Universidade Federal do Rio de Janeiro

UMA METODOLOGIA PARA PREDIÇÃO DE PRODUTIVIDADE DA CANA-DE AÇÚCAR NO ESTADO DE SÃO PAULO

Jeferson Lobato Fernandes

2015

UMA METODOLOGIA PARA PREDIÇÃO DE PRODUTIVIDADE DA CANA-DE AÇÚCAR NO ESTADO DE SÃO PAULO

Jeferson Lobato Fernandes

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro
Dezembro de 2015

UMA METODOLOGIA PARA PREDIÇÃO DE PRODUTIVIDADE DA CANA-DE AÇÚCAR NO ESTADO DE SÃO PAULO

Jeferson Lobato Fernandes

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

_____

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____

Prof[a]. Beatriz de Souza Leite Pires Lima, D.Sc.

_____

Prof. Elton Fernandes, Ph.D.

_____

Prof. Helio José Correa Barbosa, D.Sc.

_____

Dr. Júlio Cesar Dalla Mora Esquerdo, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2015

Fernandes, Jeferson Lobato

Uma Metodologia para Predição de Produtividade da Cana-de-açúcar no Estado de São Paulo/ Jeferson Lobato Fernandes. – Rio de Janeiro: UFRJ/COPPE, 2015.

XIII, 92p.: il.; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2015.

Referências Bibliográficas: p. 81-88.

1. Predição de produtividade. 2. Multi-Layer Perceptron. 3. Comitês. I. Ebecken, Nelson Francisco Favilla. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

Dedicatória

Dedico este trabalho à minha esposa, Juliana, que está sempre ao meu lado me apoiando e me encorajando a descobrir novos caminhos.

Agradecimentos

Agradeço,

Ao Programa de Engenharia Civil da COPPE/UFRJ pela possibilidade de realização deste doutorado;

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, pela bolsa concedida;

Ao meu orientador Prof. Dr. Nelson Francisco Favilla Ebecken, pela oportunidade concedida e pelas valiosas contribuições ao longo do trabalho;

Ao Dr. Júlio César Dalla Mora Esquerdo, amigo que muito me apoiou na execução deste trabalho;

À banca examinadora da qualificação, Dras. Beatriz de Souza Leite Pires Lima e Myriam Christina Aragão, que avaliaram meu projeto e me alertaram sobre o excesso de conteúdo, possibilitando assim obter uma proposta mais realista;

Aos funcionários do PEC;

Aos membros da banca;

À minha esposa Juliana, pela paciência e compreensão ao longo deste trabalho;

e à Deus, sempre.

Obrigado!

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.).

UMA METODOLOGIA PARA PREDIÇÃO DE PRODUTIVIDADE DA CANA-DE AÇÚCAR NO ESTADO DE SÃO PAULO

Jeferson Lobato Fernandes

Dezembro/2015

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

O atual sistema de previsão de safras para a cultura da cana-de-açúcar utilizado no Brasil depende, em boa parte, de informações subjetivas, baseadas no conhecimento de técnicos do setor e informações da cadeia produtiva. O objetivo deste estudo é avaliar o uso de dados espectrais do sensor *MODerate-resolution Imaging Spectroradiometer* (MODIS) e dados meteorológicos provenientes do modelo do *European Center for Medium-Range Weather Forecast* (ECMWF) aplicados a um comitê de redes neurais para predizer a produtividade da cana-de-açúcar em escala municipal e estadual no estado de São Paulo, Brasil. Foram selecionados 60 municípios e nove safras entre os anos de 2003 e 2012. Foi feita a seleção de variáveis usando um *wrapper* com rede neural Multi-Layer Perceptron (MLP) e, a partir do banco de dados selecionado, foi feita a predição da produtividade usando os métodos *Bagging* e *Stacking* de criação de comitês. Foi estimada também a produtividade para todo o Estado baseado na média aritmética dos resultados obtidos com o *Stacking* para os 60 municípios, comparando estes resultados com as estimativas feitas pela Companhia Nacional de Abastecimento – Conab. Foi possível predizer a produtividade da cana-de-açúcar para o estado de São Paulo com menor RMSE (1,4 t ha^{-1}) que as estimativas da Conab (cerca de 4 t ha^{-1}), com três meses de antecipação ao final da safra. O método objetivo e de baixo custo proposto neste estudo pode contribuir para a melhoria do atual método oficial de estimativa da safra de cana-de-açúcar realizado pela Conab.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

METHODOLOGY TO PREDICT SUGARCANE YIELD IN THE STATE OF SÃO PAULO, BRAZIL

Jeferson Lobato Fernandes

December/2015

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

The current crop yield forecasting system for sugarcane used in Brazil is dependent on subjective information which are based on personal knowledge of technicians and information from the industry. The objective of this study is to evaluate the use of spectral data of MODerate-resolution Imaging Spectroradiometer (MODIS) sensor and meteorological data provided by the European Center for Medium-Range Weather Forecast (ECMWF) model in an ensemble of Artificial Neural Network (ANN) to predict sugarcane yield at municipal and State level in the State of São Paulo, Brazil. Sixty municipalities and nine crop seasons between 2003 and 2012 were selected in the State of São Paulo. An ANN wrapper with sequential backward selection was applied in order to eliminate irrelevant and/or redundant features from the initial data set. Sugarcane yield was predicted using Bagging and Stacking ensemble methods. Finally, it was estimated the yield for the State based on the average of the 60 municipalities, which were then compared to surveys of the *Companhia Nacional de Abastecimento* - Conab. Stacking was capable of estimating sugarcane yield for the São Paulo State with smaller RMSE (1.4 t ha^{-1}) than Conab surveys (about 4 t ha^{-1}), with anticipation of three months before the end of the crop season. The low cost and objective method proposed in this study can contribute to improve the current official method of estimating sugarcane production provided by Conab.

Sumário

Capítulo 1 1
1.1. Introdução 1
1.2. Objetivos 4
Capítulo 2 – Revisão Bibliográfica 5
2.1. A cana-de-açúcar 5
2.1.1. O ciclo de desenvolvimento da cana-de-açúcar 7
2.1.2. Condições meteorológicas e a produtividade da cana-de-açúcar 9
2.1.2.1. Condições hídricas 10
2.1.2.2. Radiação solar 11
2.1.2.3. Temperatura 12
2.2. Previsão de safras 12
2.2.1. Métodos governamentais de previsão de safras 13
2.2.1.1. Previsão de safras na América do Norte 13
2.2.1.2. Previsão de safras na Europa 14
2.2.1.3. Previsão de safras no Brasil 15
2.2.2. Predição de produtividade para a cana-de-açúcar 16
2.2.2.1. Modelos de simulação baseados em processos fisiológicos 16
2.2.2.2. Sensoriamento remoto na predição de produtividade 17
2.2.2.3. Dados meteorológicos em modelos de predição de produtividade 19
2.2.3. Descoberta de conhecimento na predição da produtividade agrícola 19
2.3. Descoberta de conhecimento em banco de dados 21
2.3.1. Redes Neurais Artificiais 22
2.3.1.1. Parada antecipada 25
2.3.2. Seleção de variáveis 26
2.3.2.1. *Wrapper* 27
2.3.3. Comitê de modelos preditores 29
2.3.3.1. *Bagging* 31
2.3.3.2. *Stacked Generalization* 32
2.4. O futuro da previsão de safras para a cana-de-açúcar no Brasil 33
Capítulo 3 – Material e Métodos 36
3.1. Preparação dos dados 37

3.1.1. Definição da área e período de estudo 37
3.1.2. Obtenção dos dados e construção de variáveis 39
3.1.2.1. Dados de produtividade de biomassa 39
3.1.2.2. Dados espectrais 40
3.1.2.3. Dados meteorológicos 43
3.1.3. Pré-processamento dos dados 46
3.2. Processamento dos dados 47
3.2.1. Quatro diferentes abordagens para a predição da produtividade 47
3.2.1.1. Modelo espectral municipal 48
3.2.1.2. Modelo espectral-meteorológico municipal 48
3.2.1.3. Modelo reduzido municipal 48
3.2.1.4. Modelo estadual 49
3.2.2. Parâmetros da seleção de variáveis e dos comitês de modelos 49
3.2.2.1. Seleção de variáveis 50
3.2.2.2. Comitês de modelos preditores 50
Capítulo 4 – Resultados e Discussão 52
4.1. Análise exploratória das variáveis construídas 52
4.2. Modelo espectral municipal 55
4.2.1. Seleção de variáveis 55
4.2.2. Predição por comitês 58
4.3. Modelo espectral-meteorológico municipal 61
4.3.1. Seleção de variáveis 61
4.4. Modelo reduzido municipal 65
4.4.1. Seleção de variáveis 66
4.4.2. Predição por comitês 69
4.5. Modelo estadual 72
4.6. Discussões gerais 74
Capítulo 5 – Conclusões e Recomendações 78
Referências Bibliográficas 81
Anexo 1 90
Anexo 2 91
Anexo 3 92

Lista de Figuras

Figura 1: Volume de produção dos cinco principais produtores mundiais de cana-de-açúcar. Fonte: FAOSTAT (2015)..... 5
Figura 2: Porcentagem de área plantada das cinco principais culturas cultivadas no Brasil. Fonte: SIDRA (2015)..... 6
Figura 3: Porcentagem de área plantada das cinco principais culturas cultivadas no estado de São Paulo. Fonte: SIDRA (2015)..... 6
Figura 4: Ciclo de uma cultura de cana-de-açúcar com cinco cortes. Fonte: Modificado de Rodrigues (2013)..... 8
Figura 5: Ciclo de desenvolvimento e acúmulo de matéria seca da cana de ano e cana de ano-e-meio e variações climáticas da região centro sul do Brasil. Fonte: Modificado de Fortes (2003)..... 9
Figura 6: Modelos de redes neurais existentes, de acordo com o paradigma de aprendizagem, arquitetura, conectividade e método de aprendizagem..... 23
Figura 7: Neurônio artificial de McCulloch-Pitt..... 24
Figura 8: Ilustração de uma rede neural MLP com uma camada intermediária..... 25
Figura 9: Exemplo de treinamento com parada antecipada..... 26
Figura 10: Procedimento de seleção de variáveis do *wrapper para frente*..... 28
Figura 11: Processo de seleção de variáveis do *wrapper para trás*..... 29
Figura 12: Criação de um comitê de modelos por meio de *Bagging*..... 31
Figura 13: Criação de um comitê de modelos por meio de *Stacking*..... 32
Figura 14: Série histórica da produção brasileira de cana-de-açúcar, entre os anos de 1990 e 2013. Fonte SIDRA (2015)..... 34
Figura 15: Fluxograma geral das etapas realizadas..... 36
Figura 16: Mapeamento da área plantada com cana-de-açúcar na safra 2012-2013 no estado de São Paulo..... 37
Figura 17: Classificação climática de Koeppen do estado de São Paulo. Fonte: Adaptado de CEPAGRI (2014)..... 38
Figura 18: Mapa de localização dos 60 municípios selecionados no estado de São Paulo..... 39
Figura 19: Exemplo de séries temporais de NDVI médio obtidas para o município de Altinópolis, para as safras 2004, 2005 e 2006..... 41

Figura 20: Alguns dos parâmetros gerados pelo TIMESAT: a) início da safra, b) final da safra, c) duração da safra, d) valor máximo, e) amplitude e f) integral. Fonte: Adaptado de Jönsson e Eklundh (2004)..... 42
Figura 21: Mapa da distribuição da precipitação acumulada entre os dias 01 e 10 de janeiro de 2004 no Brasil proveniente de imagem decendial do modelo do ECMWF...45
Figura 22: Histograma da variável "rendimento médio da produção"..... 53
Figura 23: Box-plot das variáveis que apresentaram pontos extremos..... 53
Figura 24: Matriz de correlação contendo todas as variáveis iniciais..... 54
Figura 25: Gráficos de dispersão entre dados observados e dados preditos de produtividade, para uma simples MLP e o banco de dados completo, para uma simples MLP e as variáveis selecionadas (b), para o *Bagging* de 30 membros (c) e para o *Stacking* de dois membros (d)..... 60
Figura 26: Gráfico de dispersão entre dados observados e dados preditos de produtividade para o *Stacking* de cinco membros no modelo reduzido municipal..... 71

Lista de Tabelas

Tabela 1: Variáveis espectrais geradas a partir da série temporal municipal de NDVI..... 43
Tabela 2: Variáveis meteorológicas criadas a partir dos dados meteorológicos iniciais..... 46
Tabela 3: Estatísticas básicas da variável "rendimento médio da produção", em [t ha^{-1}]..... 52
Tabela 4: Impacto (I) no desempenho ao se retirar cada variável do banco de dados espectral, para cada etapa do *wrapper*..... 56
Tabela 5: Resultados da seleção de variáveis aplicando o *wrapper* de seleção sequencial para trás sobre o banco de dados espectral..... 56
Tabela 6: Variáveis selecionadas para o modelo espectral municipal, em ordem decrescente de importância..... 57
Tabela 7: Resultados de RMSE [t ha^{-1}] e R^2 para as diferentes quantidades de membros no *Bagging* do modelo espectral municipal..... 58
Tabela 8: Resultados de RMSE [t ha^{-1}] e R^2 para diferentes quantidades de membros no *Stacking* do modelo espectral municipal..... 59
Tabela 9: Impacto (I) no desempenho ao se retirar cada variável do banco de dados espectral-meteorológico, para cada etapa do *wrapper*..... 62
Tabela 10: Resultados da seleção de variáveis aplicando o *wrapper* de seleção sequencial para trás sobre o banco de dados espectral-meteorológico..... 64
Tabela 11: Variáveis selecionadas para o modelo espectral-meteorológico municipal, em ordem decrescente de importância..... 64
Tabela 12: Impacto (I) no desempenho ao se retirar cada variável do banco de dados espectral-meteorológico reduzido, para cada etapa do *wrapper*..... 66
Tabela 13: Resultados da seleção de variáveis aplicando o *wrapper* de seleção sequencial para trás sobre o banco de dados espectral-meteorológico reduzido..... 68
Tabela 14: Variáveis selecionadas para o modelo reduzido municipal, em ordem decrescente de importância..... 69
Tabela 15: Resultados de RMSE [t ha^{-1}] e R^2 para diferentes quantidades de membros no *Bagging* do modelo reduzido municipal..... 69

Tabela 16: Resultados de RMSE [t ha^{-1}] e R^2 para diferentes quantidades de membros no *Stacking* do modelo reduzido municipal..... 70
Tabela 17: Resultados da produtividade estimada [t ha^{-1}], RMSE [t ha^{-1}] e R^2 para o MEE, MRE e para os três levantamentos da Conab..... 72
Tabela 18: Resultados de cinco estudos de predição de produtividade de biomassa da cana-de-açúcar..... 76

Capítulo 1

1.1. Introdução

O Brasil é o maior produtor mundial de cana-de-açúcar, produzindo cerca de 40% da produção mundial na safra 2013 (Faostat, 2015). E grande parte da produção canavieira brasileira está concentrada no estado de São Paulo, responsável por aproximadamente 60% da produção nacional (SIDRA, 2015). Além de ser uma fonte de alimento, a cana-de-açúcar também é usada como importante fonte de energia renovável, por meio da produção do etanol e da queima do bagaço. Os produtos da cana-de-açúcar já ocupam a segunda posição na composição da matriz energética brasileira, correspondendo a 18,1% da produção de energia primária (EPE, 2015).

Devido a esta importância, é fundamental que o governo brasileiro apresente estatísticas de previsão de safras internacionalmente confiáveis, uma vez que estas informações subsidiarão, entre outras atividades, a regulação dos preços internacionais. Além disso, estimar a produção é extremamente importante ao planejamento e execução de uma série de atividades no campo, indústria e mercado, auxiliando diversas tomadas de decisões.

Os órgãos responsáveis pela previsão de safras no Brasil são a Companhia Nacional de Abastecimento (Conab) e o Instituto Brasileiro de Geografia e Estatística (IBGE). A Conab realiza regularmente levantamentos sobre o acompanhamento da safra brasileira de cana-de-açúcar, onde técnicos visitam as unidades de produção e os dados são coletados por meio de questionários. O IBGE realiza a estimativa da produção da cana-de-açúcar por meio do Levantamento Sistemático da Produção Agrícola (LSPA), método similar ao da Conab (Conab, 2015; IBGE, 2015). Trata-se de métodos onerosos e demorados, além de conterem um grau de subjetividade. É necessário que se desenvolvam métodos objetivos de previsão de safras em nível regional. E para estimar a produção agrícola primeiramente se necessita predizer a produtividade. No entanto, predizer a produtividade da cana-de-açúcar não é tarefa fácil, pois se trata de um sistema produtivo complexo. Condições meteorológicas, características dos solos, variáveis de manejo – como data de plantio/rebrota, idade do canavial, variedade

genética e aplicação de insumos – além de ocorrência de pragas e doenças influenciam a produtividade final da cultura.

Existem alguns modelos de simulação baseados em processos fisiológicos em uso ao redor do mundo que permitem simular para a cana-de-açúcar a produção de biomassa, de sacarose, uso de água, captação de nitrogênio, entre outras. Trata-se de modelos complexos e que necessitam de um grande número de variáveis medidas em campo, além de terem sido desenvolvidos em condições diferentes e que ainda necessitam de estudos para verificar a viabilidade de uso no Brasil. Além destes modelos, diversos estudos têm sido realizados com o intuito de relacionar os valores de produtividade da cana-de-açúcar com dados sobre condições hídricas, radiação solar, temperatura, dados de solo, dados de manejo e dados espectrais. Porém, de maneira geral, são estudos que costumam englobar pequenas áreas e/ou poucas safras. As relações na maioria das vezes são modeladas por meio de modelos lineares.

Uma alternativa para viabilizar a construção de modelos de produtividade em grande escala espacial é o uso de dados extraídos de imagens de satélites, como os índices de vegetação, que podem fornecer informações importantes sobre a interação solo-planta-atmosfera. Estes índices vêm sendo cada vez mais usados em estudos agrícolas, como mapeamento de culturas e caracterização do desenvolvimento fisiológico e apresentam grande potencial em estudos de produtividade. Dados de imagens de satélite são de fácil aquisição, e existem imagens adequadas a estudos agrícolas disponibilizadas gratuitamente, o que aumenta o interesse em se incorporar estes dados em modelos de predição de produtividade. Além dos índices de vegetação, existem modelos numéricos capazes de gerar dados meteorológicos, como temperatura, precipitação e radiação solar, a partir de imagens de satélite.

Métodos baseados na descoberta de conhecimento em banco de dados, com o uso de algoritmos de aprendizagem, são uma excelente ferramenta na tarefa de predição. A partir de um conjunto de dados, modelos computacionais com capacidade de aprendizagem podem identificar a estrutura desconhecida dos processos de desenvolvimento das culturas sob diferentes condições. Algoritmos de aprendizagem, como as Redes Neurais Artificiais, são bastante adequadas para o ajuste de relações não-lineares, possibilitando modelar a predição da produtividade. Técnicas de seleção de variáveis possibilitam selecionar um conjunto de variáveis de forma a eliminar aquelas irrelevantes e/ou redundantes que possam existir no conjunto de dados inicial. E

o uso de comitês de modelos pode aumentar a capacidade de generalização e a acurácia do modelo de predição. Trata-se de métodos ainda pouco explorados para a predição de produtividade agrícola, especialmente para a cana-de-açúcar.

O objetivo deste estudo é explorar técnicas de descoberta de conhecimento em banco de dados para obter um modelo de predição de produtividade de biomassa da cana-de-açúcar no estado de São Paulo. A relevância deste trabalho se justifica pelo fato de ainda não haver no Brasil um método objetivo de predição de safras para a cana-de-açúcar que possa ser aplicado em escala regional, o que é de grande necessidade devido à importância do sistema produtivo. Como contribuição, o estudo explorou métodos ainda pouco usados na predição de produtividades agrícolas, métodos estes que possibilitaram avaliar a relevância de novas variáveis a serem incorporadas em modelos de predição de produtividade da cana-de-açúcar como também o ganho em se aplicar um comitê de modelos preditores.

1.2. Objetivos

O objetivo deste trabalho é desenvolver um modelo de predição da produtividade de biomassa da cana-de-açúcar para o sistema produtivo do estado de São Paulo por meio de comitês de modelos de aprendizagem e a partir de um banco de dados gratuito.

Os objetivos específicos são:

- extrair métricas da série temporal do Índice de Vegetação por Diferença Normalizada (*Normalized Difference Vegetation Index* – NDVI) do sensor *MODerate-resolution Imaging Spectroradiometer* (MODIS) que possam ser usadas como variáveis espectrais de entrada em um modelo de predição de produtividade da cana-de-açúcar;

- gerar variáveis meteorológicas a partir de dados meteorológicos simulados pelo modelo do *European Center for Medium-Range Weather Forecast* (ECMWF) que também possam ser usadas, em conjunto com as variáveis espectrais, em um modelo de predição de produtividade da cana-de-açúcar;

- selecionar as variáveis relevantes para uso em um comitê de modelos de redes neurais *Multi-Layer Perceptron* (MLP) para predição da produtividade;

- criar diferentes configurações de comitês de modelos preditores, levando em consideração o algoritmo de aprendizagem MLP e os métodos *Bagging* e *Stacking* para criação de comitês;

- verificar qual o modelo mais adequado para predição da produtividade de biomassa municipal da cana-de-açúcar no estado de São Paulo;

- comparar os resultados do modelo com as previsões de produtividade realizadas pela Conab.

