilidade de informação das variáveis de ambiente de produção contribuam bastante para as estimativas de rendimento relativo de produção, a informação trazida pelos índices de estado pode sobrepor a informação daquelas variáveis. Em outras palavras, os índices de estado são uma forma de verificar como a cultura respondeu as condições do ambiente de produção. Isto é importante quando se quer considerar a variabilidade espacial de tecnologia de produção, por exemplo.

A indenização estimada para o quarto cenário é 2.12, enquanto que para o quinto cenário ela é 0.26122. Elas refletem a informação trazida pelas figuras recém apresentadas, que no quarto cenário a cultura foi menos prejudicada pelos fatores ambientais de produção explicados pelo modelo em comparação com o quinto cenário.

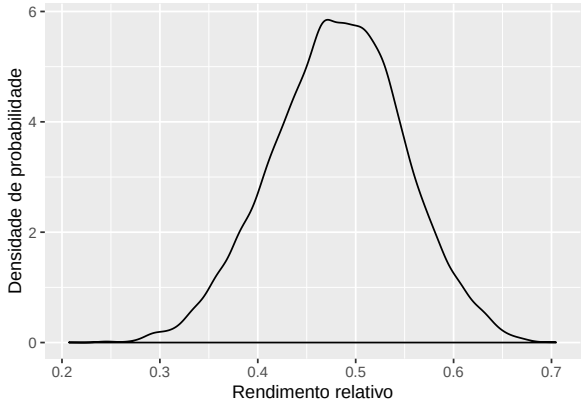


Figura 3.23 – Densidade de probabilidade aproximada da variável rendimento relativo para o cenário 4

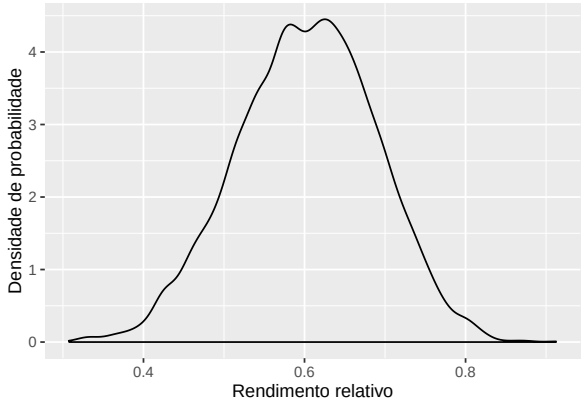


Figura 3.24 – Densidade de probabilidade aproximada da variável rendimento relativo para o cenário 5

3.4 Conclusão

Considerando as limitações da abordagem de modelagem, como disponibilidade de dados e simplicidade do modelo, algumas conclusões podem ser tiradas.

A modelagem baseada em redes bayesianas apresenta vantagens de flexibilidade de interpretação da estrutura do modelo. Em especial, a modulação do modelo a partir de um grafo da rede bayesiana permite dar uma interpretação física e agronômica a um modelo empírico. Se a abordagem de ajuste é aceitável, ganha-se grande simplicidade no ajuste de um modelo complexo.

A modulação do modelo ainda permite a aplicação do mesmo modelo em momentos diferentes do ciclo de produção agrícola para obter informações diferentes (úteis para cada momento), e associar dados meteorológicos a dados de sensoriamento remoto sem que haja necessidade de observações de todas as variáveis estarem disponíveis.

Referências

Caicedo, D. R.; Torres, J. M. C.; Cure, J. R. Comparison of eight degree-days estimation methods in four agroecological regions in Colombia. **Bragantia**, Campinas, SP, v. 71, n. 2, p. 299 –307, 2012.

Camargo, M. B. P. de; Brunini, O.; Miranda, M. A. C. de. Temperatura-base para cálculo dos graus-dia para cultivares de soja em São Paulo. **Pesquisa Agropecuária Brasileira**, Brasilia, DF, v. 22, n. 2, p. 115–121, 1987.

Delignette-muller, M. L.; Dutang, C. fitdistrplus : An R Package for Fitting Distributions. **Journal of Statistical Software**, Los Angeles, US, v. 64, n. 4, p. 1–34, 2015.

Diggle, P. J.; Ribeiro, P. J. **Model-based Geostatistics**. New York, US: Springer, 2007. p. 248.

Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B. **Bayesian Data Analysis**. 3ª ed. Boca Raton, US: Chapman e Hall/CRC, 2013. p. 675.

Goodwin, B.; Ker, A. Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. **American Journal of Agricultural Economics**, Ames, US, v. 80, n. 1, p. 139–153, 1998.

Grün, B; Kosmidis, I; Zeileis, A. Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. **Journal of Statistical Software**, Los Angeles, US, v. 48, p. 1–25, 2012.

Hengl, T.; Heuvelink, G. B. M.; Tadie, M. P.; Pebesma, E. J. Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. **Theoretical and Applied Climatology**, Wien, AT, v. 107, n. 1-2, p. 265–277, 2012.