Capítulo 2 – Revisão Bibliográfica

2.1. A cana-de-açúcar

Originária do sudeste asiático, a cana-de-açúcar foi introduzida no Brasil no início do século XVI, no período colonial, e se transformou em uma das principais culturas da economia brasileira. O Brasil é o maior produtor mundial de cana-de-açúcar, produzindo na safra 2012/2013 em torno de 768 milhões de toneladas, seguido por Índia com 341, China com 128, Tailândia com 100 e Paquistão com 64 milhões de toneladas, conforme ilustrado na Figura 1, em que é possível observar a importância da produção brasileira no cenário mundial, produzindo mais que o dobro em relação ao segundo colocado (FAOSTAT, 2015).

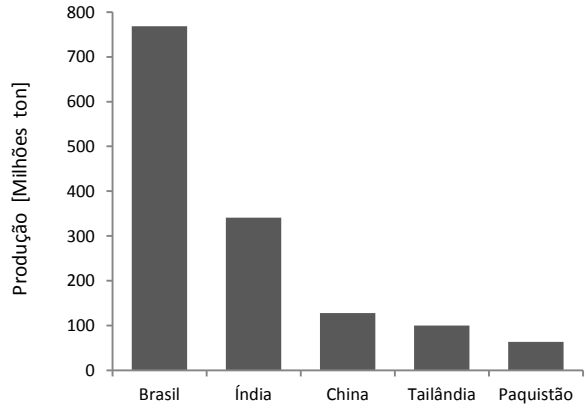


Figura 1: Volume de produção dos cinco principais produtores mundiais de cana-de-açúcar. Fonte: FAOSTAT (2015).

No Brasil a cana-de-açúcar é a terceira cultura com maior área plantada, após soja e milho, conforme mostrado na Figura 2, em que é apresentada a porcentagem de área plantada, em relação à área total ocupada pelas principais culturas brasileiras, das cinco culturas com maior área plantada na safra 2013 (SIDRA, 2015).

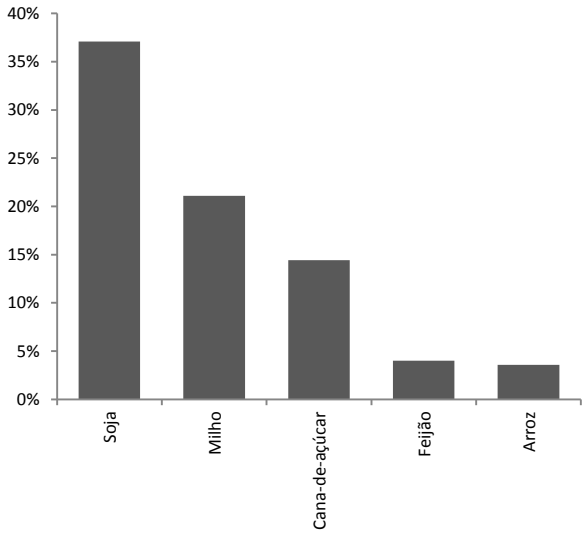


Figura 2: Porcentagem de área plantada das cinco principais culturas cultivadas no Brasil. Fonte: SIDRA (2015).

O estado de São Paulo é responsável por cerca de 60% de toda a produção brasileira de cana-de-açúcar. Na Figura 3 é apresentada a porcentagem de área plantada, em relação à área total ocupada pelas principais culturas no estado de São Paulo, para as cinco culturas com maior extensão de área plantada na safra 2013, onde se pode verificar o destaque para a cana-de-açúcar (SIDRA, 2015).

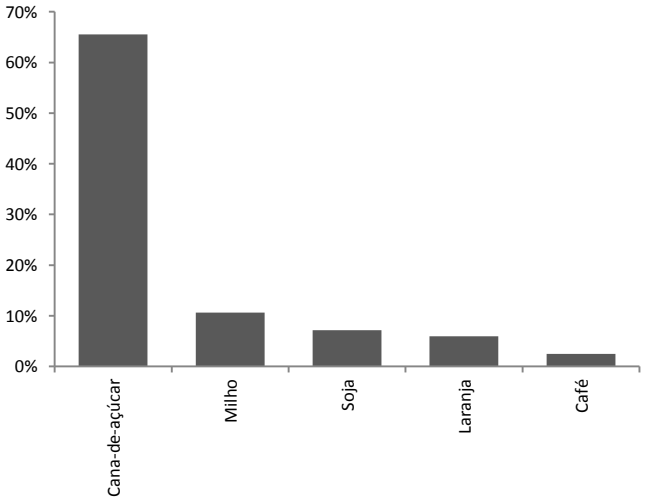


Figura 3: Porcentagem de área plantada das cinco principais culturas cultivadas no estado de São Paulo. Fonte: SIDRA (2015).

Os dois principais derivados da cana-de-açúcar são o açúcar e o etanol. Segundo a Conab (2015) a estimativa para a safra 2015/2016 é que, do total de cana-de-açúcar produzido, aproximadamente 44% sejam destinados à produção de açúcar e 56% para a produção do etanol. O Brasil produziu 28,66 bilhões de litros de etanol na safra 2014/2015 (Conab, 2015).

O etanol produzido a partir da cana-de-açúcar é uma importante fonte de energia renovável e sua produção no Brasil apresenta vantagens ambientais quando comparada à produção de outros combustíveis. Os resíduos da queima deste combustível são menos poluentes quando comparados aos da gasolina. Outra vantagem se refere à energia necessária para a fase industrial da produção do etanol, que provém do bagaço da moagem, inclusive gerando excedentes de energia elétrica que são incorporados ao sistema elétrico nacional. Isso não ocorre, por exemplo, com a produção norte americana de etanol produzido a partir do milho, que exige em sua produção o uso de energia proveniente em grande parte do carvão (BNDES, 2008). Outro aspecto a ser considerado é que ao longo do crescimento da cana-de-açúcar no campo ocorre a absorção do carbono pelas plantas, o que contribui para a redução do carbono atmosférico.

Devido à importância do Brasil na produção mundial de cana-de-açúcar e derivados, é fundamental que se desenvolvam métodos de previsão de safras confiáveis, a fim de subsidiar importantes atividades do setor sucroalcooleiro, como o desenvolvimento de políticas públicas, a organização e planejamento da cadeia produtiva e a regulação de preços internacionais.

2.1.1. O ciclo de desenvolvimento da cana-de-açúcar

A cana-de-açúcar é uma cultura semi-perene[1], com um ciclo que permite em torno de cinco estágios de corte de produtividade rentável economicamente. Geralmente é feita a reforma do canavial após o quinto corte. O ciclo se inicia com o plantio da

[1] Possui características de uma cultura temporária, ou anual – que se extinguem pela colheita, sendo seguidas de um novo plantio – e de uma cultura perene – aquela de duração superior a um ano ou que proporcionam mais de uma colheita, sem a necessidade de novo plantio.

muda, que brota, perfilha[2], matura e recebe o primeiro corte. Essa cana é denominada de cana-planta. A partir do primeiro corte um novo estagio de corte se inicia, dando origem à chamada cana-soca. Na Figura 4 é representado um Ciclo de Reforma com cinco cortes.

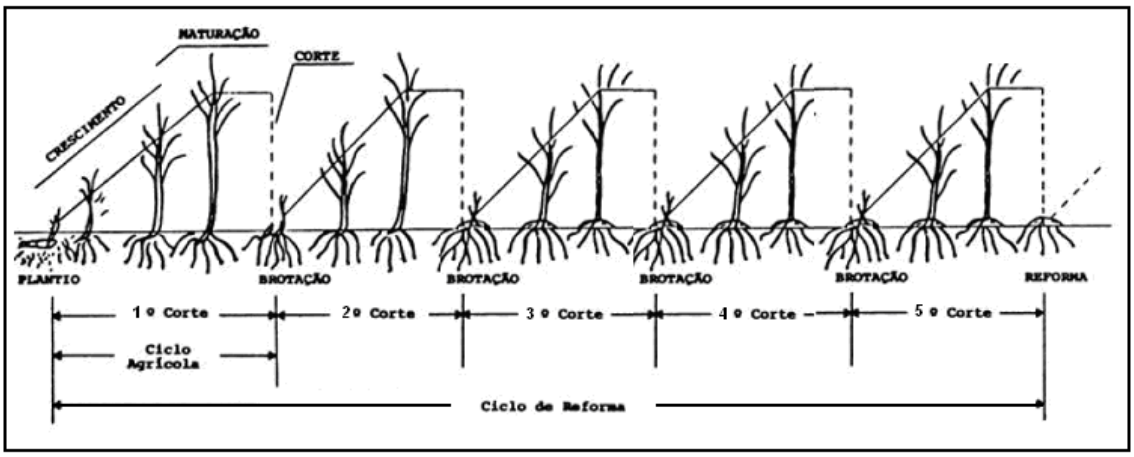


Figura 4: Ciclo de uma cultura de cana-de-açúcar com cinco cortes. Fonte: Modificado de Rodrigues (2013).

O plantio pode ser feito em duas diferentes épocas. No Centro-Sul do Brasil, se o plantio ocorrer em setembro e outubro o primeiro estágio de corte terá duração de aproximadamente um ano, denominada então de cana de ano. Se o plantio ocorrer entre janeiro e abril o primeiro estágio de corte terá duração de cerca de um ano e meio, denominada assim de cana de ano e meio. A partir do primeiro corte os estágios passam a ter duração anual para ambos os tipos. Na Figura 5 é apresentada a dinâmica de cultivo e o acúmulo de matéria seca, para a cana de ano (CA) e a cana de ano e meio (CAM), ao longo do período de crescimento, bem como as variações climáticas para a região Centro-Sul do Brasil.

Cada estágio de corte é dividido em diferentes fases de desenvolvimento. Rodrigues (2013) cita haver três fases ao longo do desenvolvimento: 1-perfilhamento e crescimento lento; 2- crescimento rápido; 3- crescimento lento. Wiedenfeld (2000) divide o estagio de corte em quatro fases: 1- germinação e emergência; 2- perfilhamento

[2] Surgimento de ramos laterais que se desenvolvem a partir das gemas axilares dos nós que se localizam abaixo da superfície do solo.

e estabelecimento do dossel; 3- crescimento rápido; 4- maturação. Marin e Carvalho (2012) dividiram o estagio de corte em três fases: 1- crescimento vegetativo 1; 2- crescimento vegetativo 2; 3- maturação.

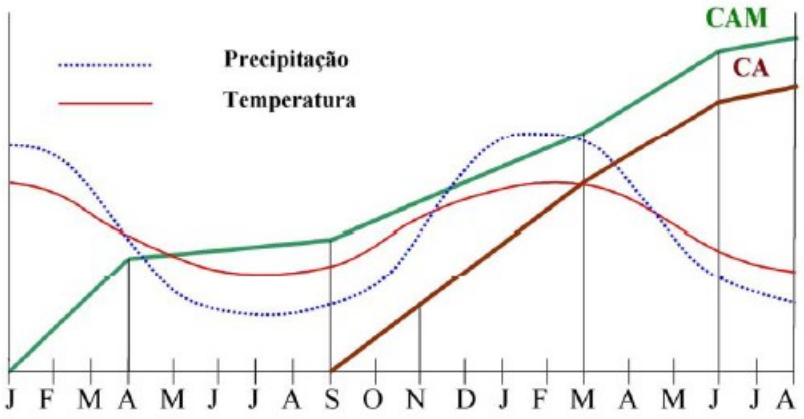


Figura 5: Ciclo de desenvolvimento e acúmulo de matéria seca da cana de ano e cana de ano-e-meio e variações climáticas da região centro sul do Brasil. Fonte: Modificado de Fortes (2003).

Em relação às variedades genéticas, existem aquelas de ciclo precoce, médio e tardio, de acordo com a duração da fase de maturação de cada variedade (Rodrigues, 2013). Isso permite que a colheita seja realizada durante um longo período, que normalmente varia de maio (precoce) a novembro (tardio).

A produtividade de biomassa da cana-de-açúcar é medida pela quantidade de biomassa por unidade de área, normalmente em toneladas por hectare [t ha^{-1}].

2.1.2. Condições meteorológicas e a produtividade da cana-de-açúcar

As condições meteorológicas exercem grande influência no desenvolvimento da cana de açúcar e, consequentemente, nos resultados finais de produtividade. Marin e Carvalho (2012) avaliaram a variabilidade espaço-temporal da eficiência produtiva da cana-de-açúcar no estado de São Paulo e verificaram que os fatores climáticos explicaram 43% da variabilidade da eficiência produtiva, na seguinte ordem de importância: radiação solar, deficiência hídrica, temperatura máxima do ar, precipitação e temperatura mínima do ar.

Avaliando os levantamentos sobre a produção de cana-de-açúcar no Centro-Sul do Brasil entre 2005 e 2012, disponíveis em Conab (2015), verificou-se que o clima foi apontado como principal causador de variações na produtividade. Portanto, o monitoramento de variáveis meteorológicas ao longo da safra é interessante para que se diminuam as incertezas relacionadas às estimativas de produtividade final.

2.1.2.1. Condições hídricas

Segundo Conab (2015), na safra 2005 chuvas inadequadas prejudicaram o desempenho das lavouras no Centro-Sul do Brasil, acarretando em redução de 15% de produtividade no Paraná. Nas safras 2008 e 2009 o excesso de chuva atrasou a colheita, impossibilitando sua ocorrência em algumas áreas. Na safra de 2010 a falta de chuva no final do ciclo de desenvolvimento beneficiou o acúmulo de sacarose, porém prejudicou o desenvolvimento do canavial para a safra seguinte. Em 2011 a estiagem da safra anterior prejudicou o desempenho dos canaviais.

Van den Berg *et al.* (2000) citam que a disponibilidade de água é apontada por produtores do estado de São Paulo como uma das principais causas de variações na produtividade interanual, sendo os veranicos – mais que duas semanas consecutivas sem chuva, entre novembro e março – a maior causa de reduções na produtividade de biomassa. Singels *et al.* (2005) avaliaram o efeito da data de plantio/rebrota, da classe – cana planta ou cana soca – e do cultivar no desenvolvimento do dossel[3] da cana-de-açúcar e interceptação da radiação e concluíram que a data de plantio/rebrota foi a variável que apresentou maior efeito no desenvolvimento do dossel. Sabe-se que o momento da rebrota está diretamente relacionado com a ocorrência das chuvas.

Inman-Bamber e Smith (2005) realizaram uma revisão sobre trabalhos que abordam as relações hídricas na cana-de-açúcar e resposta a déficits hídricos. Os autores verificaram que a resposta da cana-de-açúcar quanto ao uso da água pode ser dividida em três fases: desenvolvimento inicial do dossel, desenvolvimento final do dossel e alongamento da haste e maturação. No desenvolvimento inicial do dossel o stress hídrico apresenta baixo risco, tendo pouco impacto na produtividade final de biomassa e sacarose, embora o aspecto visual pareça severo. Já na fase de desenvolvimento final do

[3] Parte superior da plantação que irá receber a radiação solar diretamente, o topo da plantação.

dossel e alongamento da haste o stress hídrico pode prejudicar consideravelmente a produtividade final, sendo esta a fase que apresenta melhores respostas em plantios irrigados. Na fase de maturação, períodos secos são benéficos à medida que contribuem para o acúmulo de sacarose, mesmo com redução de biomassa. Os resultados de Wiedenfeld (2000) confirmam que as reduções na produtividade da cana-de-açúcar são mais frequentes e severas quando o stress hídrico ocorre na chamada "fase de grande crescimento vegetativo", quando a evapotranspiração[4] é máxima. Stresses hídricos ocorridos durante as outras fases de desenvolvimento (germinação e emergência, perfilhamento e estabelecimento do dossel e maturação) são menos impactantes para a produtividade e qualidade da cultura.

2.1.2.2. Radiação solar

Um dossel bem desenvolvido beneficia a produtividade. Keating *et al.* (1999) em estudo sobre modelagem do sistema de produção de cana-de-açúcar verificaram que informações sobre a diferença na eficiência do uso da radiação e no desenvolvimento do dossel são de grande importância na simulação da produtividade. Em Lisson *et al.* (2005) é relatado que uma melhor captura e utilização da radiação e otimização da duração do ciclo até a colheita poderia incrementar a produtividade. Park *et al.* (2005a), em estudo sobre o declínio do crescimento da cana-de-açúcar na segunda metade do ciclo de desenvolvimento, citam que, uma vez que as demandas por água e nutrientes estão supridas em quantidades suficientes, o recebimento de radiação solar e sua conversão em biomassa irá determinar o potencial produtivo no ambiente. Uma das variáveis de entrada foi a fração de radiação interceptada, estimada a partir de uma equação baseada na lei de Beer, que relaciona a fração de radiação interceptada com o Índice de Área Foliar[5] (IAF). Nassif *et al.* (2012), em estudo de parametrização de modelo de simulação para a cana de açúcar testaram cinco diferentes variedades genéticas e verificaram que os resultados de produtividade poderiam estar associados ao IAF. E Simões *et al.* (2005a) demonstraram haver relação entre o IAF da cana de açúcar e índices de vegetação, como o *Normalized Difference Vegetation Index* – NDVI,

[4] Perda de água de uma plantação para a atmosfera, causada pela evaporação e pela transpiração das plantas.
[5] Índice que se refere à razão entre a área total da superfície das folhas superiores e a unidade de superfície projetada no solo.

extraídos de imagens de satélites. Conclui-se que é possível inferir sobre o desenvolvimento do dossel da cana-de-açúcar a partir de imagens de satélite.

2.1.2.3. Temperatura

A temperatura também é uma variável que apresenta importância no desenvolvimento da planta, embora o estado de São Paulo apresente temperaturas médias bastante adequadas ao cultivo da cana durante toda a safra. Rodrigues (2013) afirma que para temperaturas abaixo de 20°C o crescimento é praticamente nulo e que temperaturas elevadas, acima de 35°C, também diminuem a taxa de crescimento.

Em estudo para a estimativa da maturação por meio de parâmetros edafo-climáticos[6] no município de Pontal, estado de São Paulo, Scarpari e Beauclair (2004) verificaram que a ocorrência de temperaturas abaixo de 20°C (parâmetro denominado graus-dia-negativo) beneficiava o acúmulo de Açúcar Total Recuperável[7] (ATR) quando estas ocorriam com cinco meses antes da colheita. Liu e Bull (2001) citam que, não havendo condições de stress causadas pelos outros fatores, a temperatura é o fator principal que irá determinar o particionamento da fotossíntese entre acúmulo de sacarose e crescimento vegetativo.

2.2. Previsão de safras

A previsão de safras agrícolas pode ser realizada em diferentes escalas espaciais. Órgãos governamentais possuem interesse em previsões de safras regionais, ou seja, para todo um município, estado ou país. Já os produtores agrícolas estão interessados em prever a safra em escala espacial local, compreendido por talhões[8], sítios, fazendas, ou seja, nas áreas plantadas de seu interesse. Neste tópico são apresentados estudos que abordam estas duas escalas espaciais.

[6] Parâmetros definidos através de fatores do meio, tais como o clima, o relevo, a litologia, a temperatura, a umidade do ar, a radiação, o tipo de solo, o vento, a composição atmosférica e a precipitação pluvial.
[7] Quantidade de açúcar disponível na matéria-prima subtraída das perdas no processo industrial.
[8] Unidade mínima de cultivo de uma propriedade.

2.2.1. Métodos governamentais de previsão de safras

Alguns países possuem um sistema de previsão de safras para avaliação da produção agrícola interna e, em alguns casos, da produção externa. Neste tópico são apresentados os métodos governamentais de previsão de safras norte americano, europeu e brasileiro.

2.2.1.1. Previsão de safras na América do Norte

Nos Estados Unidos da América o *National Agricultural Statistics Service* (NASS), pertencente ao *United States Department of Agriculture* (USDA), é o órgão responsável pelas estatísticas agrícolas. O NASS faz as previsões de produtividade baseadas em informações sobre operações nas fazendas, por meio de relatórios dos fazendeiros, contendo uma projeção sobre a produção, medições em campo e também usando um modelo objetivo de medição (USDA, 2011). No modelo objetivo de medição são realizados levantamentos mensais durante a safra sobre a produção nos estados mais representativos em produção. A partir de um mapa de uso da terra e imagens de satélite são obtidas amostras do território e, para as propriedades que estiverem contidas nas amostras, são levantadas informações a respeito do desenvolvimento da cultura, características das plantas, como número de espigas e tamanho (para o milho), quantidade e peso das vagens (para a soja), entre outras. As previsões são feitas usando médias históricas e um modelo de regressão linear simples. Os parâmetros do modelo são ajustados usando 15 anos de dados. É considerado que as condições meteorológicas se manterão normais. Caso haja alguma anormalidade meteorológica as previsões sofrem ajustes. Trata-se de um levantamento de alto custo, por isso realizado apenas nos estados mais representativos (USDA, 2011). Fontes de dados de outras organizações, tanto públicas quanto privadas, também são disponibilizadas para avaliar a acurácia das estimativas de produção e determinar as estimativas finais.

O USDA também monitora e estima a produção nos principais países produtores das commodities agrícolas através do *Foreign Agricultural Service* (FAS). Para isso são usados dados obtidos em levantamentos de especialistas alocados no país de interesse, imagens de satélite e dados meteorológicos.

2.2.1.2. Previsão de safras na Europa

O MARS (*Monitoring Agriculture with Remote Sensing*) e o MCYFS (*Mars Crop Yield Forecasting System*) têm como objetivo monitorar o comportamento da cultura no campo e produzir previsões de produtividade da cultura. Este sistema cobre toda a região da União Européia, além de Maghreb (noroeste da África) e Turquia. O MARS está experimentalmente sendo estendido para outras quatro áreas, entre elas a América do Sul (METAMP, 2011).

De acordo com METAMP (2011) o sistema pode ser dividido em três níveis:

1 – dados meteorológicos são coletados, checados quanto à qualidade, processados e analisados;

2 – um sistema de simulação do desenvolvimento da cultura é aplicado a fim de transformar os dados meteorológicos em efeitos sobre o desenvolvimento da biomassa da cultura. O sistema usado é o CGMS (*Crop Growth Monitoring System*), que faz uso do modelo WOFOST (*WOrld FOod STudies*), que é um modelo de simulação para a análise quantitativa do crescimento e produção de plantações anuais. Trata-se de um modelo determinístico baseado em regressão linear. Nesta etapa são acrescentadas informações auxiliares, como parâmetros do solo, calendário agrícola, práticas de manejo da cultura e parâmetros da cultura. Diversos indicadores específicos sobre cada cultura são produzidos e irão subsidiar a etapa de análise estatística e produção quantitativa da previsão da safra. É neste segundo nível que dados provenientes de imagens de satélites são incorporados ao sistema e usados para gerar índices de desenvolvimento da cultura no campo, das condições do solo e outros indicadores;

3 – Os indicadores produzidos pelos dados meteorológicos, agrometeorológicos[9] e de sensoriamento remoto[10] são relacionados com séries temporais de dados oficiais sobre produtividade e analisados. Nesta etapa diferentes ferramentas estatísticas são usadas: análise de tendências, análise de regressão e análise de cenários/similaridades. O resultado é uma previsão quantitativa da produtividade que, junto com outras informações, são publicados por meio de boletins.

[9] Dados que consideram as características do solo e a influência das condições meteorológicas e climáticas sobre a produção agrícola.
[10] Ver tópico 2.2.2.2.

Para tratar com a incerteza gerada pelo desconhecimento sobre a evolução da safra no momento em que se fez a predição até o momento da colheita, cenários agrometeorológicos são produzidos e analisados. Os cenários são baseados em anos agrometeorológicos similares detectados através de clusters e análise de componentes principais. Uma vez detectada a safra de condições agrometeorológicas similares, a faixa de resultado de produtividade final é atribuída para a safra corrente. (METAMP, 2011).

2.2.1.3. Previsão de safras no Brasil

Os órgãos responsáveis pela previsão de safras no Brasil são a Companhia Nacional de Abastecimento (Conab) e o Instituto Brasileiro de Geografia e Estatística (IBGE).

A Conab realiza sistematicamente levantamentos das safras agrícolas para quantificar e acompanhar a produção brasileira (Conab, 2015). São feitos levantamentos mensais para grãos, com deslocamentos a campo bimestralmente. Em todos os estados produtores da Federação técnicos entram em contato com produtores rurais, agrônomos e técnicos de cooperativas, secretarias de agricultura, órgãos de assistência técnica e extensão rural (oficiais e privados), agentes financeiros e revendedores de insumos (sementes, fertilizantes, corretivos, etc.), informantes e parceiros, com o objetivo de buscar informações sobre a intenção de plantio. Estes levantamentos contêm uma análise das condições climáticas, estimativa da área plantada, estimativa da produção, análise do mercado de insumos e avaliação individual das culturas do algodão, arroz, canola, feijão, milho, soja e trigo. As análises são baseadas em dados históricos sobre as condições meteorológicas e a produtividade (Conab, 2015). Para a avaliação da safra e estimativa da produção de cana-de-açúcar em escala estadual, a Conab baseia-se na pesquisa de campo nos municípios produtores, com aplicação de questionários diretamente aos detentores das informações sobre o sistema agroindustrial sucroalcooleiro, gerando assim os dados que são apresentados em cada levantamento. São realizados três levantamentos para a cana-de-açúcar ao longo de cada safra. São feitas visitas às usinas com periodicidade quadrimestral (Conab, 2015).

O IBGE realiza a estimativa da produção da cana-de-açúcar por meio do Levantamento Sistemático da Produção Agrícola (LSPA), com periodicidade mensal e

abrangendo todos os municípios. Com método similar ao da Conab, são levantadas informações sobre a demanda de insumos agrícolas junto aos produtores, associações de classe e cooperativas, utilização de dados históricos de produtividade e levantamentos e observações de campo sobre o comportamento da cultura em face de ocorrências climáticas e/ou fitossanitárias (IBGE, 2015).

São métodos que não se baseiam em modelos, possuindo um grau de subjetividade. Além disso, existem divergências entre os números apresentados pelos diferentes órgãos, como na safra brasileira 2011/2012 em que o IBGE (SIDRA, 2015) divulgou uma produção brasileira de cana-de-açúcar de 734.006.059 toneladas, enquanto a Conab (2015) divulgou 560.955.200 toneladas.

Analisando os três métodos governamentais de previsão de safras verifica-se que o norte americano e o europeu são métodos objetivos, que fazem uso de modelos determinísticos. Já o método brasileiro é alimentado por informações obtidas por terceiros, não fazendo uso de modelos matemáticos.

2.2.2. Predição de produtividade para a cana-de-açúcar

Diversos estudos e modelos vêm sendo desenvolvidos com o objetivo de se prever a produtividade da cana-de-açúcar. Basicamente, existem os modelos baseados em processos fisiológicos, que geralmente requerem um grande número de variáveis detalhadas de campo, e os modelos empíricos adaptativos, que buscam modelar a produtividade a partir de dados que possam apresentar relações com a produtividade. Neste tópico é apresentado o estado da arte sobre os estudos em predição da produtividade da cana-de-açúcar considerando estas duas abordagens.

2.2.2.1. Modelos de simulação baseados em processos fisiológicos

Existem dois principais modelos de simulação baseados em processos fisiológicos para a predição de produtividade da cana-de-açúcar em uso ao redor do mundo: APSIM (Keating *et al.*, 1999) e CANEGRO (Inman-Bamber, 1991; Jones e

Singels, 2008). Trata-se de modelos dinâmicos complexos, que apresentam embutido um grande número de outros modelos de regressões que simulam os processos fisiológicos da planta, envolvendo diversas variáveis e parametrizações. Tais modelos possibilitam simular qual seria a produção de biomassa, de sacarose, uso de água, captação de nitrogênio, entre outras. A simulação do desenvolvimento é feita para uma planta e então extrapolada para toda a área, permitindo predições em escala local, ou seja, no âmbito da propriedade agrícola.

Lisson *et al.* (2005) apontam algumas lacunas sobre o uso do conhecimento fisiológico nos modelos de simulação, especificamente para os modelos CANEGRO E APSIM. Uma das lacunas seria sobre a magnitude da redução no acúmulo de biomassa da cana-de-açúcar que ocorre ao longo do ciclo de desenvolvimento, e que pode estar associada à senescência, florescimento, tombamento, redução na fotossíntese ou aumento na taxa de respiração. Outra lacuna está relacionada ao stress hídrico, embora este seja um dos aspectos mais estudados na cana-de-açúcar. Park *et al.* (2005b) relatam que procedimentos de calibração-verificação são os principais fatores limitantes ao amplo uso destes modelos. São necessários cuidadosos procedimentos de calibração e verificação para outras variedades genéticas e diferentes condições climáticas. Marin *et al.* (2011) e Nassif *et al.* (2012) parametrizaram e avaliaram o uso do DSSAT/CANEGRO para sete variedades genéticas brasileiras. Verificou-se nestes dois estudos que ainda não é possível aplicar este modelo no Brasil, sendo ainda necessárias adaptações para as condições brasileiras. Vale lembrar que estes modelos não fazem uso de dados espectrais remotos, como imagens de satélite, que podem fornecer informações importantes sobre o desenvolvimento da cultura em toda área plantada.

2.2.2.2. Sensoriamento remoto na predição de produtividade

Sensoriamento remoto é a obtenção de informações sem o contato direto com o objeto de estudo, como nas imagens de satélites. A cultura da cana-de-açúcar possui características favoráveis ao uso de sensoriamento remoto, pois se trata de uma cultura que ocupa extensas áreas e que apresenta um dossel uniforme. Dados espectrais[11] provenientes de imagens de satélites vêm sendo cada vez mais usados em estudos com a cana-de-açúcar em diferentes aplicações, como mapeamento de áreas plantadas (Xavier

[11] Dados obtidos por meio da interação de objetos com a radiação eletromagnética do Sol.

et al., 2006; Rudorff *et al*., 2010; Vicente *et al*., 2012), diferenciação de variedades genéticas (Galvão *et al*., 2006; Fortes e Demattê, 2006), caracterização de práticas de colheita (Mello *et al*., 2013) e predição de produtividade (Simões *et al.*, 2005b; Almeida *et al*., 2006; Fortes *et al.*, 2009; Simões *et al*., 2009; Fernandes *et al*., 2011; Duveiller *et al*., 2013). A partir dos resultados destes estudos verifica-se haver um potencial uso de dados espectrais em modelos de predição de produtividade. A partir de índices de vegetação obtidos em uma série temporal de imagens de satélite, Ramme *et al.* (2010) e Fernandes *et al.* (2011) obtiveram uma curva que possibilitou inferir sobre a dinâmica da cultura no campo e o desenvolvimento da cana-de-açúcar ao longo da safra.

O sensor *MODerate-resolution Imaging Spectroradiometer* (MODIS), a bordo das plataformas Terra (lançada em 1999) e Aqua (lançada em 2002) possui grande potencial de monitoramento da cana-de-açúcar, além de disponibilidade gratuita. As imagens de NDVI (*Normalized Difference Vegetation Index*) do MODIS apresentam resolução espacial de 250m e o uso combinado das imagens das plataformas Terra e Aqua permitem obter uma imagem para a mesma área a cada 8 dias, o que o torna bastante adequado ao monitoramento de culturas agrícolas.

O MODIS apresenta resoluções espacial e espectral superiores ao sensor NOAA/AVHRR para aplicações na agricultura. Também apresenta resoluções espectrais e temporais superiores em relação ao SPOT e Landsat TM. Devido às bandas espectrais do MODIS serem estreitas, o impacto da absorção de vapor de água na banda do infravermelho próximo é minimizado, a banda do vermelho é mais sensível à clorofila, beneficiando a qualidade dos dados de NDVI (Ren *et al.*, 2008).

O índice de vegetação NDVI do sensor MODIS tem sido usado com sucesso para estimar a produtividade agrícola (Ren *et al.*, 2008; Mkhabela *et al*., 2011; Kogan *et al*., 2013; Mulianga *et al*., 2013; Sakamoto *et al*., 2014). Os trabalhos de Picoli *et al.* (2009) e Ramme *et al.* (2010) são exemplos de aplicação de imagens MODIS em estudos de predição de produtividade da cana-de-açúcar.

2.2.2.3. Dados meteorológicos em modelos de predição de produtividade

Scarpari e Beauclair (2004) propuseram um modelo, baseado em regressão linear, para estimativa do Açúcar Total Recuperável[12] (ATR) da cana-de-açúcar a partir de dados de armazenamento de água no solo e de temperatura mínima do ar. Foram usados dados de 74 talhões durante as safras 1999/2000, 2000/2001 e 2001/2002. O modelo foi treinado usando os dados das duas primeiras safras e avaliado com os dados da terceira safra. O R^2 para a etapa de treinamento foi de 0,86. Já para o teste o R^2 caiu para 0,24. Scarpari e Beauclair (2009) propuseram um novo modelo fisiológico para estimar a maturação da cana-de-açúcar. Por meio de regressão linear se buscou estimar o ATR da cana-de-açúcar a partir de dados de temperatura, armazenamento de água no solo e produção de fotoassimilados. Foram obtidos baixos valores de R^2 na etapa de treinamento. Os autores citam que um maior número de dados poderia incrementar o modelo.

Oliveira *et al*. (2012) usaram o método da zona agroecológica (ZAE) para estimar a produtividade da cana-de-açúcar nas regiões produtoras de Minas Gerais. Neste método a produtividade potencial de biomassa é calculada por meio de dados de temperatura, radiação solar e duração do ciclo da cultura. Em seguida a produtividade potencial é penalizada pelo stress hídrico, calculado pela razão entre evapotranspiração real (ETr) e evapotranspiração máxima da cultura (ETc), obtendo-se a produtividade da cana-de-açúcar limitada pela disponibilidade hídrica ao longo do ciclo. Durante a etapa de parametrização e calibração do método o modelo foi bem ajustado (R^2=0,89) aos valores de produtividade observados. Já na fase de teste foram usados dados de produtividade do IBGE, mudando assim as características do banco de dados e dificultando a análise dos resultados.

2.2.3. Descoberta de conhecimento na predição da produtividade agrícola

Existem diversos estudos que fazem uso de algoritmos de mineração de dados na modelagem e predição de produtividade na agricultura, como redes neurais artificiais

[12] Representa a qualidade da cana-de-açúcar, ou seja, a quantidade de açúcar na planta que será convertida em açúcar ou etanol.

(Liu *et al.*, 2005; Miao *et al.*, 2006; Savin *et al.*, 2007; Alvarez, 2009), árvore de decisão (Zhang *et al.*, 2006) e *fuzzy cognitive maps* (Papageorgiou *et al.*, 2009).

Park *et al.* (2005b) afirmam que os modelos empíricos adaptativos são uma alternativa interessante aos modelos de simulação baseados em processos fisiológicos (como o CANEGRO e APSIM citados no Tópico 2.2.2.1) e aos modelos determinísticos por meio de regressão linear, podendo desempenhar importante papel na identificação da estrutura desconhecida dos processos de desenvolvimento das culturas sob diferentes condições de manejo. Kaul *et al.* (2005) relatam que os modelos de predição baseados em redes neurais são uma alternativa mais simples aos modelos de simulação baseados em processos fisiológicos. Park *et al.* (2005b) também sugerem que os modelos baseados em processos fisiológicos e os modelos empíricos adaptativos poderiam ser usados de forma complementar.

Há poucos estudos de modelos que aplicam algoritmos de aprendizagem na predição de produtividade da cana-de-açúcar. Ferraro *et al.* (2009) analisaram os fatores que influenciam a produtividade da cana-de-açúcar aplicando classificação por meio de árvore de regressão (*Classification and Regression Trees* – CART) em estudo conduzido no norte da Argentina. Foram definidas três classes para o atributo classe "produtividade" (baixo, médio e alto). Segundo Ferraro *et al.* (2009) o método CART classificou corretamente 56% das ocorrências. Souza *et al.* (2010) analisaram os atributos do solo e a produtividade da cana-de-açúcar por meio de classificação por árvore de decisão. Fernandes *et al.* (2011) verificaram a relação entre dados espectrais, meteorológicos e a produtividade da cana-de-açúcar por meio de classificação por árvore de decisão.

Everingham *et al.* (2009) criaram um comitê para predição da produtividade da cana-de-açúcar na Austrália contendo 840 modelos, cada um consistindo em uma configuração diferente do modelo APSIM (modelo de predição baseado em processos fisiológicos, citado no Tópico 2.2.2.1). Foram avaliados três diferentes métodos de agregação das saídas dos membros do comitê: média simples, média entre os membros selecionados pelo algoritmo *forward stagewise* com pesos iguais para cada membro e média entre os membros selecionados pelo algoritmo *forward stagewise* com pesos padronizados para cada membro. O objetivo foi selecionar os modelos com melhores resultados. Os autores verificaram ganhos significativos com o uso do comitê, sendo o

comitê com número reduzido de membros e pesos ponderados o que apresentou melhor resultado, superando o melhor modelo individual.

2.3. Descoberta de conhecimento em banco de dados

Uma alternativa aos métodos baseados em processos fisiológicos para se obter um modelo de predição de produtividade é o uso de algoritmos de aprendizagem em um processo de descoberta de conhecimento. De acordo com Fayyad *et al.* (1996) a descoberta de conhecimento em banco de dados (*Knowledge Discovery in Databases* – KDD) se refere ao processo de extração de informação útil (conhecimento) em um grande volume de dados. O termo "processo" implica que KDD compreende vários passos, repetidos em múltiplas interações.

Segundo Han e Kamber (2006), o processo de descoberta de conhecimento consiste numa sequência interativa das seguintes etapas:

1. Limpeza dos dados (remoção de ruídos e dados inconsistentes);
2. Integração dos dados (onde múltiplas fontes de dados podem ser combinadas);
3. Seleção dos dados (onde dados relevantes para a tarefa de análise são recuperados da base de dados);
4. Transformação dos dados (onde os dados são transformados ou consolidados em formatos apropriados para mineração por meio de operações de agregação ou sumarização, por exemplo);
5. Mineração dos dados (etapa essencial onde métodos inteligentes são aplicados de forma a extrair padrões);
6. Avaliação de padrões (identificação de padrões verdadeiros e interessantes que representem o conhecimento, baseado em algumas medidas);
7. Apresentação do conhecimento (onde técnicas de visualização e representação de conhecimento são usadas para apresentar o conhecimento extraído).

As etapas 1, 2, 3 e 4 são chamadas de etapas de pré-processamento e as etapas 5, 6 e 7 são etapas de processamento e apresentação dos dados. Fayyad *et al.* (1996) alertam que, embora a etapa de mineração de dados tenha ganhado maior atenção no

processo de KDD, os passos adicionais, especialmente no pré-processamento, são essenciais para assegurar que conhecimentos úteis serão derivados dos dados.

Nas etapas de pré-processamento o banco de dados é preparado adequadamente para a aplicação dos algoritmos de mineração. Esta é a etapa que demanda maior dedicação de tempo, sendo de fundamental importância na qualidade dos resultados alcançados. Nesta etapa também são realizadas operações como: Tratamento de valores ausentes, registros repetidos e *outliers*; Discretizações de variáveis; Construção de novas variáveis a partir das variáveis de entrada e Padronizações de variáveis (normalização, por exemplo).

De acordo com Han e Kamber (2006), novas variáveis construídas a partir das variáveis de entrada podem ser adicionadas ao banco de dados de forma a ajudar aprimorar a acurácia e entendimento da estrutura dos dados. A construção de novas variáveis pode contribuir para o descobrimento de informações escondidas referentes aos relacionamentos entre variáveis. No âmbito da predição de safras, a construção de novas variáveis levando em consideração o conhecimento sobre aspectos fisiológicos da cana-de-açúcar é uma maneira de embutir o conhecimento fisiológico no modelo de aprendizagem, o que pode beneficiar sua acurácia. A construção de novas variáveis foi uma etapa explorada nesta tese.

As predições são realizadas por meio de classificação ou regressão. A classificação é usada para predizer valores categóricos (discretizados e não-ordenados). Regressões são usadas para predizer valores numéricos contínuos (ordenados).

2.3.1. Redes Neurais Artificiais

Dentre os algoritmos usados em mineração de dados se destacam as redes neurais artificiais (*Artificial Neural Networks* – ANN). Existem diversos modelos de redes neurais, que podem classificadas de acordo com o tipo de problema, estrutura das conexões e algoritmo de treinamento. Em relação ao tipo de problema, existem modelos preditivos ou descritivos. Quanto à estrutura das conexões, estas podem ser organizadas em camadas ou em estruturas recorrentes. Há diferentes algoritmos de ajuste de parâmetros adequados ao tipo de rede neural que se pretende usar. Na Figura 6 são

apresentados diversos modelos de redes neurais existentes, com destaque para o modelo usado neste estudo.

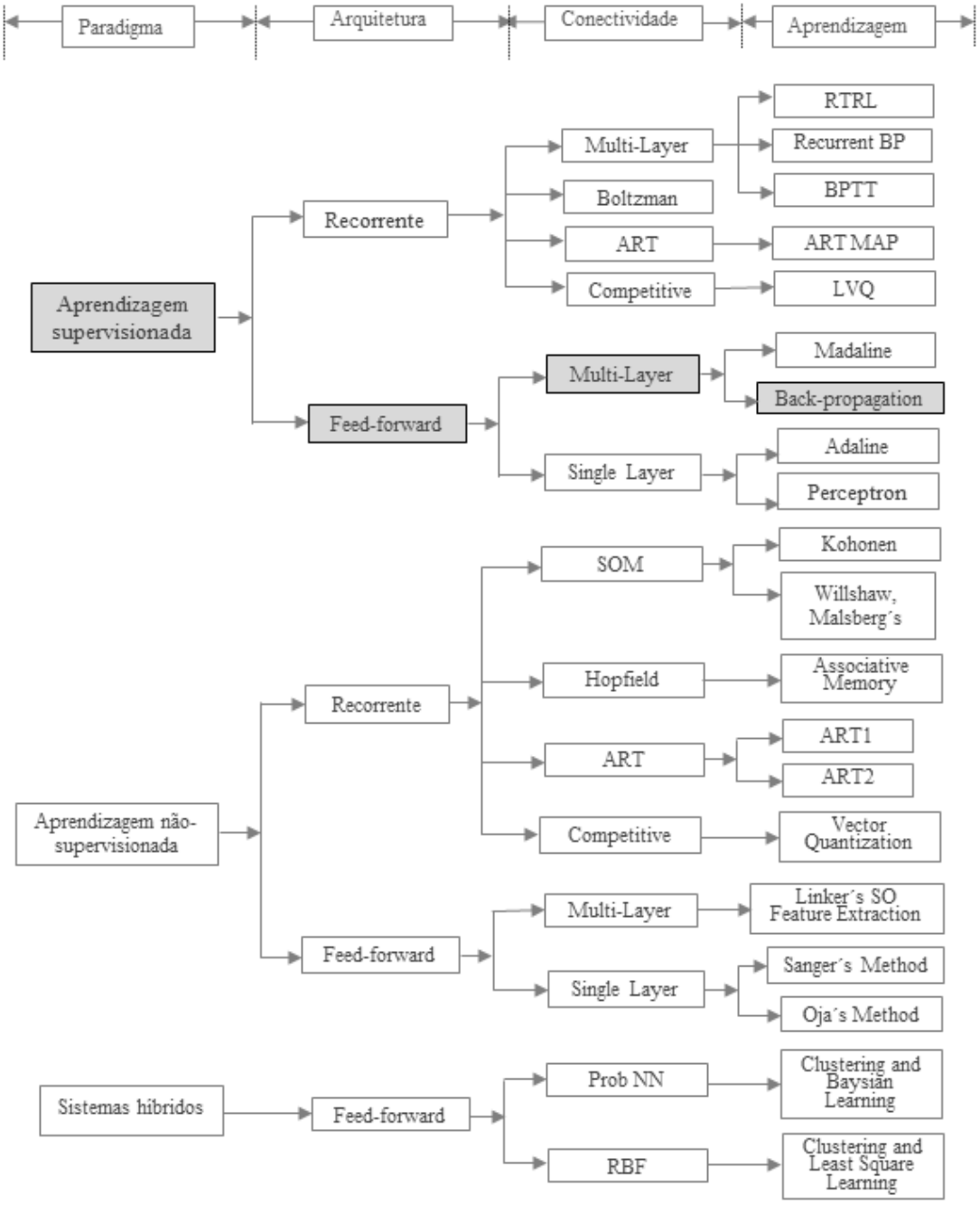


Figura 6: Modelos de redes neurais existentes, de acordo com o paradigma de aprendizagem, arquitetura, conectividade e método de aprendizagem.