Huete, a.; Didan, K.; Miura, T.; Rodriguez, E. P.; Gao, X.; Ferreira, L. G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, New York, US, v. 83, n. 1-2, p. 195–213, 2002.

Jensen, J. R. **Sensoriamento remoto do ambiente:** uma perspectiva em recursos terrestres. São José dos Campos: Parêntese Editora, 2009. p. 598.

Kilibarda, M.; Hengl, T.; Heuvelink, G.; Gräler, B.; Pebesma, E.; Tadi'c, M. P.; Bajat, B. Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. **Journal of Geophysical Research: Atmospheres**, Washington, US, v. 119, n. 5, p. 2294–2313, 2014.

Mello, M. P.; Risso, J.; Atzberger, C.; Aplin, P.; Pebesma, E.; Vieira, C. A. O.; Rudorff, B. F. T. Bayesian Networks for Raster Data (BayNeRD): Plausible Reasoning from Observations. **Remote Sensing**, Postfach, CH, v. 5, n. 11, p. 5999–6025, 2013.

Ozaki, V. A. Pricing farm-level agricultural insurance: a Bayesian approach. **Empirical Economics**, Heildenberg, DEU, v. 36, n. 2, p. 231–242, 2009.

Pereira, A. R.; Angelocci, L. R.; Sentelhas, P. C. **Agrometeorologia:** fundamentos e aplicações práticas. Guaíba: Agropecuária, 2002. p. 478.

Pupin Mello, M.; Rudorff, B. F. T.; Adami, M.; Rizzi, R.; Aguiar, D. a.; Gusso, A.; Fonseca, L. M. G. A simplified Bayesian Network to map soybean plantations. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYPOSIUM, 2010, Honolulu, US. **Anais ...** New York, US: IEEE, 2010. p. 351–354.

Rasmussen, S.; Madsen, A. L.; Lund, M. *Bayesian network as a modelling tool for risk management in agriculture*. IFRO Working Paper 2013/12. University of Copenhagen, Department of Food e Resource Economics, 2013.

Robert, C. P.; Casella, G. **Monte Carlo Statistical Methods**. 2ª ed. Springer Texts in Statistics. New York, NY: Springer, 2004. p. 649.

Steduto, P; Raes, D; Hsiao, T.; Fereres, E. AquaCrop: a new model for crop prediction under water deficit conditions. **Options Méditerranéennes : Série A. Séminaires Méditerranéens**, Rome, IT, v. 80, p. 285–292, 2008.

Wu, T.; Li, Y. Spatial interpolation of temperature in the United States using residual kriging. **Applied Geography**, Oxford, UK, v. 44, n. 10, p. 112–120, 2013.

4 CONSIDERAÇÕES FINAIS GERAIS

Nesta seção algumas considerações gerais sobre a pesquisa são apresentadas, enfatizando os pontos fracos que podem ser abordados em pesquisas futuras.

Risco modelado – a metodologia desta pesquisa estima a distribuição de probabilidade da variável rendimento utilizando algumas variáveis meteorológicas que estão intimamente relacionadas aos impactos de eventos de secas. Outros riscos climáticos, como o de incêndios, granizo, vendaval, ou excesso de chuva na colheita não são abordados pelo modelo.

Interpolação espacial dos dados pontuais – Os dados meteorológicos interpolados não refletem a realidade das variáveis meteorológicas locais. A interpolação limita o domínio da variável meteorológica por ser resultante de uma combinação dos valores observados em outros lugares. Além disso, desconsidera-se a dinâmica que pode ser modelada por modelos físicos de clima.

Identificação de ciclo da cultura – Embora o uso de um algoritmo de identificação de ciclo contribua para aumentar a representatividade dos dados usados na modelagem, há um risco considerável de prejuízo na parametrização do modelo. Tal algoritmo é mais eficiente para identificar ciclos de culturas com desenvolvimento normal. Em anos com quebra de safra, o comportamento das séries de índices de vegetação é alterado requerendo maior sensibilidade do algoritmo. É possível que ciclos de culturas em anos de quebra de safra sejam identificados com menor precisão levando a redução da representatividade dos dados para quebras de safras. Esse fato não é desejável para um modelo que requer dados dessas condições. Além do erro de estimativa decorrente do algoritmo, a precisão da identificação de ciclo é dependente da resolução temporal dos dados de índices de vegetação usados no processo.

Estádios críticos da cultura – É sabido que alguns estádios de desenvolvimento da cultura da soja são mais críticos que outros. Em certos estádios a cultura é mais sensível a déficits hídricos que em outros. O presente modelo não leva em conta a modelagem de estádios das culturas.

Agrupamento de dados de rendimento – O agrupamento dos dados de rendimento por município reduz a quantidade de observações para a parametrização do

modelo, bem como transforma a variável rendimento em rendimento relativo médio municipal. A representatividade da variabilidade do rendimento para um certo produtor é reduzida.