O neurônio de McCulloch-Pitt é um modelo matemático do neurônio biológico proposto em 1943 por Walter Pitts e Warren McCulloch. Na Figura 7 é mostrado o neurônio artificial de McCulloch-Pitt.

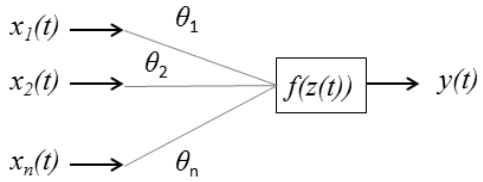


Figura 7: Neurônio artificial de McCulloch-Pitt.

A formulação matemática do neurônio é bastante simples e o valor de saída é calculado em função das entradas, conforme Equação 1:

$$y(t) = f\left(\sum_{i=1}^{N} x_i(t)\theta_i\right) \tag{1}$$

Em que:

$y(t)$: variável de saída
f: função de ativação
x_i: variáveis de entrada
θ_i: parâmetros de entrada (pesos)

A estrutura das redes neurais artificiais é inspirada no cérebro humano, em que os neurônios são conectados por meio de sinapses. Cada neurônio é representado por uma equação (função de ativação) e as sinapses são pesos ou funções que propagam os sinais. As funções de ativação mais utilizadas são a linear, a sigmoide e a tangente hiperbólica.

Um modelo de rede neural artificial amplamente usado em mineração de dados é o *Perceptron* de Múltiplas Camadas (*Multi Layer Perceptron* – MLP). Trata-se de uma rede *feed-forward*, em que a informação se propaga para frente. A topologia da rede é definida pelo número de unidades na camada de entrada, número de unidades na camada intermediária e número de unidades na camada de saída. Na Figura 8 é ilustrada uma rede neural artificial MLP com uma camada intermediária.

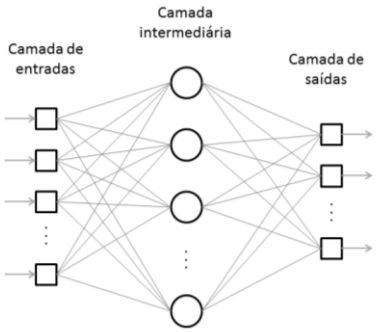


Figura 8: Ilustração de uma rede neural MLP com uma camada intermediária.

Durante a etapa de treinamento é realizada a aprendizagem da rede por meio de algoritmos, como o *Backpropagation*, e os pesos entre conexões são ajustados. O ajuste dos parâmetros é realizado pela minimização do Erro Médio Quadrático (EMQ). Ocorrida a convergência da rede, uma função de aproximação é desenvolvida e utilizada para predições futuras.

Embora as MLPs sejam poderosas ferramentas de predição, elas apresentam uma desvantagem que se refere à interpretabilidade dos resultados, o que dificulta a descrição de como as variáveis de entrada influenciaram no resultado. Outro ponto que pode representar desvantagem ao uso de redes neurais MLP se refere ao tempo necessário na etapa de treinamento da rede, onde é feito o ajuste dos pesos num processo de retropropagação do erro. Novas pesquisas buscam superar esta dificuldade, como o *Extreme Learning Machine* – EML (Huang *et al*., 2006). ELM são redes neurais alimentadas para frente (*feed-forward*) para classificação ou regressão com uma única camada escondida, onde os pesos conectando os inputs aos nós escondidos são gerados aleatoriamente e nunca atualizados. Os pesos entre a camada escondida e outputs são treinados em um único passo, como em modelos lineares. Estes modelos podem produzir bom desempenho de generalização e aprender muito mais rápido do que as redes treinadas usando retropropagação.

2.3.1.1. Parada antecipada

Um problema associado ao uso de algoritmos de aprendizagem, como redes neurais MLP, é a ocorrência de sobre-ajuste dos dados, mais comumente conhecido como *overfitting*, embora este problema também possa estar associado a outros fatores,

como a presença de variáveis irrelevantes e/ou redundantes no banco de dados de entrada. *Overfitting* é quando, após convergir durante a etapa de treinamento, minimizando o erro de predição, o método não apresenta a mesma capacidade para predizer registros que não participaram do treinamento, ou seja, não apresenta capacidade de generalização.

Uma forma de evitar a ocorrência de *overfitting* é por meio da parada antecipada. Trata-se de uma validação do método durante a etapa de treinamento, em que, a cada época, verifica-se o erro que o método apresenta ao predizer novos registros que não participaram do treinamento. O método de convergência continua enquanto houver minimização do erro na parcela de dados de validação. O método é interrompido quando não houver mais minimização do erro na parcela de validação. A partir deste ponto o método poderá continuar a minimizar o erro na parcela de treinamento, porém, com aumento do erro na parcela de validação, ou seja, ocorrência de *overfitting*. Na Figura 9 é mostrado um exemplo de treinamento com parada antecipada.

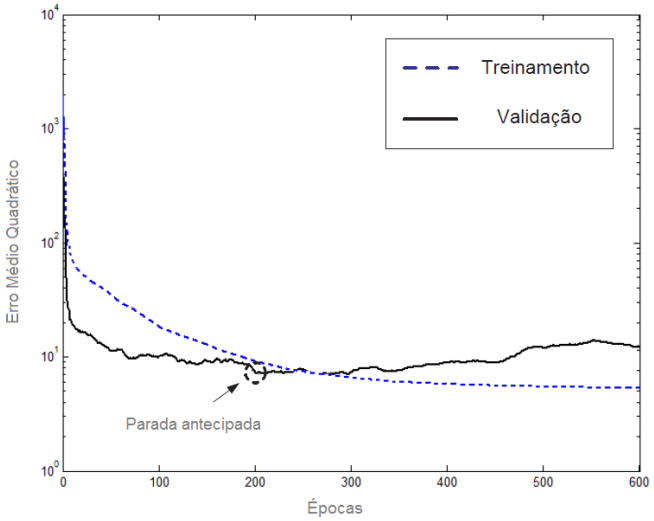


Figura 9: Exemplo de treinamento com parada antecipada.

2.3.2. Seleção de variáveis

Existe uma relação entre o número de variáveis e a capacidade de generalização. Se por um lado um maior número de variáveis implica em maior quantidade de informação, o que em tese beneficiaria o modelo de aprendizagem, por outro lado o

número limitado de registros pode não ser suficiente para um aprendizado adequado do modelo, trazendo prejuízos ao processo de aprendizagem. Esta relação entre o número de registros e o número de variáveis é chamada de dimensionalidade do banco de dados e, para um melhor desempenho do algoritmo de aprendizagem, é necessário que haja um equilíbrio na dimensionalidade.

Por meio da seleção de variáveis é possível selecionar aquelas mais relevantes para a predição do resultado final e também eliminar aquelas que não contribuem para um melhor o resultado, muitas vezes piorando o desempenho dos algoritmos de predição quando mantidas no banco de dados. Por meio da seleção de variáveis é possível avaliar também a relevância de novas variáveis construídas a partir do banco de dados inicial. A seleção de variáveis pode ser usada como forma de aumentar a diversidade e melhorar a acurácia de comitês de modelos (Kim, 2009).

Os métodos de seleção de variáveis podem ser classificados em três categorias: filtros, *wrappers* e híbridos. Os filtros realizam a seleção de variáveis baseada nas características do banco de dados, não envolvendo nenhum tipo de modelo de aprendizagem. Este método oferece uma pontuação para cada variável, ou grupo de variáveis, por meio de medidas estatísticas, como correlação, variância, entropia, teste do Chi-quadrado, entre outros. A seleção de variáveis por meio de *wrapper* requer um modelo de aprendizagem predeterminado e a seleção de variáveis tem o objetivo de melhorar o desempenho de generalização do modelo de aprendizagem em questão. Embora o método *wrapper* tenha maior custo computacional se comparado aos filtros, o desempenho de generalização é melhor (Kabir *et al.*, 2010). Os métodos híbridos buscam tirar vantagem dos métodos filtro e *wrapper*, explorando suas forças complementares. Neste estudo foi usado um método *wrapper* de seleção de variáveis.

2.3.2.1. *Wrapper*

O *wrapper* é um método que emprega o próprio algoritmo preditor como método de seleção. Para um determinado banco de dados com N variáveis, o método *wrapper* inicia o processo de seleção de variáveis a partir de uma amostra F_0, que pode ser uma amostra vazia, uma amostra contendo todas as variáveis ou algum tipo de seleção aleatória de variáveis, e percorre o espaço de busca usando uma estratégia particular de

busca. Posteriormente uma amostra F_i é gerada e avaliada aplicando-se a esta um modelo de aprendizagem. Se o desempenho do modelo de aprendizagem em F_i for melhor, então F_i é considerada a melhor amostra atual. O método *wrapper* então modifica F_i adicionando ou retirando variáveis em F_i e a busca segue de forma iterativa até que algum critério de parada pré-definido seja alcançado.

Existem diferentes estratégias em relação à forma com que as amostras podem ser geradas e seguir no processo de busca. A busca pode se iniciar em uma amostra vazia e sucessivamente ir adicionando variáveis (Peng *et al.*, 2003; Guan *et al.*, 2004), sendo este método denominado "seleção sequencial para frente" (*sequential forward selection*) mostrado na Figura 10.

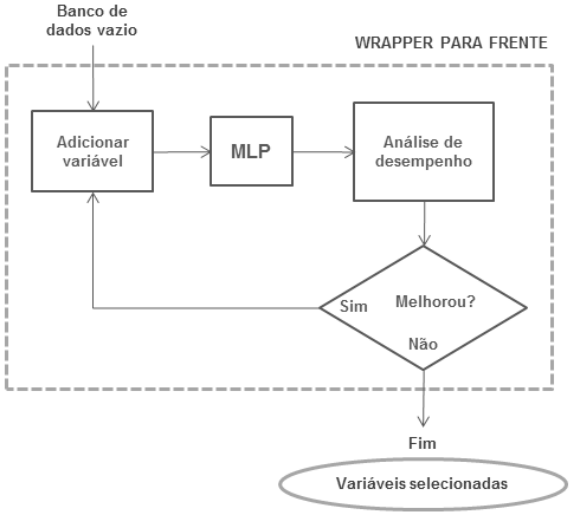


Figura 10: Procedimento de seleção de variáveis do *wrapper para frente*.

Outra maneira é iniciar o processo com uma amostra contendo todas as variáveis e então removê-las sucessivamente (Gasca *et al.*, 2006; Romero e Sopena, 2008), chamado de "seleção sequencial para trás" (*sequential backward selection*), mostrado na Figura 11.

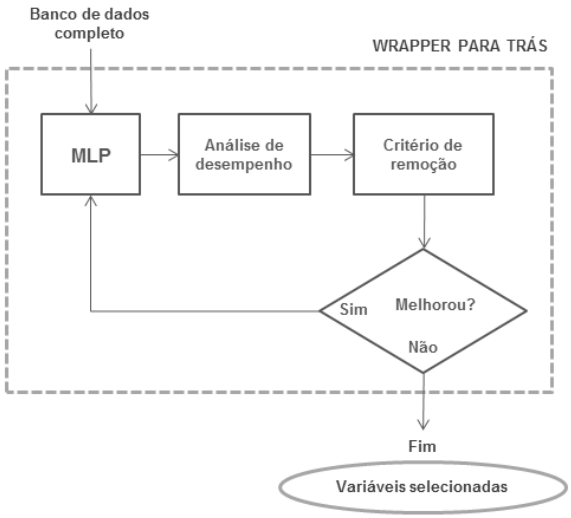


Figura 11: Processo de seleção de variáveis do *wrapper para trás*.

Estes métodos de seleção de variáveis, que envolvem uma estratégia sequencial, são simples de implementar e rápidos. Entretanto, esta estratégia pode causar o efeito de aninhamento (*nesting effect*), uma vez que se uma variável é adicionada (ou retirada) esta não poderá ser retirada (ou adicionada) posteriormente (Kabir *et al.*, 2010).

Os trabalhos de Maldonado e Weber (2009) e Kabir *et al.* (2010) são exemplos de uso de *wrapper* tendo SVM e Redes Neurais Artificiais como modelo de aprendizagem, respectivamente.

Moutinho (2011) propôs um método denominado "*wrapper* de análise de sensibilidade", onde, a cada rodada do *wrapper*, é feita uma análise de sensibilidade para verificar a importância de cada variável na determinação do resultado final, retirando assim variáveis com pouca ou nenhuma importância. Foram usados SVM e Redes Neurais como modelos de aprendizagem. Segundo Moutinho, a cada repetição da análise de sensibilidade o resultado pode ser diferente, uma vez que os valores iniciais dos algoritmos de aprendizagem são definidos aleatoriamente. Desta forma, é necessário que o processo seja repetido diversas vezes, a fim de garantir que a variável candidata a remoção não resulte em queda de desempenho do modelo.

2.3.3. Comitê de modelos preditores

Um comitê de modelos se refere ao uso de um conjunto de modelos (*Ensemble*) combinados ao invés de um único modelo. Os comitês têm o potencial de melhorar o

desempenho de um modelo preditor individual por meio da combinação das saídas de um conjunto de modelos complementares. A explicação teórica é de que diferentes modelos, gerados em condições variadas, cometem erros em diferentes pontos do banco de testes, embora mantendo resultados similares. Isso possibilitará ao comitê uma maior diversidade se comparado a um modelo simples, e a união dos resultados aumentará a capacidade de generalização do modelo. A diversidade faz com que se diminua o erro do modelo na etapa de teste, quando são usados dados que não participaram do treinamento. Esta propriedade é fornecida pelo chamado dilema "viés-variância", baseado em uma decomposição adequada do erro de predição.

Os métodos por comitês abrangem duas fases principais. A primeira se refere à produção de diferentes modelos. Os comitês podem ser compostos de modelos homogêneos ou heterogêneos. Os modelos homogêneos são aqueles que fazem uso de um mesmo algoritmo de aprendizagem. Já em modelos heterogêneos os modelos fazem uso de diferentes algoritmos. A segunda fase se trata da combinação das saídas de seus membros, que pode ser feita de diferentes maneiras, como por exemplo, média simples, média ponderada, votação ou *stacking*. *Stacking* é um método onde se aplica um algoritmo de aprendizagem para combinar as saídas dos membros do comitê.

Alguns requisitos são essenciais para o bom funcionamento de um comitê. Primeiramente, os membros individuais do comitê devem ter dados suficientes para o treinamento adequado. Em segundo lugar, os membros do comitê devem ser diversos ou complementares, isto é, cada modelo deve apresentar diferentes propriedades de predição. Isso quer dizer que cada membro do comitê deve cometer erros em diferentes partes do espaço de busca. Brown *et al.* (2005) relatam que apenas a diversidade não é suficiente, é necessário haver um equilíbrio entre diversidade e acurácia individual, de forma a obter o menor erro global. Em terceiro lugar, deve-se ter uma boa estratégia de união das diferentes saídas de cada membro, de forma a obter melhor desempenho.

A diversidade em um comitê pode ser gerada de diferentes formas, como através do treinamento de cada modelo por meio de diferentes amostragens de treinamento, diferentes estruturas internas do algoritmo de aprendizagem ou o uso de diferentes algoritmos de aprendizagem nos membros que compõem o comitê. Alguns métodos buscam produzir diversidade fornecendo diferentes amostras do banco de dados para o treinamento de cada membro do comitê. Este é provavelmente o método mais

investigado de se treinar comitês. É possível criar amostras pela divisão dos registros do banco de dados (*pattern set*), mantendo todas as variáveis presentes em cada amostra, como também criar amostras pela divisão das variáveis (*feature set*), mantendo todos os registros em cada amostra. Em ambos pode haver sobreposição de dados, ou seja, dados contidos em uma amostra também podem estar contidos em outra amostra. A divisão pelos registros (*pattern set*) é a forma mais usada para se gerar diversidade no comitê, e *Bagging* é um dos métodos mais populares.

2.3.3.1. *Bagging*

Bagging (Breiman, 1996), acrônimo derivado de *bootstrap aggregation*, é um método de se criar comitês onde a diversidade é obtida implicitamente. Trata-se de um método de predição que busca diminuir a variância. Em *Bagging* cada membro do comitê é construído a partir de uma amostragem diferente. Os membros do comitê, que fazem uso de um mesmo algoritmo de aprendizagem, trabalham de forma paralela e suas saídas (predições individuais) são combinadas por meio de uma média aritmética. Na Figura 12 é apresentado o funcionamento de um comitê de modelos criados usando *Bagging*. Estas amostras são geradas por meio de uma redistribuição aleatória com reposição (*random with replacement*) do banco de dados de treinamento. Para cada amostra, registros podem estar repetidos, enquanto outros podem ser excluídos.

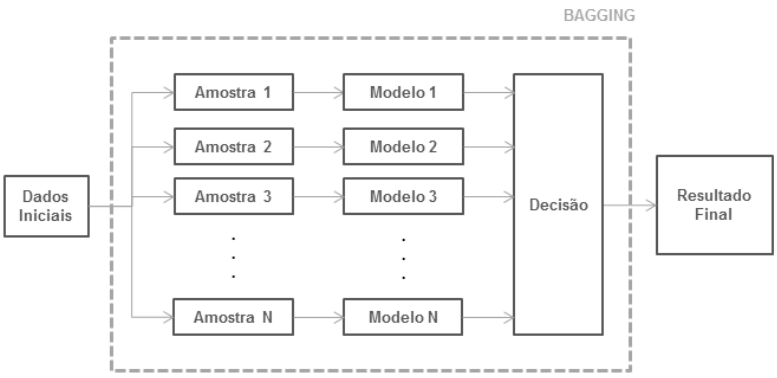


Figura 12: Criação de um comitê de modelos por meio de *Bagging*.

Bagging é um algoritmo de aprendizado bastante robusto, que pode apresentar bom desempenho mesmo em bancos de dados que apresentem ruídos (Quinlan, 1996;

Dietterich, 2000). De modo geral, o acréscimo de membros no comitê não causa *overfitting* (Breiman, 2001).

Outra observação importante do método *Bagging* é quanto ao tamanho do banco de dados. Se o banco de dados for muito pequeno, os ganhos do método por comitê podem não compensar as perdas em acurácia de cada membro. Já em grandes bancos de dados, um modelo simples pode ser adequado (Kotsiantis e Pintelas, 2004).

A probabilidade de sucesso de comitês treinados por meio de diferentes amostragens de registros, como o *Bagging*, está diretamente relacionada com a instabilidade do algoritmo de aprendizagem (Breiman, 1996). Neste contexto se destacam os algoritmos árvores de regressão e redes neurais artificiais como sendo mais adequados (Kim 2009). Os resultados de King *et al.* (2013) confirmam esta melhor adequabilidade do uso de algoritmos como redes neurais e árvores de regressão em comitês de modelos.

2.3.3.2. *Stacked Generalization*

Stacking (Wolpert, 1992) é outra forma de prover diversidade ao comitê. Trata-se de um método onde se aplica um algoritmo de aprendizagem para combinar as saídas dos membros do comitê, ao invés de calculá-la por meio de média simples ou ponderada. Dessa forma, as predições dos membros do comitê (fase 1) serão registros em um novo espaço de variáveis (Wolpert, 1992). E estes registros serão usados para treinar um novo algoritmo preditor (fase 2) que fornecerá o resultado final do comitê. Na Figura 13 está ilustrado o procedimento de criação de um comitê de modelos por meio de *Stacking*.

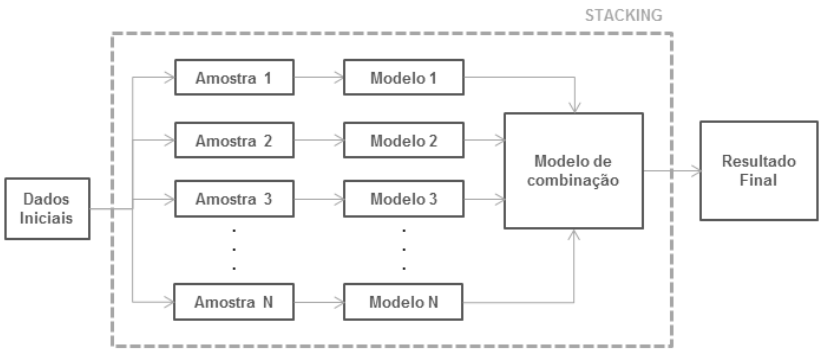


Figura 13: Criação de um comitê de modelos por meio de *Stacking*.

Espinosa *et al.* (2006) propuseram um método de combinação de resultados baseado na idéia de *Stacked Generalization*, treinando uma rede neural para combinar os vetores de saída gerados por um comitê de redes neurais. Foram avaliados comitês com 3, 9, 20 e 40 membros e oito métodos de combinação dos vetores de saída do comitê: média simples, votação, *winner takes all*, *borda count*, combinação bayesiana, *dinamically averaged networks*, *stacked generalization* e *stacked generalization plus*. Os métodos foram avaliados em seis bancos de dados do UCI, repositório de dados para uso em aprendizado de máquina. Os autores concluíram que o ganho de desempenho dos dois métodos de *Stacking Generalization* depende do banco de dados e do número de membros no comitê usados, mas que, de forma geral, os métodos de *stacking* foram superiores ao método por média simples.

King *et al.* (2013) avaliaram o uso de comitês em modelo de predição de dias favoráveis à prática de esqui. Foram avaliados três modelos preditores (regressão linear múltipla, árvore de regressão e redes neurais) e quatro métodos de comitês (*Bagging*, *Random Subspace*, *Stacking* e Votação), obtendo dez diferentes formas de comitês: três *Bagging* (um para cada algoritmo preditor); três *Randon Subspace* (um para cada algoritmo preditor); três *Stacking*, (considerando na primeira etapa os três algoritmos preditores de forma conjunta e na segunda fase, de união dos resultados, um dos algoritmos preditores como agregador); e um Votação. Todos os comitês apresentaram resultados superiores aos modelos simples. Porém, o *Stacking*, fazendo uso dos três algoritmos de predição como membros do comitê (heterogêneo), e Redes Neurais como algoritmo agregador, foi o que apresentou melhor resultado.

2.4. O futuro da previsão de safras para a cana-de-açúcar no Brasil

A demanda por fontes de energia renováveis continuará a aumentar, mantendo a tendência de aumento de produção de cana-de-açúcar verificada no Brasil (Figura 14).

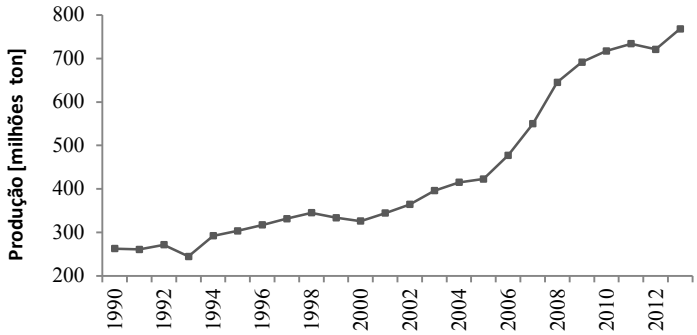


Figura 14: Série histórica da produção brasileira de cana-de-açúcar, entre os anos de 1990 e 2013. Fonte SIDRA (2015).

Os sistemas produtivos serão constantemente otimizados, necessitando assim de informações sobre o desenvolvimento da cultura no campo de forma cada vez mais rápida. Dessa forma, haverá uma permanente necessidade por modelos de previsão de safras que forneçam resultados cada vez mais rápidos e acurados.

Além do monitoramento da produção, modelos de predição de produtividade serão cada vez mais usados para verificar o impacto das mudanças climáticas nos sistemas produtivos, auxiliando também nas pesquisas de novas variedades genéticas adaptadas.

A geração de dados para o monitoramento de áreas plantadas - como dados provenientes de satélites, radares, veículos aéreos não tripulados (VANTs ou DRONES), sensores instalados nas áreas cultivadas, dados meteorológicos simulados por modelos - como o *European Centre for Medium-Range Weather Forecasts* (ECMWF, 2012) - serão constantemente aprimorados, aumentando a quantidade de dados possíveis de serem analisados e incorporados em modelos de predição de produtividade, bem como a rapidez na disponibilização.

Modelos acurados e confiáveis de previsão de safras brasileiras serão cada vez mais importantes para assegurar confiança às estatísticas sobre a produção brasileira em nível internacional, evitando especulações provenientes de países concorrentes. Outros países já monitoram a produção agrícola do Brasil. O *United States Department of Agriculture* (USDA) monitora e estima a produção nos principais países produtores das commodities agrícolas através do *Foreign Agricultural Service* (FAS) por meio de dados obtidos em levantamentos de especialistas alocados no Brasil, imagens de satélite e dados meteorológicos (USDA, 2011). A União Europeia – por meio do MARS

(*Monitoring Agriculture with Remote Sensing*) e do MCYFS (*Mars Crop Yield Forecasting System*), que monitoram o comportamento da cultura no campo e produzem previsões de produtividade da cultura - está experimentalmente estendendo sua área de abrangência para outras quatro áreas, entre elas a América do Sul (METAMP, 2011). Seguramente estes países continuarão a aperfeiçoar seus modelos de predição sobre a produção brasileira.

Todos estes aspectos aumentam a necessidade de se pensar em modelos dinâmicos de predição de produtividade, que possibilitem a avaliação sobre a relevância de novas variáveis bem como o uso de um volume de dados cada vez maior.

Capítulo 3 – Material e Métodos

A execução deste trabalho está dividida em duas etapas: Preparação dos dados (tópico 3.1 desta tese) e Processamento dos dados (tópico 3.2).

Na etapa de preparação foram definidas as áreas e período de estudo, obtidos os dados espectrais, meteorológicos e de produtividade para a área e período selecionados e construídas as novas variáveis a partir dos dados iniciais. No processamento foram feitas a seleção de variáveis relevantes e a predição da produtividade por meio de comitês de modelos. Este procedimento foi realizado para quatro abordagens diferentes: usando somente variáveis espectrais ao longo de toda a safra (modelo espectral municipal), usando variáveis espectrais e meteorológicas ao longo de toda a safra (modelo espectral-meteorológico municipal), usando variáveis espectrais e meteorológicas referentes aos dois terços iniciais da safra (modelo reduzido municipal) e calculando a produtividade para uma escala estadual (modelo estadual). Na Figura 15 está apresentado um fluxograma geral sobre as etapas realizadas.

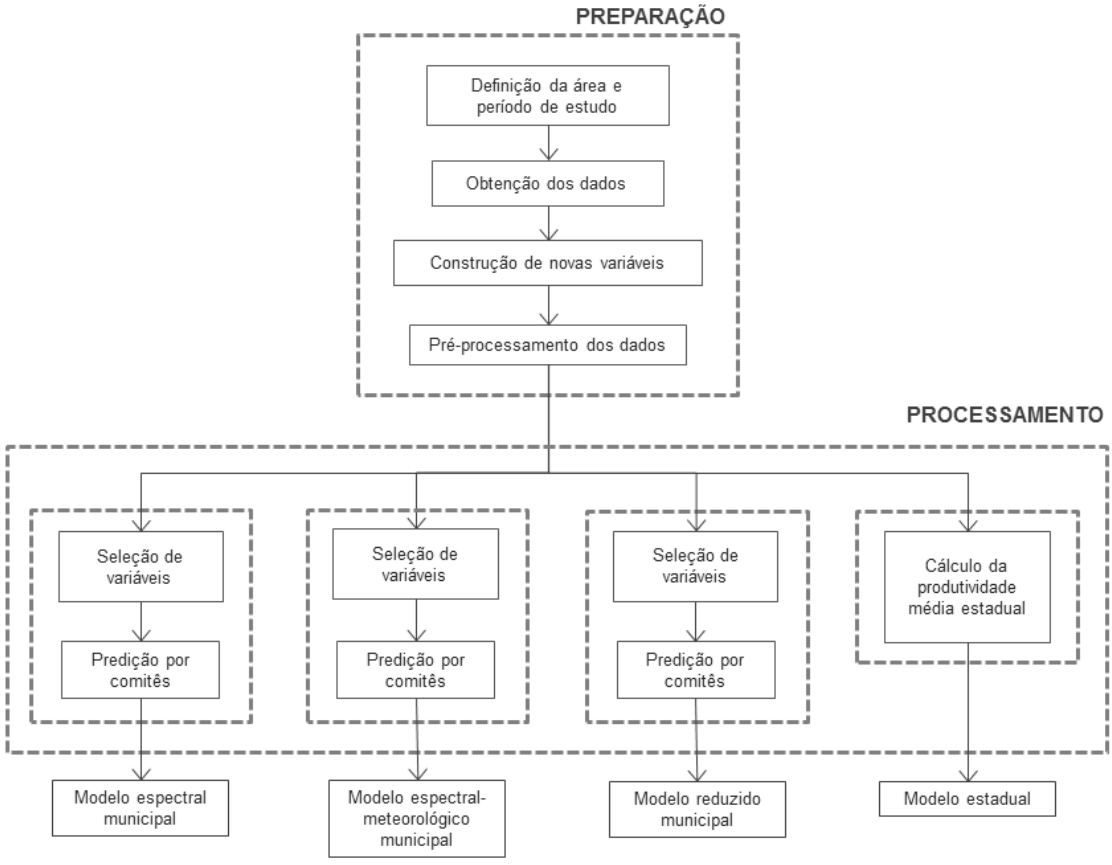


Figura 15: Fluxograma geral das etapas realizadas.

3.1. Preparação dos dados

3.1.1. Definição da área e período de estudo

O estado de São Paulo foi escolhido como área de estudo devido à sua importância produtiva. Os municípios de maior produção do Estado apresentam uma grande proporção de área ocupada com cana-de-açúcar e extensas áreas contínuas de plantio. Isso beneficia a seleção de *pixels* puros nas imagens de satélite, ou seja, *pixels* que contenham somente área plantada com cana-de-açúcar, propiciando assim séries temporais do NDVI representativas do comportamento da cultura no campo. Dessa forma, para a seleção dos municípios, foi considerada a distribuição espacial das áreas plantadas com cana-de-açúcar em cada município e selecionados aqueles que, além da importância produtiva, possibilitaram a seleção de uma amostra representativa de *pixels* em áreas canavieiras. Os dados sobre produção foram obtidos em SIDRA (2015).

Na Figura 16 é apresentado um mapa da área plantada com cana-de-açúcar na safra 2012-2013 no estado de São Paulo, elaborado a partir de mapeamento via imagens de satélite realizado pelo projeto Canasat (Rudorff *et al*., 2010).

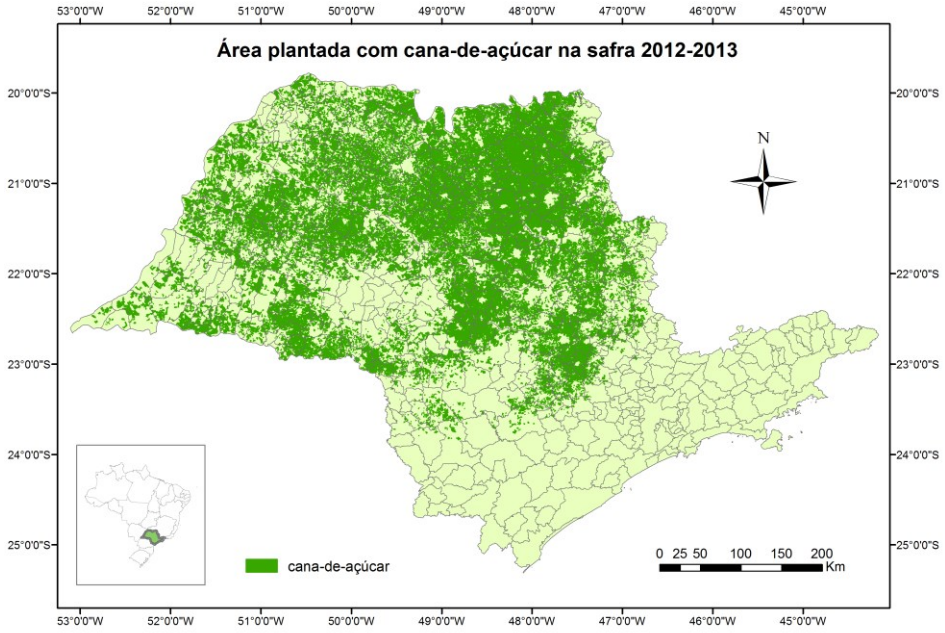


Figura 16: Mapeamento da área plantada com cana-de-açúcar na safra 2012-2013 no estado de São Paulo.

Para a escolha dos municípios, além da distribuição espacial da área plantada e do volume de produção de cada município, foi considerada também a classificação climática do estado de São Paulo, de forma a selecionar municípios que contemplassem diferentes classes de clima. De acordo com a classificação climática de Koeppen, o estado de São Paulo abrange sete tipos climáticos distintos, conforme mostrado na Figura 17. As áreas canavieiras do Estado contemplam basicamente duas classes climáticas: o **Cwa**, que abrange toda a parte central do Estado e é caracterizado pelo clima tropical de altitude, com chuvas no verão e seca no inverno, com a temperatura média do mês mais quente superior a 22°C; e o **Aw** nas regiões a Noroeste, mais quentes, com clima tropical chuvoso com inverno seco e mês mais frio com temperatura média superior a 18ºC. O mês mais seco tem precipitação inferior a 60mm e com período chuvoso que se atrasa para o outono. Em alguns poucos municípios canavieiros também ocorre o clima **Am**, que caracteriza o clima tropical chuvoso, com inverno seco onde o mês menos chuvoso tem precipitação inferior a 60mm. O mês mais frio tem temperatura média superior a 18°C.

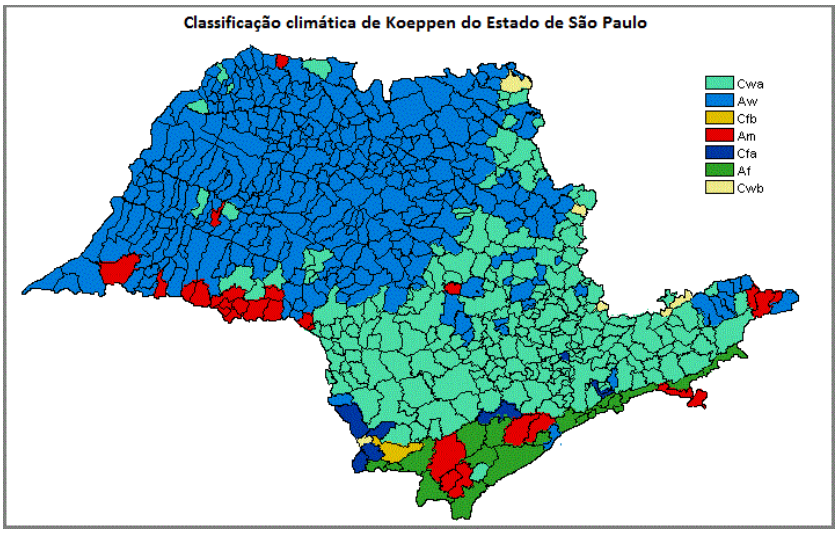


Figura 17: Classificação climática de Koeppen do estado de São Paulo. Fonte: Adaptado de CEPAGRI (2014).

Dessa forma, buscando considerar a importância produtiva e a homogeneidade de área plantada, entre 2003 e 2012, como também a diferenciação climática, foram selecionados os 60 municípios apresentados na Figura 18.

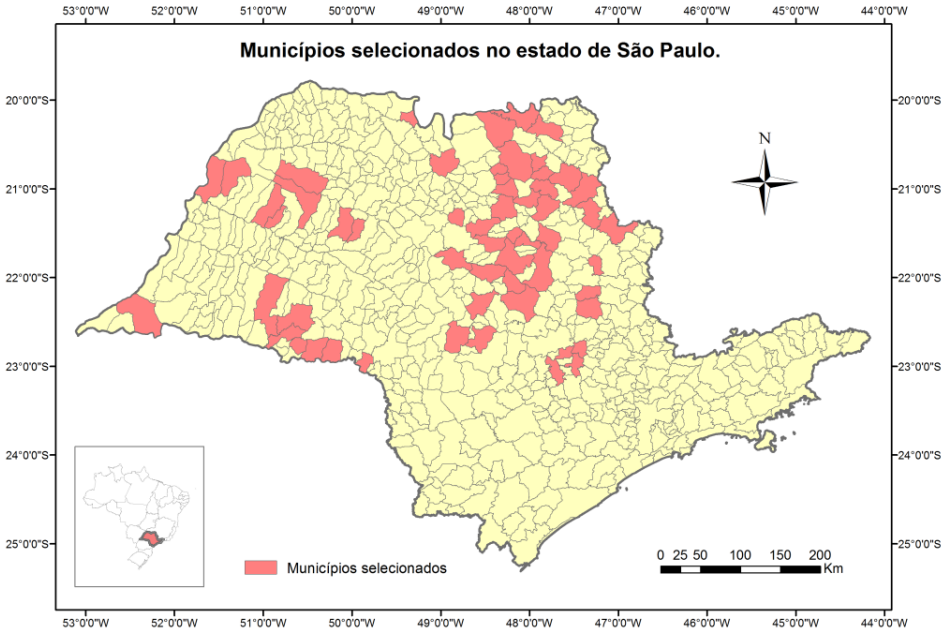


Figura 18: Mapa de localização dos 60 municípios selecionados no estado de São Paulo.

Em discussões preliminares sobre predição de produtividade de safras, Jame e Cutforth (1996) argumentaram sobre a importância de se ter mais de 10 anos de dados consecutivos para predições baseadas em modelos empíricos. Considerando isso e o período de disponibilidade dos dados, definiu-se o período de estudo entre 2003 e 2012, totalizando 10 anos consecutivos de dados. Com isso foi possível obter nove safras, uma vez que a safra para a cana de ano na Região Centro-Sul do Brasil se inicia aproximadamente entre setembro-novembro de um ano e se estende até aproximadamente novembro do ano seguinte.

Assim, a partir dos 60 municípios e nove safras consecutivas, foi possível obter um banco de dados contendo 540 registros.

3.1.2. Obtenção dos dados e construção de variáveis

3.1.2.1. Dados de produtividade de biomassa

Os dados municipais de produtividade de biomassa de cana-de-açúcar foram obtidos na base de dados do IBGE, disponível em SIDRA (2015). O IBGE denomina

esta variável de rendimento médio da produção, em quilogramas por hectare [kg ha^{-1}]. Foram obtidos os dados de rendimento médio da produção para os 60 municípios selecionados e entre as safras de 2003 a 2012.

3.1.2.2. Dados espectrais

A partir de imagens de satélite é possível obter o Índice de Vegetação por Diferença Normalizada (*Normalized Difference Vegetation Index* – NDVI), proposto por Rouse *et al*. (1973). O NDVI é obtido de acordo com a equação (2).

$$NDVI = \frac{(IVP - V)}{(IVP + V)} \qquad (2)$$

Em que:

IVP = refletância no comprimento de onda do infravermelho próximo;

V = refletância no comprimento de onda do vermelho.

Este índice varia entre zero (-1) e um (1). Valores baixos de NDVI indicam pouca biomassa e vice-versa. No início da safra as áreas canavieiras apresentam NDVI baixo e, conforme a cultura vai se desenvolvendo, o índice aumenta (Figuras 19 e 20).

Foram usados dados de NDVI extraídos de imagens do sensor MODIS, a bordo dos satélites Terra e Aqua, disponíveis em LP-DAAC (2015). Cada satélite permite a obtenção de uma nova imagem, para um mesmo alvo na Terra, a cada 16 dias. O uso combinado das plataformas Terra e Aqua permitiu obter imagens de NDVI com periodicidade de oito dias, para cada município e safra em estudo. Foi aplicado um primeiro método de suavização nesta série temporal de NDVI, baseado no filtro conservativo proposto por Wardlow *et al*. (2006), com o objetivo de reduzir ruídos causados pela presença de nuvens.

O seguinte procedimento foi realizado para a obtenção da série temporal do NDVI médio para cada município e safra:

a) Download da série de imagens de NDVI, para as plataformas Terra (MOD13Q1) e Aqua (MYD13Q1), de 2003 a 2012, totalizando aproximadamente 460 imagens (46 imagens por ano);

b) Seleção dos pixels puros localizados em áreas canavieiras, para cada município e safra, fornecidos por mapeamento realizado pelo projeto Canasat (Rudorff *et al*., 2010);

c) Extração do valor de NDVI para cada pixel selecionado no item "b", em cada uma das 460 imagens;

d) Cálculo do valor médio de NDVI entre todos os pixels de cada município, em cada uma das 460 imagens.

Os itens c e d do procedimento citado foram realizados por meio de uma rotina automática desenvolvida em linguagem IDL. Com isso, foi possível obter a série temporal do NDVI médio municipal para cada safra. Esta curva corresponde ao comportamento espectral médio das áreas canavieiras de cada município, que representa o ciclo de desenvolvimento da planta no campo ao longo da safra, bem como seu vigor vegetativo. Neste estudo foram selecionados pixels referentes apenas a áreas com cana de ano, em que o plantio / rebrota ocorre aproximadamente em setembro / novembro e tem duração de até um ano. Na Figura 19 são mostrados três exemplos de séries temporais de NDVI, para o município de Altinópolis, obtidas por meio do procedimento citado. O eixo *x* do gráfico da Figura 19 se inicia em junho.

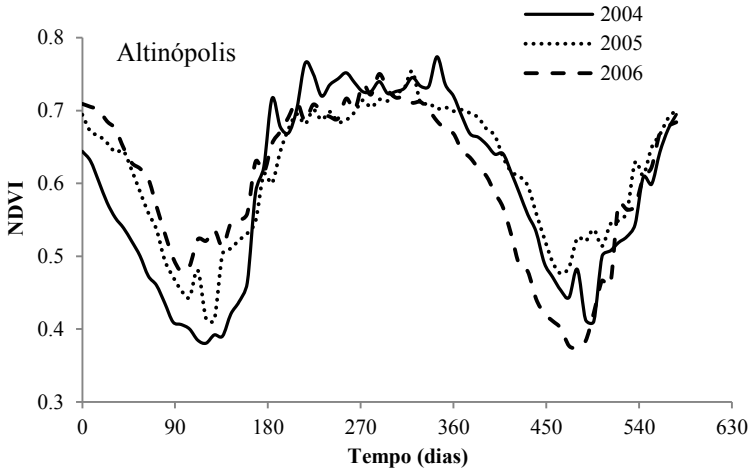


Figura 19: Exemplo de séries temporais de NDVI médio obtidas para o município de Altinópolis, para as safras 2004, 2005 e 2006.