Referências espaciais incorretas – Em um estudo que envolve processamento de dados espacializados, há necessidade do estabelecimento de correspondências espaciais entre variáveis. Para tanto supõe-se a disponibilidade de referências espaciais (coordenadas geográficas) corretas. Um exemplo é a correspondência entre dados de temperatura provenientes de estações meteorológicas e os provenientes de sensoriamento remoto. Se as coordenadas que indicam a posição da estação meteorológica não são corretas, haverá perda de qualidade da análise baseada na relação entre tais variáveis. É possível relevar que erros pequenos de posicionamento possam ser ignorados se a pressuposição de que na dimensão espacial objetos próximos tendem a ser parecidos é válida.

Variáveis agregadas para o período do ciclo da cultura – A identificação do ciclo da cultura é fundamental para amostragem dos dados em séries temporais de dados relacionados a tal ciclo. No entanto, cuidado deve ser tomado. Variáveis agregadas, como soma térmica e precipitação acumulada, são extremamente influenciadas pelo resultado do algoritmo de identificação de ciclos de culturas. Tratando-se de observações em escala diária, alguns dias a mais que o ciclo compreende podem resultar em variabilidade nas variáveis de interesse que não é descrita pelos modelos propostos. Isso interfere nas análises relacionadas ao modelo de interesse. Portanto, alguma metodologia de estimação de uma estatística representativa deve ser usada ao invés de usar a simples variável agregada.

Soma térmica e comprimento de ciclo – Recorrentemente na literatura verifica-se a utilização de soma térmica para estimar o comprimento do ciclo da cultura. Pressupõe-se que uma cultura (variedade) possua um comprimento de ciclo dependente de soma térmica. Dessa forma, usar soma térmica como variável explicativa do rendimento de uma cultura e a identificação de comprimento de ciclo aparentam ser redundantes.

Redes bayesianas na modelagem agronômica – As redes bayesianas possuem grande potencial na modelagem de produtividade de culturas com incerteza. Em especial, o diagrama acíclico direto (grafo) de uma rede bayesiana permite representar

as relações de causalidade no modelo probabilístico de forma intuitiva e com semelhanças com modelos determinísticos. Alem disso, a rede bayesiana dá flexibilidade no processo de inferência quando observações de algumas variáveis que compõe o modelo não estão disponíveis.

Variáveis que compõe o modelo, e seu ajuste – As variáveis usadas na modelagem de rendimento agrícola relativo juntamente com a falta de representatividade dos dados usados nos ajustes dos modelos deixaram o modelo simplista e subestimaram o risco de seca nos cenários simulados.

Por fim, é possível dizer que há muito a ser melhorado em metodologias de processamento de dados meteorológicos conjuntamente com dados de sensoriamento remoto para aplicação em seguro agrícola e monitoramento de culturas. A seleção de dados nas dimensões temporal e espacial têm grande apelo no contexto de acompanhamento de culturas agrícolas, em especial quando análise de dados agrupados estão em questão, e a representação de modelos de produtividade por redes bayesianas pode facilitar a estruturação e ajuste desses modelos e a inferência com eles.

ANEXOS

ANEXO A - Tabela de municípios

A seguinte tabela apresenta os municípios nos quais as análises desta pesquisa se basearam

Tabela 1 – Tabela de municípios utilizados na pesquisa.

(continua)

Estado	Município	Geocódigo
PR	Altamira do Paraná	4100459
PR	Araruna	4101705
PR	Barbosa Ferraz	4102505
PR	Boa Esperança	4103008
PR	Boa Ventura de São Roque	4103040
PR	Brasilândia do Sul	4103370
PR	Campo Mourão	4104303
PR	Cândido de Abreu	4104402
PR	Candói	4104428
PR	Cantagalo	4104451
PR	Corumbataí do Sul	4106555
PR	Engenheiro Beltrão	4107504
PR	Faxinal	4107603
PR	Fênix	4107702
PR	Goioerê	4108601
PR	Guarapuava	4109401
PR	Iretama	4110805
PR	Ivaiporã	4111506
PR	Janiópolis	4112207
PR	Juranda	4112959
PR	Luiziana	4113734
PR	Mamborê	4114005
PR	Manoel Ribas	4114500
PR	Moreira Sales	4116109
PR	Nova Santa Rosa	4117222

Tabela 1 – Tabela de municípios utilizados na pesquisa.

(conclusão)

Estado	Município	Geocódigo
PR	Ouro Verde do Oeste	4117453
PR	Palmital	4117800
PR	Peabiru	4118808
PR	Pinhão	4119301
PR	Pitanga	4119608
PR	Quarto Centenário	4120655
PR	Quinta do Sol	4121109
PR	Rancho Alegre	4121307
PR	Roncador	4122503
PR	São Jorge do Ivaí	4125308
PR	São Pedro do Iguaçu	4125753
PR	Toledo	4127700
PR	Tupãssi	4127957
SC	Abelardo Luz	4200101
SC	Ouro Verde	4211850
SC	São Domingos	4216107