O software TIMESAT (Jönsson e Eklundh, 2004) foi usado para identificar a data de início e final de cada safra e também extrair algumas métricas da série temporal do NDVI. O método de suavização Savitzy-Golay foi aplicado à série temporal para eliminar variações abruptas possivelmente causados por ruídos e que poderiam dificultar a extração das métricas. Assim, a partir do software TIMESAT, foram geradas oito variáveis espectrais para cada safra e município: data de início da safra, data de final da safra, duração da safra, valor máximo de NDVI, amplitude da série temporal, taxa de crescimento (derivada do lado esquerdo), taxa de senescência (derivada do lado direito) e integral da série temporal. Na Figura 20 são apresentados alguns parâmetros extraídos da série temporal do NDVI pelo TIMESAT. No Anexo 1 é apresentada a interface gráfica do TIMESAT com o exemplo de uma série temporal de NDVI para um município e ao longo de 10 safras, sem a aplicação do filtro (em azul) e após a aplicação do filtro Savitzy-Golay (em amarelo). Os pontos em amarelo se referem ao início e final de cada safra.

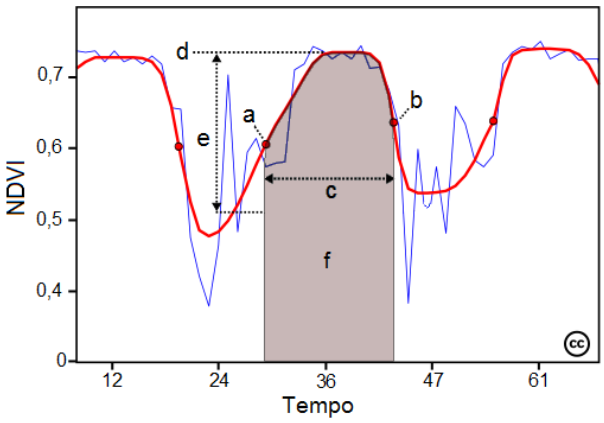


Figura 20: Alguns dos parâmetros gerados pelo TIMESAT: a) início da safra, b) final da safra, c) duração da safra, d) valor máximo, e) amplitude e f) integral. Fonte: Adaptado de Jönsson e Eklundh (2004).

Além das variáveis obtidas pelo software TIMESAT, valores de NDVI também foram extraídos da série temporal para 11 períodos específicos: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10 meses após o início da safra. Foram obtidas assim 19 variáveis espectrais derivadas da série temporal de NDVI. O nome de cada variável espectral e sua descrição estão detalhados na Tabela 1.

Verificam-se na Figura 19 diferentes comportamentos da série temporal de NDVI para cada uma das três safras exemplificadas. O objetivo das variáveis

apresentadas na Tabela 1 é "medir" estas diferenças, a fim de correlacioná-las com a produtividade média municipal da cana-de-açúcar.

Tabela 1: Variáveis espectrais geradas a partir da série temporal municipal de NDVI.

nº	Nome da variável	Descrição
1	*sos*	Data de início da safra
2	*eos*	Data do final da safra
3	*lenght*	Duração da safra
4	*max_value*	Valor máximo do NDVI ao longo da safra
5	*amplitude*	Amplitude da série temporal de NDVI
6	*left_derivative*	Taxa de crescimento da série temporal de NDVI
7	*right_derivative*	Taxa de senescência da série temporal de NDVI
8	*integral*	Integral da série temporal de NDVI
9	*ndvi0*	Valor do NDVI no início da safra (sos)
10	*ndvi1*	Valor do NDVI após um mês do início da safra
11	*ndvi2*	Valor do NDVI após dois meses do início da safra
12	*ndvi3*	Valor do NDVI após três meses do início da safra
13	*ndvi4*	Valor do NDVI após quatro meses do início da safra
14	*ndvi5*	Valor do NDVI após cinco meses do início da safra
15	*ndvi6*	Valor do NDVI após seis meses do início da safra
16	*ndvi7*	Valor do NDVI após sete meses do início da safra
17	*ndvi8*	Valor do NDVI após oito meses do início da safra
18	*ndvi9*	Valor do NDVI após nove meses do início da safra
19	*ndvi10*	Valor do NDVI após 10 meses do início da safra

3.1.2.3. Dados meteorológicos

No Brasil existem dificuldades de utilização de dados meteorológicos provenientes de estações meteorológicas para o monitoramento de culturas, como, por exemplo, o atraso existente entre o momento de coleta e de divulgação do dado e a desuniformidade na densidade de estações distribuídas pelo território brasileiro, além de um grande número de dados faltantes nas séries de dados disponíveis. Uma alternativa aos dados de estações meteorológicas é o uso de dados simulados por modelos matemáticos, que calculam as variáveis meteorológicas a partir de dados espectrais provenientes de imagens de satélites. O *Joint Research Centre* (JRC) disponibiliza gratuitamente um banco de dados meteorológicos simulados, oriundos do modelo do

ECMWF (*European Center for Medium-Range Weather Forecast*). A utilização dos dados simulados pelo ECMWF apresenta grandes vantagens em relação aos dados provenientes de estações meteorológicas, como rapidez na obtenção dos dados (em tempo quase real) e cobertura para toda a superfície terrestre.

Moraes *et al*. (2012) verificaram a precisão e acurácia do modelo ECMWF pela comparação de dados decendiais simulados de precipitação, temperatura máxima e mínima do ar aos dados observados por mapas interpolados de estações meteorológicas do estado de São Paulo no período entre 2005 e 2010. Como resultado, observou-se que o modelo ECMWF simula satisfatoriamente os elementos meteorológicos para o estado de São Paulo.

Os dados são disponibilizados em formato *shapefile* (.shp) no *website* do JRC (2014), e apresentam resolução espacial de aproximadamente 25 km e resolução temporal decendial (intervalo de 10 dias). São disponibilizados os seguintes dados: temperatura máxima, temperatura média, temperatura mínima, radiação global acumulada, precipitação acumulada, evapotranspiração acumulada (Penman-Monteith), balanço hídrico climático. Porém, após uma análise prévia, verificou-se que algumas variáveis eram altamente correlacionadas, devido à relação natural existente entre as variáveis, como:

- temperatura máxima ***X*** temperatura média;

- temperatura máxima ***X*** temperatura mínima;

- temperatura média ***X*** temperatura mínima;

- radiação global acumulada ***X*** evapotranspiração acumulada;

- precipitação acumulada ***X*** evapotranspiração acumulada;

- precipitação acumulada ***X*** balanço hídrico climático;

- evapotranspiração acumulada ***X*** balanço hídrico climático;

Dessa forma, considerando as correlações existentes e os fatores climáticos mais críticos para a produção da cana-de-açúcar no estado de São Paulo, optou-se por trabalhar com as seguintes variáveis: temperatura mínima (*TMIN*) em [°C], radiação global acumulada (*RAD*) em [kJ/m^2] e precipitação acumulada (*RRR*) em [mm]. Foram então adquiridos dados entre os anos de 2003 a 2012 para estas três variáveis meteorológicas.

Para exemplificar o modelo de dados meteorológicos disponibilizados por JRC (2014), na Figura 21 é apresentado um mapa de distribuição da precipitação acumulada entre os dias 01 e 10 de janeiro de 2004 no Brasil, obtido por meio dos dados decendiais de precipitação acumulada do modelo do ECMWF.

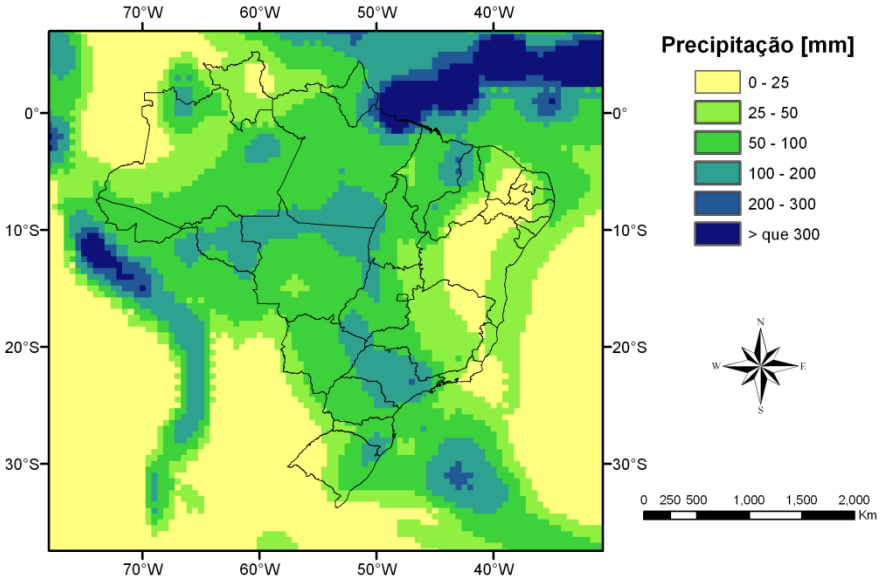


Figura 21: Mapa da distribuição da precipitação acumulada entre os dias 01 e 10 de janeiro de 2004 no Brasil proveniente de imagem decendial do modelo do ECMWF.

A partir dos dados obtidos para as três variáveis meteorológicas consideradas - temperatura mínima, radiação global acumulada e precipitação acumulada – foram construídas as variáveis meteorológicas usadas no estudo.

O ciclo fisiológico foi segmentado com o objetivo de possibilitar a criação de variáveis em diferentes etapas do desenvolvimento da cultura e assim verificar se, por exemplo, a média da precipitação ocorrida no início, no meio, ou no final da safra teve maior relação com a produtividade final. Considerando as divisões do ciclo fisiológico da cana-de-açúcar feitas por Wiedenfeld (2000), Marin e Carvalho (2012) e Rodrigues (2013), descritas no tópico 2.1.1 desta tese, optou-se por dividir o ciclo da cultura - período entre o início da safra (*sos*) e final da safra (*eos*) - em três partes. Dessa forma, foram calculados o valor máximo, mínimo e médio de cada variável e para cada terço do ciclo de desenvolvimento. Na Tabela 2 estão descritas as variáveis meteorológicas obtidas a partir dos dados do ECMWF, bem como suas descrições.

Tabela 2: Variáveis meteorológicas criadas a partir dos dados meteorológicos iniciais.

nº	Nome da variável	Descrição
1	*TMIN_min1*	Mínimo da temperatura mínima na fase 1
2	*TMIN_min2*	Mínimo da temperatura mínima na fase 2
3	*TMIN_min3*	Mínimo da temperatura mínima na fase 3
4	*TMIN_max1*	Máximo da temperatura mínima na fase 1
5	*TMIN_max2*	Máximo da temperatura mínima na fase 2
6	*TMIN_max3*	Máximo da temperatura mínima na fase 3
7	*TMIN_avg1*	Média da temperatura mínima na fase 1
8	*TMIN_avg2*	Média da temperatura mínima na fase 2
9	*TMIN_avg3*	Média da temperatura mínima na fase 3
10	*RAD_min1*	Mínimo da radiação global acumulada na fase 1
11	*RAD_min2*	Mínimo da radiação global acumulada na fase 2
12	*RAD_min3*	Mínimo da radiação global acumulada na fase 3
13	*RAD_max1*	Máximo da radiação global acumulada na fase 1
14	*RAD_max2*	Máximo da radiação global acumulada na fase 2
15	*RAD_max3*	Máximo da radiação global acumulada na fase 3
16	*RAD_avg1*	Média da radiação global acumulada na fase 1
17	*RAD_avg2*	Média da radiação global acumulada na fase 2
18	*RAD_avg3*	Média da radiação global acumulada na fase 3
19	*RRR_max1*	Máximo da precipitação acumulada na fase 1
20	*RRR_max2*	Máximo da precipitação acumulada na fase 2
21	*RRR_max3*	Máximo da precipitação acumulada na fase 3
22	*RRR_avg1*	Média da precipitação acumulada na fase 1
23	*RRR_avg2*	Média da precipitação acumulada na fase 2
24	*RRR_avg3*	Média da precipitação acumulada na fase 3

3.1.3. Pré-processamento dos dados

Nesta etapa foi feita uma análise exploratória dos dados. Em seguida os dados foram preparados para posterior seleção de variáveis e criação dos comitês de modelos preditores.

a) Caracterização e visualização dos dados

Foram calculadas as estatísticas básicas das variáveis construídas e verificadas as distribuições por meio de histogramas.

b) Verificação de outliers

Foi analisada a ocorrência de possíveis *outliers* por meio de *box-plot*. Posteriormente verificou-se o efeito nos resultados ao se retirar os candidatos a *outliers* do banco de dados.

c) Análise de correlação entre variáveis

Foi feita uma análise de correlação entre variáveis a partir de uma matriz de correlação, a fim de identificar variáveis independentes altamente correlacionadas bem como a correlação entre variáveis independentes e a variável dependente "rendimento médio" (rend).

d) Padronização dos dados

Os dados foram padronizados pelo método *min-max* (função do Matlab) com o objetivo de possibilitar o correto funcionamento das redes neurais MLPs.

3.2. Processamento dos dados

Nesta etapa foram aplicados os algoritmos de aprendizagem para selecionar as variáveis relevantes para a predição da produtividade e criar modelos de predição por meio de comitês. Os scripts necessários para as etapas de seleção de variáveis e predição por comitês, descritas nos próximos tópicos, foram desenvolvidos no *software* Matlab R2013a (8.1.0.604). Foram definidas quatro diferentes abordagens, detalhadas no tópico a seguir.

3.2.1. Quatro diferentes abordagens para a predição da produtividade

A predição da produtividade foi realizada para quatro diferentes abordagens: em escala municipal usando apenas variáveis espectrais, denominado "modelo espectral municipal", em escala municipal usando variáveis espectrais e meteorológicas, denominado "modelo espectral-meteorológico municipal", em escala municipal considerando apenas variáveis que podem ser obtidas antes do término da safra, denominado "modelo reduzido municipal" e em escala estadual considerando a média

dos resultados dos modelos municipais, denominado "modelo estadual". Estas abordagens estão detalhadas nos tópicos a seguir.

3.2.1.1. Modelo espectral municipal

Nesta primeira abordagem somente variáveis espectrais foram usadas. Aqui o objetivo foi verificar o potencial das variáveis espectrais em predizer a produtividade de biomassa da cana-de-açúcar em nível municipal. Iniciou-se com o procedimento de seleção de variáveis, descrito no tópico 3.2.2.1, considerando as 19 variáveis espectrais descritas na Tabela 1. Feita a seleção de variáveis, aplicaram-se os métodos de criação de comitês descritos no tópico 3.2.2.2.

3.2.1.2. Modelo espectral-meteorológico municipal

Nesta segunda abordagem foram usadas variáveis espectrais e meteorológicas com o objetivo de verificar se a inclusão de variáveis meteorológicas melhoraria o desempenho do modelo espectral municipal citado no tópico anterior. Assim, iniciou-se o procedimento de seleção de variáveis (tópico 3.2.2.1) considerando as 19 variáveis espectrais da Tabela 1 e as 24 variáveis meteorológicas da Tabela 2. Feita a seleção de variáveis, aplicaram-se os métodos de criação de comitês (tópico 3.2.2.2).

3.2.1.3. Modelo reduzido municipal

Nesta abordagem foram usadas apenas as variáveis que podem ser obtidas, no mínimo, três meses antes do final da safra com o objetivo de verificar a capacidade de se estimar a produtividade antecipadamente ao final da safra. Para isso, foram excluídas todas as variáveis que dependessem dos últimos três meses da safra para então iniciar os procedimentos de seleção de variáveis e predição por comitês de modelos citados nos tópicos 3.2.2.1 e 3.2.2.2.

3.2.1.4. Modelo estadual

Neste modelo buscou-se predizer a produtividade média para todo o estado de São Paulo. O objetivo foi de comparar este resultado de predição com as estimativas oficiais de produtividade fornecidas pelos Levantamentos para Acompanhamento da Safra Brasileira de Cana-de-açúcar, realizados pela Conab e disponíveis em Conab (2015), que são realizados somente em escala estadual.

Para isso, a partir dos resultados de produtividade encontrados nos modelos em escala municipal foi calculado o resultado médio de produtividade entre todos os 60 municípios em estudo, obtendo assim a produtividade média estadual por safra, para as safras entre 2004 e 2012.

A partir dos dados oficiais de produtividade municipal e estadual, disponibilizados pelo IBGE (SIDRA, 2015), os resultados dos quatro modelos propostos foram avaliados em relação à sua acurácia, por meio da Raiz do Erro Médio Quadrático (*Root Mean Square Error* – RMSE), e precisão, usando o Coeficiente de Determinação (R^2).

Também foram calculados o RMSE e o R^2 para as três estimativas que a Conab realiza ao longo da safra de cana-de-açúcar com o objetivo de comparar os resultados do modelo estadual com as estimativas estaduais da Conab.

3.2.2. Parâmetros da seleção de variáveis e dos comitês de modelos

Tanto para o procedimento de seleção de variáveis quanto para a predição por comitês foram definidas as proporções de 70% do banco de dados para a etapa de treinamento do algoritmo de aprendizagem, 10% para a validação (parada antecipada, a fim de evitar *overfitting*) e 20% para o teste, ou seja, dados que não participaram das etapas anteriores.

Redes neurais MLP foram usadas para a seleção de variáveis e para a predição por comitês de modelos. Foi usado o algoritmo *Levenberg-Marquardt backpropagation* como função de treinamento para o ajuste dos pesos na otimização da rede neural

devido ao seu melhor desempenho entre as funções de treinamento disponíveis no Matlab. As funções de ativação usadas foram a sigmoide para a camada escondida e a linear para a camada de saída.

3.2.2.1. Seleção de variáveis

A seleção de variáveis foi feita usando um *wrapper* com uma rede neural MLP. A vantagem deste método é o fato da seleção ser feita usando o mesmo algoritmo de aprendizagem da etapa seguinte, de predição da produtividade por comitês de redes MLP. Após testes preliminares entre os métodos *wrapper* com seleção sequencial para frente e seleção sequencial para trás, optou-se por usar o *wrapper* com seleção sequencial para trás, em que se inicia com o banco de dados contendo todas as variáveis e, a cada rodada, retira-se variáveis que apresentam impacto negativo no desempenho da MLP. O impacto de retirada de cada variável foi calculado sobre a variação do RMSE a cada rodada, em porcentagem. O processo foi interrompido quando não houve mais ganho de desempenho ao se retirar variáveis.

A escolha das porções do banco de dados destinadas a cada uma das parcelas de treinamento (70%), validação (10%) e teste (20%) foi feita de forma aleatória. Este fato faz com que, a cada repetição do modelo, diferentes resultados possam ser encontrados. Outro fator que pode causar diferenças nos resultados a cada repetição é o fato dos pesos iniciais do algoritmo MLP serem definidos aleatoriamente. Dessa forma, para se ter maior segurança sobre o desempenho da MLP a cada rodada do *wrapper*, o método foi repetido 100 vezes para cada variável. Ao final foram desconsiderados os três melhores e os três piores desempenhos e calculado o resultado médio entre as repetições restantes. Para cada repetição variou-se aleatoriamente a quantidade de neurônios na camada intermediária, entre 25% e 100% do número de variáveis em cada rodada.

3.2.2.2. Comitês de modelos preditores

Feita a escolha das variáveis relevantes para a predição da produtividade usando uma MLP simples, avaliou-se o uso de comitês de modelos preditores como forma de melhorar o desempenho encontrado com o uso de uma MLP simples. Similar ao

procedimento feito na etapa de seleção de variáveis, para se obter o desempenho de cada membro do comitê de forma mais segura foram feitas 100 repetições da MLP para cada membro. Em seguida foram desconsiderados os três melhores e três piores desempenhos e então calculado o desempenho médio entre as repetições restantes. Para cada repetição também se variou aleatoriamente a quantidade de neurônios na camada intermediária entre 25% e 100% do número de variáveis.

Foram testados os seguintes métodos de criação de comitês:

- *Bagging*, aplicando média simples na combinação dos resultados de cada membro MLP do comitê;

- *Stacking*, usando um modelo MLP para combinação dos resultados de cada membro MLP do comitê.

Tanto para *Bagging* quanto para *Stacking* foram testados diferentes tamanhos de comitês, avaliando as seguintes quantidades de membros: 2, 3, 5, 10, 20, 30, 50, 100, 200 e 300.

Capítulo 4 – Resultados e Discussão

Primeiramente são apresentados no tópico 4.1 os resultados da análise exploratória feita sobre as variáveis independentes obtidas neste estudo, bem como da variável a ser predita. Em seguida são apresentados os resultados da seleção de variáveis e das predições por comitês de modelos para os seguintes modelos propostos: modelo espectral municipal no tópico 4.2, modelo espectral-meteorológico municipal no tópico 4.3 e modelo reduzido municipal no tópico 4.4. E ao final são apresentados os resultados do modelo estadual no tópico 4.5. As discussões sobre os resultados foram feitas ao longo da apresentação dos resultados e no tópico 4.6.

4.1. Análise exploratória das variáveis construídas

Na Tabela 3 são apresentadas as estatísticas básicas da variável a ser predita, rendimento médio da produção (*rend*), e, na Figura 22, o respectivo histograma. As estatísticas básicas e histogramas de todas as variáveis espectrais e meteorológicas construídas neste estudo estão apresentadas nos Anexos 2 e 3, respectivamente.

Tabela 3: Estatísticas básicas da variável "rendimento médio da produção", em [t ha^{-1}].

Mínimo	50,0
Máximo	130,2
Média	83,5
Mediana	83,0
Desvio padrão	9,2

Analisando a Tabela 3 e a Figura 22 verifica-se que a variável a ser predita, rendimento médio da produção (rend), apresenta uma distribuição próxima da distribuição normal. Verifica-se no Anexo 2 que grande parte das variáveis espectrais também apresentaram este padrão de distribuição.

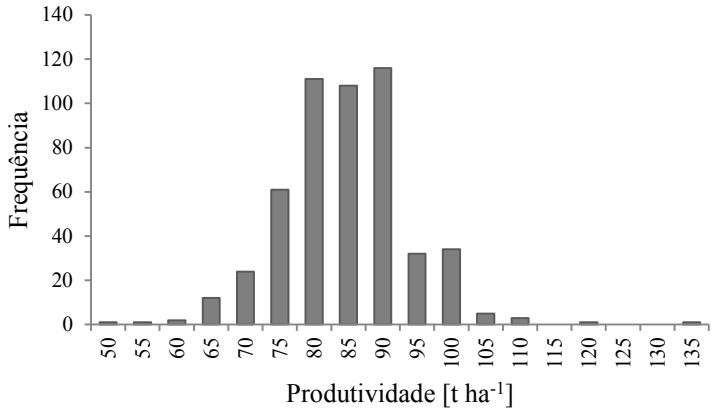


Figura 22: Histograma da variável "rendimento médio da produção".

Na Figura 23 estão mostrados os *box-plot* das variáveis que apresentaram pontos extremos (que excederam três vezes a distância interquartil). Nota-se na Figura 23 que as variáveis meteorológicas *RRR_avg3* e *RAD_avg1*, as variáveis espectrais *ndvi4*, *ndvi5* e *ndvi6* e variável alvo "*rend*" apresentaram pontos extremos (marcados com uma cruz), considerados neste trabalho candidatos a *outlier*.

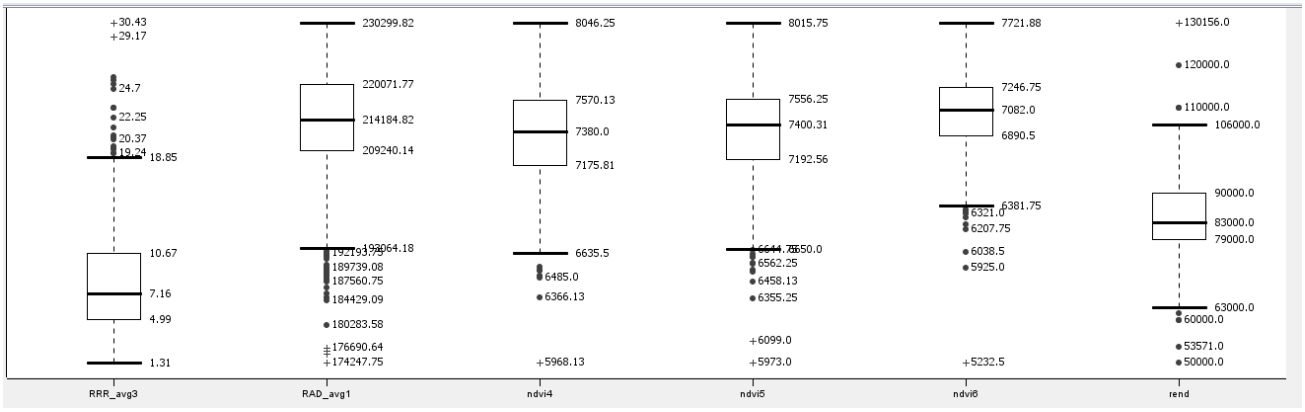


Figura 23: Box-plot das variáveis que apresentaram pontos extremos.

Na etapa de seleção de variáveis (tópicos 4.2.1, 4.3.1 e 4.4.1) o processo foi realizado considerando todos os registros do banco de dados inicial e, posteriormente, retirados os registros que continham pontos extremos (mostrados na Figura 23). Verificou-se que não houve ganho ao se retirar os registros candidatos a *outlier*, inclusive havendo em alguns casos queda de desempenho da rede neural MLP após a

remoção dos registros com pontos extremos. Estes registros então não foram considerados *outliers* e foram mantidos no banco de dados.

Foi feita uma matriz de correlação para a verificação de correlações entre variáveis, contendo todas as variáveis espectrais e meteorológicas construídas neste estudo, além da variável alvo "*rend*", apresentada na Figura 24. De modo geral, verifica-se na Figura 24 que as variáveis espectrais apresentaram maior correlação com a variável alvo "*rend*" do que as variáveis meteorológicas, embora as variáveis construídas neste estudo, em sua maioria, apresentaram pouca correlação com a variável alvo. A variável mais correlacionada com a variável alvo foi a variável espectral "*max value*", que se refere ao valor máximo do NDVI ao longo da safra. Verifica-se também que as variáveis meteorológicas derivadas de *TMIN* apresentaram correlações entre si.

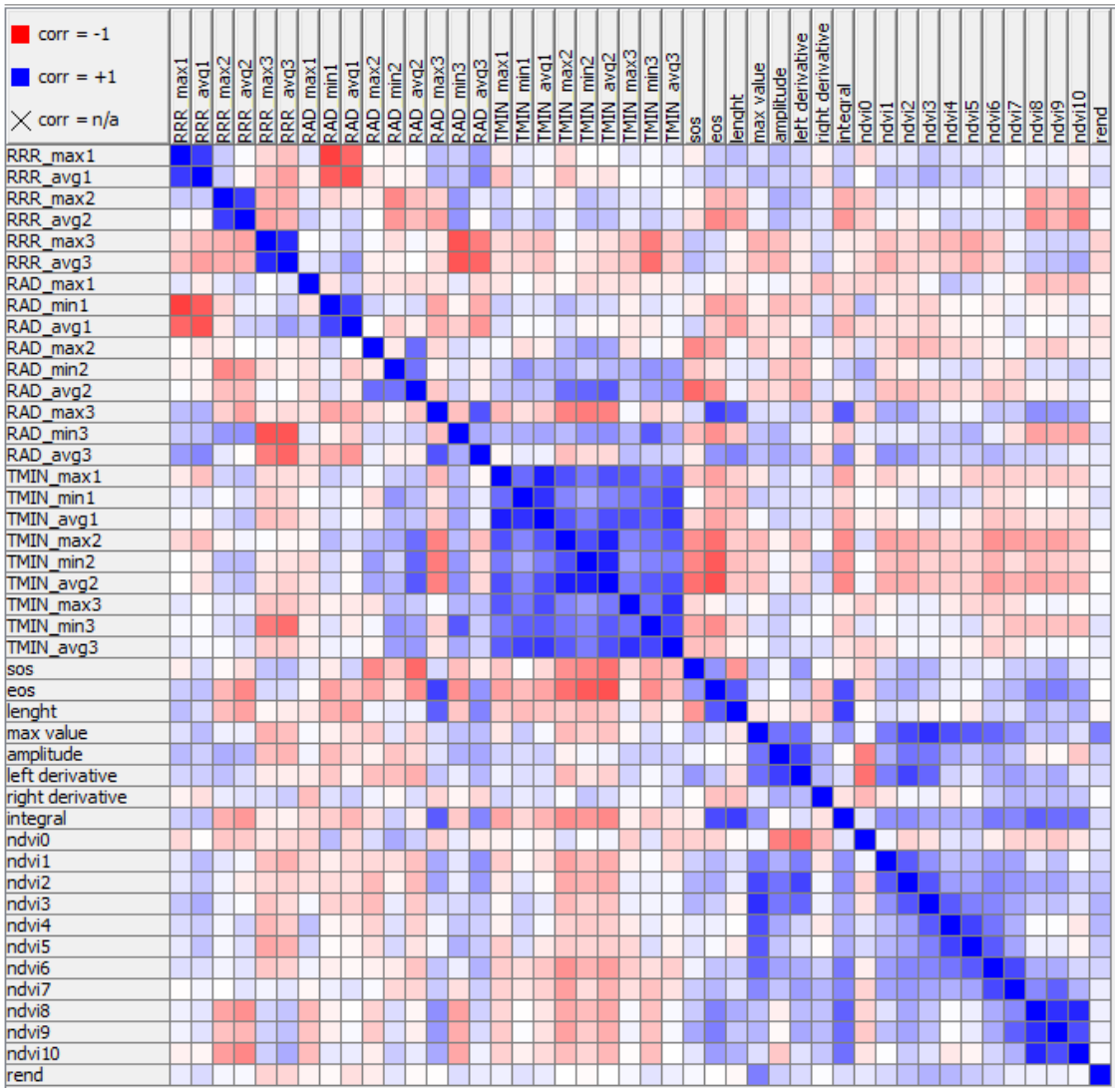


Figura 24: Matriz de correlação contendo todas as variáveis iniciais.

4.2. Modelo espectral municipal

Neste modelo foram consideradas apenas as variáveis espectrais descritas na Tabela 1 com o objetivo de verificar a capacidade em se predizer a produtividade média municipal usando apenas variáveis obtidas por meio das séries temporais do NDVI provenientes de imagens do sensor MODIS. O banco de dados inicial consistiu então em 540 registros e 20 variáveis, sendo 19 variáveis espectrais independentes e a variável dependente "*rend*".

4.2.1. Seleção de variáveis

O *wrapper* com o algoritmo MLP e seleção sequencial para trás foi aplicado ao banco de dados espectral inicial. Na Tabela 4 são apresentados os impactos (I), em porcentagem, de cada variável ao ser removida do banco de dados, para as cinco etapas do *wrapper* realizadas. As variáveis com I positivo foram aquelas que ao serem removidas a cada etapa melhoraram o desempenho da rede neural MLP, ou seja, variáveis que prejudicavam o desempenho do algoritmo de aprendizagem e poderiam ser excluídas do banco de dados. Com o objetivo de se ter maior segurança sobre a exclusão de variáveis, decidiu-se por não extrair, a cada etapa, todas as variáveis com I positivo. Dentre as variáveis com I positivo, optou-se por extrair somente os 30% que representavam as variáveis com maior impacto positivo. Assim, a convergência pôde ser realizada de forma mais lenta e conservadora. Na Tabela 4 são apresentados, para cada etapa do *wrapper*, os impactos (I) de cada variável ao ser removida do banco de dados. As variáveis excluídas a cada etapa (pertencentes aos 30% de maior impacto positivo) estão indicadas em vermelho. Após quatro etapas do *wrapper*, 12 variáveis espectrais foram excluídas. Verifica-se na Tabela 4 que na etapa cinco nenhuma variável apresentou impacto positivo ao ser excluída do banco de dados, finalizando então o processo de seleção de variáveis.

Tabela 4: Impacto (I) no desempenho ao se retirar cada variável do banco de dados espectral, para cada etapa do *wrapper*.

	Etapas do *wrapper* com seleção sequencial para trás									
	1		2		3		4		5	
	Var	I	Var	I	Var	I	Var	I	Var	I
1	*ndvi9*	9,4	*ndvi2*	9,9	*ndvi3*	4,4	*sos_c*	4,6	*ndvi0*	-3,9
2	*ndvi7*	9,2	*ndvi10*	9,8	*eos_c*	3,8	*eos_c*	2,0	*ndvi5*	-4,4
3	*ndvi4*	8,2	*ndvi8*	9,7	*sos_c*	2,8	*ndvi5*	1,3	*right_derivative*	-9,7
4	*ndvi6*	7,1	*ndvi1*	9,7	*lenght*	2,1	*lenght*	0,5	*amplitude*	-34,7
5	*ndvi2*	5,9	*left_derivative*	9,5	*ndvi0*	0,3	*ndvi0*	-6,0	*lenght*	-37,4
6	*eos_c*	4,3	*ndvi3*	6,8	*ndvi5*	-0,6	*right_derivative*	-11,3	*integral*	-38,2
7	*ndvi10*	3,1	*eos_c*	3,6	*right_derivative*	-7,2	*amplitude*	-32,2	*max_value*	-74,1
8	*ndvi1*	3,1	*ndvi5*	3,2	*amplitude*	-23,2	*integral*	-33,9		
9	*ndvi8*	2,8	*lenght*	2,5	*integral*	-25,6	*max_value*	-75,6		
10	*lenght*	1,5	*sos_c*	1,8	*max_value*	-73,8				
11	*ndvi3*	0,0	*right_derivative*	0,8						
12	*sos_c*	-0,7	*ndvi0*	-0,2						
13	*left_derivative*	-2,1	*integral*	-8,9						
14	*right_derivative*	-5,8	*amplitude*	-21,3						
15	*ndvi0*	-6,5	*max_value*	-75,7						
16	*ndvi5*	-7,7								
17	*integral*	-12,5								
18	*amplitude*	-27,5								
19	*max_value*	-73,6								

Na Tabela 5 são apresentados os resultados de desempenho da MLP para o banco de dados selecionado em cada etapa. O RMSE é expresso em t ha^{-1}. A etapa zero (0) representa o desempenho da MLP usando o banco de dados inicial contendo as 19 variáveis espectrais.

Tabela 5: Resultados da seleção de variáveis aplicando o *wrapper* de seleção sequencial para trás sobre o banco de dados espectral.

Etapa	**Treinamento**		**Validação**		**Teste**		**Todos**		**Variáveis excluídas**
	RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	
0	5,7	0,62	7,6	0,42	8,2	0,34	6,3	0,52	Sem exclusão
1	5,8	0,60	7,1	0,42	7,7	0,37	6,3	0,53	*ndvi9, ndvi7, ndvi4, ndvi6*
2	5,8	0,60	6,8	0,47	7,1	0,43	6,2	0,55	*ndvi2, ndvi10, ndvi8, ndvi1, left_derivative*
3	6,1	0,57	6,5	0,46	7,1	0,44	6,3	0,53	*ndvi3*
4	6,1	0,56	6,6	0,50	**6,7**	**0,48**	6,3	0,54	*sos, eos*

Verifica-se na Tabela 5 que, ao se considerar todas as variáveis iniciais (etapa 0), o resultado da MLP para a parcela de teste foi de um RMSE de 8,2 t ha^{-1} e R^2 de 0,34. Após quatro etapas, 12 variáveis que prejudicavam o desempenho da MLP foram excluídas e o desempenho para a parcela de teste melhorou, ocorrendo uma diminuição do RMSE, de 8,2 para 6,7 t ha^{-1} e aumento do R^2, de 0,34 para 0,48. O processo de exclusão de variáveis foi interrompido na quarta etapa do *wrapper*, pois a partir desta etapa a exclusão de variáveis prejudicaria o desempenho da MLP, conforme visto na etapa 5 do *wrapper* mostrado na Tabela 4.

Observa-se também na Tabela 5 que, ao se considerar todas as variáveis (etapa 0), havia grande diferença entre o desempenho obtido na parcela de treinamento (RMSE de 5,7 e R^2 de 0,62) e na parcela de teste (RMSE de 8,2 e R^2 de 0,34), caracterizando *overfitting*. Conforme variáveis irrelevantes / redundantes foram excluídas, verificou-se na etapa 4 que, além da melhora de desempenho na parcela de teste (RMSE de 6,7 e R^2 de 0,48), houve também diminuição na diferença entre os desempenhos nas parcelas de treinamento e de teste, minimizando o *overfitting*.

As sete variáveis finais selecionadas estão apresentadas na Tabela 6 em ordem decrescente de importância.

Tabela 6: Variáveis selecionadas para o modelo espectral municipal, em ordem decrescente de importância.

	Variáveis selecionadas
1	*max_value*
2	*integral*
3	*length*
4	*amplitude*
5	*right_derivative*
6	*ndvi5*
7	*ndvi0*

Conclui-se que, das 19 variáveis espectrais usadas neste estudo, obtidas ao longo de toda a safra, somente sete variáveis apresentaram relevância para a predição da produtividade média municipal por meio de redes neurais MLP.

4.2.2. Predição por comitês

A fim de verificar se o uso de comitês de modelos melhoraria o desempenho de predição de uma MLP simples, foram aplicados os métodos *Bagging* e *Stacking* propostos no tópico 3.2.2.2 sobre o banco de dados com as variáveis selecionadas no tópico 4.2.1. Assim, o banco de dados contendo as sete variáveis espectrais selecionadas, acrescentadas da variável "*rend*", foram os dados de entrada nos modelos de predição por comitês. A seguir são apresentados os resultados para os comitês *Bagging* (a) e *Stacking* (b).

a) *Bagging*

Foram avaliados comitês *Bagging* com diferentes quantidades de membros e usando a média aritmética como método agregador das saídas de cada membro. Na Tabela 7 estão apresentados os resultados de predição usando *Bagging* de MLP para as diferentes quantidades de membros no comitê avaliadas no modelo espectral municipal.

Tabela 7: Resultados de RMSE [t ha^{-1}] e R^2 para as diferentes quantidades de membros no *Bagging* do modelo espectral municipal.

QM*	Treinamento		Validação		Teste		Todos	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
2	6,9	0,44	6,9	0,43	7,0	0,44	6,9	0,44
3	7,2	0,39	7,4	0,43	7,1	0,41	7,2	0,39
5	7,3	0,42	7,2	0,39	7,1	0,42	7,3	0,42
10	7,5	0,42	7,5	0,43	7,2	0,43	7,4	0,42
20	7,3	0,51	7,6	0,50	7,5	0,49	7,4	0,51
30	7,2	0,58	7,0	0,57	7,0	0,56	7,2	0,58
50	7,4	0,55	7,3	0,52	7,4	0,51	7,4	0,54
100	7,4	0,62	7,3	0,62	7,1	0,61	7,3	0,62
200	7,4	0,63	7,3	0,62	7,4	0,60	7,4	0,62
300	7,3	0,64	7,0	0,62	7,4	0,63	7,3	0,63

***** Quantidade de Membros.

Verifica-se na Tabela 7 que os resultados de RMSE para o *Bagging* de MLP foram piores que o resultado de uma MLP simples (6,7 t ha^{-1}), embora comitês com grande número de membros tenham apresentado valores de R^2 superiores ao valor encontrado com uma MLP simples. Isso se deve ao método de agregação dos resultados dos membros do comitê, que foi por meio da média aritmética. Na Figura 25c é

apresentado o gráfico de dispersão entre a produtividade observada e a produtividade predita pelo *Bagging* composto por 30 membros, onde é possível observar a tendência dos valores de produtividade predita se aproximarem do valor médio de produtividade, o que fez com que o *Bagging* apresentasse um RMSE superior a uma MLP simples. Dessa forma, conclui-se que o *Bagging* de MLP proposto não foi adequado para a predição da variável "*rend*" no modelo espectral municipal.

b) *Stacking*

Aqui o método *Stacking* foi aplicado ao banco de dados espectral selecionado no tópico 5.2.1. Na Tabela 8 são apresentados os resultados de predição usando *Stacking* composto de membros MLP e rede neural MLP como algoritmo agregador para o modelo espectral municipal.

Tabela 8: Resultados de RMSE [t ha^{-1}] e R^2 para diferentes quantidades de membros no *Stacking* do modelo espectral municipal.

QM*	Treinamento		Validação		Teste		Todos	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
2	5,4	0,65	5,6	0,61	**5,7**	**0,61**	5,5	0,64
3	5,8	0,61	5,9	0,57	5,9	0,59	5,8	0,60
5	5,6	0,64	5,7	0,63	6,0	0,59	5,8	0,63
10	5,4	0,66	5,7	0,61	6,1	0,58	5,6	0,64
20	5,3	0,67	6,6	0,53	6,5	0,52	5,7	0,62
30	5,2	0,68	6,1	0,56	6,3	0,53	5,5	0,64
50	5,6	0,63	6,9	0,40	9,1	0,42	6,9	0,54
100	6,3	0,52	8,7	0,28	9,1	0,23	7,3	0,41
200	6,7	0,54	9,9	0,19	10,2	0,20	8,1	0,39
300	5,8	0,58	9,8	0,21	10,6	0,16	7,8	0,40

* Quantidade de Membros,

De acordo com os resultados da Tabela 8, *Stacking* composto de poucos membros foi adequado para a predição da variável "*rend*", melhorando os resultados de uma simples MLP. O melhor resultado foi encontrado com um *Stacking* composto de dois membros, apresentando um RMSE de 5,7 t ha^{-1} e R^2 de 0,61. O desempenho do *Stacking* piorou conforme se aumentou a quantidade de membros no comitê. Percebe-se na Tabela 8 a ocorrência de *overfitting* conforme se aumenta a quantidade de membros nos comitês.

Na Figura 25 são apresentados os gráficos de dispersão entre os dados de produtividade observados (dados oficiais do IBGE) e os dados de produtividade preditos pelos modelos, para uma simples MLP e considerando todas as 19 variáveis espectrais iniciais (a), para uma simples MLP e apenas as sete variáveis espectrais selecionadas pelo *wrapper* (b), para as sete variáveis espectrais selecionadas e o *Bagging* de 30 membros MLP (c) e para as sete variáveis espectrais selecionadas e o *Stacking* de dois membros MLP e rede neural MLP como algoritmo agregador (d). Os pontos nos gráficos representam todas as instâncias do banco de dados (treinamento, validação e teste).

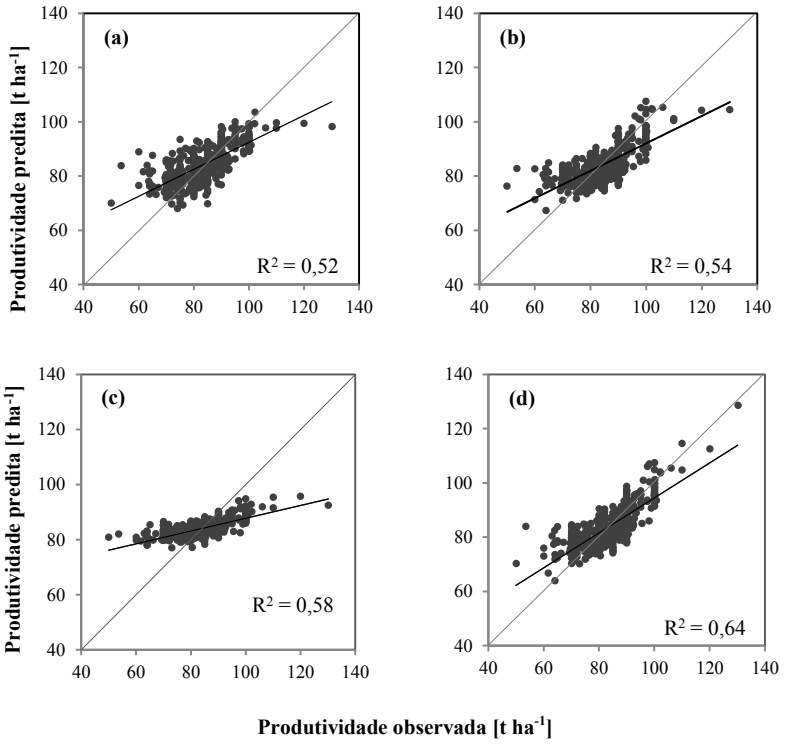


Figura 25: Gráficos de dispersão entre dados observados e dados preditos de produtividade, para uma simples MLP e o banco de dados completo, para uma simples MLP e as variáveis selecionadas (b), para o *Bagging* de 30 membros (c) e para o *Stacking* de dois membros (d).

Verifica-se na Figura 25 que a seleção de variáveis (Figura 25b) e principalmente o uso de comitê *Stacking* composto por dois membros (Figura 25d) foram capazes de melhorar o ajuste entre os dados de produtividade observados e os dados preditos.

4.3. Modelo espectral-meteorológico municipal

Com o objetivo de verificar o possível ganho no desempenho de predição da produtividade média municipal ao se incorporar variáveis meteorológicas ao modelo espectral municipal proposto no tópico 4.2, repetiu-se o procedimento de seleção de variáveis e predição por comitê de modelos, partindo-se de um banco de dados inicial contendo todas as variáveis espectrais e meteorológicas construídas neste estudo. Assim, o banco de dados inicial foi composto de todas as variáveis espectrais descritas na Tabela 1 e todas as variáveis meteorológicas descritas na Tabela 2. O banco de dados espectral-meteorológico inicial consistiu então em 540 registros e 44 variáveis, sendo 19 variáveis espectrais, 24 variáveis meteorológicas e a variável dependente "*rend*".

4.3.1. Seleção de variáveis

A seleção de variáveis ocorreu da mesma forma aplicada para o modelo espectral municipal descrito no tópico 4.2.1. O *wrapper* com o algoritmo MLP e seleção sequencial para trás foi aplicado ao banco de dados espectral-meteorológico inicial. Na Tabela 9 são apresentados os impactos (I), em porcentagem, de cada variável ao ser removida do banco de dados a cada etapa, para as 11 etapas do *wrapper* realizadas. As variáveis excluídas a cada etapa (pertencentes aos 30% de maior impacto positivo) estão indicadas em vermelho na Tabela 9.

Verifica-se na etapa 11 da Tabela 9 que nenhuma variável apresentou impacto positivo ao ser excluída do banco de dados, finalizando então o processo de seleção de variáveis.

Tabela 9: Impacto (I) no desempenho ao se retirar cada variável do banco de dados espectral-meteorológico, para cada etapa do *wrapper*.

	Etapa do *wrapper* com seleção sequencial para trás									
	1		2		3		4		5	
	Var	I	Var	I	Var	I	Var	I	Var	I
1	*TMIN_min2*	15	*RRR_avg3*	9.5	*RRR_avg2*	3	*TMIN_avg3*	13	*TMIN_min1*	10
2	*RRR_max2*	13	*TMIN_avg1*	8.6	*ndvi4*	3	*ndvi9*	13	*TMIN_min3*	10
3	*RAD_min1*	12	*ndvi7*	7.5	*ndvi6*	3	*ndvi1*	13	*ndvi2*	9.9
4	*RRR_avg1*	11	*RRR_max1*	7.2	*TMIN_max1*	0.8	*RRR_max3*	11	*RAD_max1*	7.5
5	*TMIN_avg2*	11	*RAD_avg2*	7	*RAD_max3*	-0	*RAD_avg3*	11	*TMIN_max2*	6.1
6	*TMIN_max3*	10	*ndvi10*	6	*ndvi8*	-0.6	*TMIN_max1*	7.8	*left_derivative*	6
7	*TMIN_min1*	8.8	*RRR_max3*	5.2	*ndvi2*	-0.9	*ndvi8*	7.7	*RAD_min2*	6
8	*RRR_max1*	7.7	*ndvi6*	4.9	*RAD_avg3*	-0.9	*TMIN_max2*	6.7	*sos_c*	5.9
9	*ndvi10*	7.3	*RAD_max3*	4.6	*TMIN_avg3*	-2	*right_derivative*	6.6	*RAD_min3*	5.9
10	*ndvi7*	7	*ndvi3*	4.4	*ndvi1*	-2.7	*ndvi2*	6.6	*RAD_max3*	3.7
11	*TMIN_min3*	7	*ndvi9*	4.3	*TMIN_min1*	-3.1	*TMIN_min3*	6.6	*ndvi5*	3.6
12	*ndvi3*	6.9	*RAD_min3*	4.1	*eos_c*	-3.9	*left_derivative*	6.3	*ndvi0*	3.6
13	*TMIN_max2*	6.8	*TMIN_min3*	3.4	*RAD_max2*	-3.9	*RAD_max2*	6.2	*ndvi3*	3.5
14	*eos_c*	6	*TMIN_min1*	2.7	*ndvi9*	-4	*RAD_max3*	6.1	*RAD_avg1*	3.5
15	*RAD_max1*	5.8	*TMIN_max1*	1	*ndvi0*	-4.6	*eos_c*	4.3	*RAD_max2*	1.4
16	*ndvi8*	5.7	*ndvi8*	1	*RRR_max3*	-4.8	*ndvi3*	4.2	*eos_c*	1.3
17	*ndvi4*	5.4	*RAD_max2*	0.8	*RAD_max1*	-4.9	*RAD_min2*	4.1	*right_derivative*	-7
18	*RAD_min2*	5.3	*RAD_avg3*	0.5	*RAD_min2*	-5.6	*ndvi0*	3.8	*lenght*	-34
19	*RAD_max2*	5	*eos_c*	-2	*RAD_avg1*	-5.7	*RAD_min3*	3.4	*integral*	-41
20	*TMIN_avg3*	4.9	*ndvi0*	-3	*TMIN_max2*	-6.4	*TMIN_min1*	2.9	*amplitude*	-43
21	*RAD_avg1*	4.6	*left_derivative*	-3	*sos_c*	-6.5	*RAD_max1*	2.2	*max_value*	-70
22	*RAD_min3*	3.9	*RAD_avg1*	-3	*TMIN_min3*	-6.5	*RAD_avg1*	-4.7		
23	*left_derivative*	3.6	*sos_c*	-3	*left_derivative*	-7	*sos_c*	-5		
24	*RAD_avg2*	3.2	*ndvi4*	-3	*ndvi5*	-7.9	*ndvi5*	-20		
25	*ndvi9*	2.4	*ndvi2*	-4	*right_derivative*	-9.2	*lenght*	-28		
26	*RRR_avg2*	1.9	*RRR_avg2*	-4	*RAD_min3*	-11	*integral*	-40		
27	*RRR_max3*	-0.4	*RAD_max1*	-5	*ndvi3*	-23	*amplitude*	-40		
28	*sos_c*	-1	*TMIN_avg3*	-5	*lenght*	-31	*max_value*	-74		
29	*ndvi2*	-1.1	*RAD_min2*	-6	*amplitude*	-42				
30	*ndvi5*	-1.1	*right_derivative*	-10	*integral*	-46				
31	*TMIN_max1*	-1.1	*ndvi5*	-11	*max_value*	-76				
32	*RRR_avg3*	-1.6	*TMIN_max2*	-11						
33	*ndvi6*	-2.7	*ndvi1*	-19						
34	*TMIN_avg1*	-5.9	*lenght*	-22						
35	*RAD_max3*	-8	*amplitude*	-43						
36	*ndvi0*	-8.4	*integral*	-45						
37	*ndvi1*	-9.6	*max_value*	-78						
38	*RAD_avg3*	-11								
39	*right_derivative*	-11								
40	*lenght*	-19								
41	*amplitude*	-39								
42	*integral*	-44								
43	*max_value*	-78								

Continuação da Tabela 9...

	6		7		8		9		10	
	Var	I	Var	I	Var	I	Var	I	Var	I
1	*RAD_min3*	8.5	*RAD_max3*	11	*RAD_max2*	10	*TMIN_max2*	1.8	*RAD_avg1*	2.4
2	*RAD_min2*	5.8	*eos_c*	3.9	*sos_c*	9.8	*ndvi0*	1.7	*ndvi5*	-5.7
3	*RAD_avg1*	2.8	*ndvi3*	3.8	*left_derivative*	9.3	*RAD_avg1*	-1.1	*right_derivative*	-11
4	*left_derivative*	2	*RAD_avg1*	3.6	*TMIN_max2*	8.5	*ndvi5*	-4.2	*lenght*	-38
5	*RAD_max2*	1.8	*left_derivative*	3.5	*right_derivative*	3	*right_derivative*	-6.1	*amplitude*	-42
6	*sos_c*	1.6	*RAD_max2*	3.4	*ndvi0*	2.7	*integral*	-36	*integral*	-45
7	*TMIN_max2*	1.5	*sos_c*	3	*ndvi5*	2.5	*lenght*	-37	*max_value*	-73
8	*ndvi3*	1.5	*TMIN_max2*	2.3	*RAD_avg1*	2.3	*amplitude*	-38		
9	*eos_c*	1.2	*right_derivative*	-5	*amplitude*	-31	*max_value*	-77		
10	*RAD_max3*	-0.4	*ndvi5*	-23	*lenght*	-32				
11	*ndvi0*	-1	*ndvi0*	-26	*integral*	-35				
12	*ndvi5*	-4.5	*lenght*	-35	*max_value*	-70				
13	*right_derivative*	-5.3	*amplitude*	-36						
14	*amplitude*	-38	*integral*	-46						
15	*lenght*	-39	*max_value*	-73						
16	*integral*	-39								
17	*max_value*	-70								

	11	
	Var	I
1	*ndvi5*	-3.9
2	*right_derivative*	-7.8
3	*amplitude*	-38
4	*integral*	-43
5	*lenght*	-43
6	*max_value*	-74

Na Tabela 10 são apresentados os resultados de desempenho da MLP para o banco de dados selecionado em cada etapa. A etapa zero (0) representa o banco de dados inicial, contendo as 19 variáveis espectrais e 24 variáveis meteorológicas. O procedimento de seleção de variáveis avançou até a etapa 10 do *wrapper*. A partir da etapa 10 a remoção de variáveis causaria prejuízos ao desempenho da MLP.

Tabela 10: Resultados da seleção de variáveis aplicando o *wrapper* de seleção sequencial para trás sobre o banco de dados espectral-meteorológico.

Etapa	Treinamento		Validação		Teste		Todos		Variáveis excluídas
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	
0	6,0	0,59	8,9	0,29	9,2	0,24	7,2	0,46	Sem exclusão
1	5,8	0,60	7,6	0,37	8,3	0,30	6,6	0,50	*TMIN_min2, RRR_max2, RAD_min1, RRR_avg1, TMIN_avg2, TMIN_max3*
2	6,0	0,56	7,4	0,38	7,7	0,36	6,6	0,49	*RRR_avg3, TMIN_avg1, ndvi7, RRR_max1, RAD_avg2, ndvi10*
3	5,9	0,59	7,5	0,40	7,7	0,34	6,6	0,51	*RRR_avg2, ndvi4, ndvi6*
4	6,2	0,55	7,2	0,41	7,7	0,37	6,7	0,49	*TMIN_avg3, ndvi9, ndvi1, RRR_max3, RAD_avg3, TMIN_max1, ndvi8*
5	6,1	0,57	6,8	0,42	7,3	0,40	6,4	0,52	*TMIN_min1, TMIN_min3, ndvi2, RAD_max1*
6	6,3	0,53	6,9	0,42	7,1	0,41	6,6	0,49	*RAD_min3, RAD_min2*
7	6,4	0,53	6,7	0,44	7,1	0,41	6,6	0,49	*RAD_max3, eos, ndvi3*
8	6,4	0,54	6,9	0,44	6,9	0,44	6,6	0,49	*RAD_max2, sos, left_derivative*
9	6,5	0,52	6,7	0,48	6,9	0,46	6,6	0,49	*TMIN_max2, ndvi0*
10	6,3	0,52	6,6	0,48	6,8	0,47	6,5	0,51	*RAD_avg1*

Na Tabela 11 são apresentadas as seis variáveis selecionadas, em ordem decrescente de importância.

Tabela 11: Variáveis selecionadas para o modelo espectral-meteorológico municipal, em ordem decrescente de importância.

	Variáveis selecionadas
1	*max_value*
2	*length*
3	*integral*
4	*amplitude*
5	*right_derivative*
6	*ndvi5*

Verifica-se na Tabela 11 que apenas variáveis espectrais foram selecionadas, excluindo todas as variáveis meteorológicas ao longo das etapas do *wrapper*. A não seleção de variáveis meteorológicas pode ser justificada devido a algumas condições presentes neste estudo, como a escala espacial usada (média municipal), o uso de variáveis meteorológicas ao invés de variáveis agrometeorológicas e o fato de as variáveis espectrais poderem conter implicitamente informações meteorológicas, uma vez que se referem ao desenvolvimento da cultura. Com isso, conclui-se que as variáveis meteorológicas usadas neste estudo não foram capazes de melhorar os resultados obtidos com as variáveis espectrais para a predição da produtividade média municipal por meio de redes neurais MLP. Discussões detalhadas a respeito da não seleção das variáveis meteorológicas foram feitas no tópico 4.6 desta tese.

Devido ao resultado similar ao encontrado na seleção de variáveis do modelo espectral municipal do tópico 4.2.1, não houve necessidade de realizar a predição por comitês para o modelo espectral-meteorológico, uma vez que seria o próprio modelo espectral municipal executado no tópico 4.2.2.

4.4. Modelo reduzido municipal

Com o objetivo de verificar a capacidade em se estimar a produtividade da cana-de-açúcar antecipadamente ao final da safra, foram excluídas do banco de dados espectral-meteorológico inicial usado no tópico 4.3 todas as variáveis que dependessem dos três últimos meses da safra. Assim, sete variáveis espectrais (*eos*, *length, right derivative, integral, ndvi8, ndvi9* e *ndvi10*) e oito variáveis meteorológicas (*TMIN_min3, TMIN_max3, TMIN_avg3, RAD_min3, RAD_max3, RAD_avg3, RRR_max3* e *RRR_avg3*) foram excluídas do banco de dados inicial. Este banco de dados reduzido passou a conter 540 registros e 29 variáveis, sendo 12 variáveis espectrais, 16 variáveis meteorológicas e a variável alvo "rend", e foi usado como entrada no procedimento de seleção de variáveis para o modelo reduzido municipal.

4.4.1. Seleção de variáveis

O *wrapper* com o algoritmo MLP e seleção sequencial para trás foi aplicado ao banco de dados reduzido. Na Tabela 12 são apresentados os impactos (I), em porcentagem, de cada variável ao ser removida do banco de dados a cada etapa, para as 12 etapas do *wrapper* realizadas. As variáveis excluídas a cada etapa (pertencentes aos 30% de maior impacto positivo) estão indicadas em vermelho na Tabela 12. Verifica-se que na etapa 12 nenhuma variável apresentou impacto positivo ao ser excluída do banco de dados, finalizando então o processo de seleção de variáveis.

Tabela 12: Impacto (I) no desempenho ao se retirar cada variável do banco de dados espectral-meteorológico reduzido, para cada etapa do *wrapper*.

	Etapa do *wrapper* com seleção sequencial para trás									
	1		2		3		4		5	
	Var	I	Var	I	Var	I	Var	I	Var	I
1	*TMIN_min2*	13	*sos*	6,7	*RAD_min1*	15	*RAD_min2*	7,4	*TMIN_min1*	6,9
2	*RAD_avg2*	6	*RRR_avg1*	4,9	*amplitude*	15	*ndvi2*	7,1	*RRR_max1*	4
3	*ndvi4*	5,9	*RAD_avg2*	3,9	*ndvi1*	13	*TMIN_min1*	4	*ndvi4*	2,5
4	*ndvi2*	5,5	*ndvi4*	3,8	*RAD_avg2*	11	*TMIN_avg2*	2,8	*RAD_avg1*	1,3
5	*TMIN_max1*	5	*TMIN_avg1*	2,9	*ndvi2*	10	*ndvi4*	2,5	*TMIN_avg2*	0,9
6	*RAD_min2*	3,5	*RAD_max1*	1,5	*ndvi7*	10	*RRR_avg2*	0,1	*RAD_max1*	0,6
7	*TMIN_avg2*	3,4	*ndvi1*	0,9	*left derivative*	8,8	*RAD_max2*	-0,4	*RRR_avg2*	-0,2
8	*sos*	3,2	*TMIN_min1*	0,2	*TMIN_avg2*	7,9	*ndvi7*	-1	*TMIN_max1*	-0,6
9	*RRR_avg2*	2,6	*amplitude*	-1	*RAD_avg1*	7,2	*ndvi0*	-1,5	*ndvi6*	-2
10	*ndvi5*	0,2	*ndvi7*	-2	*ndvi0*	6,1	*TMIN_max1*	-1,7	*TMIN_avg1*	-2,1
11	*TMIN_max2*	-0,2	*ndvi2*	-3	*TMIN_max1*	5,7	*RRR_max1*	-1,7	*RRR_max2*	-2,1
12	*amplitude*	-0,8	*ndvi5*	-3	*RAD_min2*	5,5	*RAD_avg1*	-2,5	*TMIN_max2*	-2,5
13	*ndvi1*	-1,2	*ndvi6*	-4	*TMIN_min1*	5,2	*TMIN_max2*	-2,6	*RAD_max2*	-4,3
14	*RRR_max2*	-1,2	*TMIN_max2*	-4	*RRR_avg2*	4,7	*ndvi6*	-4,1	*ndvi3*	-5
15	*ndvi0*	-1,8	*RAD_min1*	-5	*ndvi6*	4,4	*RRR_max2*	-4,2	*ndvi5*	-5,4
16	*ndvi7*	-3,1	*TMIN_avg2*	-7	*ndvi4*	4,1	*TMIN_avg1*	-4,5	*ndvi7*	-9,5
17	*RAD_max2*	-3,2	*left derivative*	-7	*RRR_max2*	2,2	*ndvi5*	-4,6	*left derivative*	-10
18	*RAD_min1*	-3,5	*RRR_max1*	-8	*RRR_max1*	2	*left derivative*	-7,7	*ndvi0*	-11
19	*RAD_avg1*	-3,5	*RAD_min2*	-9	*RAD_max1*	0,5	*ndvi3*	-10	*max value*	-76
20	*RRR_max1*	-3,6	*TMIN_max1*	-10	*TMIN_max2*	-0,4	*RAD_max1*	-10		
21	*RRR_avg1*	-3,6	*RAD_avg1*	-12	*ndvi3*	-1,3	*max value*	-73		
22	*RAD_max1*	-3,9	*ndvi0*	-13	*ndvi5*	-2				
23	*TMIN_min1*	-4	*RAD_max2*	-15	*TMIN_avg1*	-2,3				
24	*ndvi6*	-4,5	*RRR_max2*	-15	*RAD_max2*	-7,1				
25	*TMIN_avg1*	-8,5	*RRR_avg2*	-18	*max value*	-71				
26	*left derivative*	-10	*ndvi3*	-19						
27	*ndvi3*	-14	*max value*	-71						
28	*max value*	-73								

Continuação da Tabela 12...

	6		7		8		9		10	
	Var	I	Var	I	Var	I	Var	I	Var	I
1	*RAD_max2*	17	*TMIN_max2*	4,3	*TMIN_avg2*	7,3	*RRR_max2*	6,4	*RAD_max1*	6,3
2	*RRR_avg2*	14	*RRR_max1*	2,4	*RRR_max1*	1,4	*RRR_max1*	4,6	*ndvi6*	5,5
3	*RAD_avg1*	12	*ndvi6*	1,9	*ndvi7*	-0,2	*RAD_avg1*	4,5	*TMIN_avg1*	4,3
4	*RAD_max1*	11	*RRR_max2*	1,7	*RAD_avg1*	-2,9	*TMIN_max1*	4,4	*ndvi0*	1,3
5	*ndvi7*	10	*ndvi7*	0,2	*RAD_max1*	-3	*ndvi5*	2,7	*ndvi4*	-1,1
6	*TMIN_avg2*	9,9	*RAD_avg1*	-0	*ndvi6*	-3,3	*ndvi6*	2,5	*ndvi7*	-1,3
7	*RRR_max2*	9,8	*TMIN_avg1*	-2	*ndvi4*	-4,3	*ndvi7*	2,4	*ndvi5*	-4,7
8	*ndvi0*	7,4	*TMIN_max1*	-2	*ndvi3*	-4,7	*RAD_max1*	2	*ndvi3*	-4,9
9	*TMIN_avg1*	5,5	*TMIN_avg2*	-2	*RRR_max2*	-6,4	*ndvi0*	-0	*left derivative*	-18
10	*TMIN_max2*	4,3	*ndvi3*	-2	*TMIN_max1*	-6,6	*ndvi4*	-2,4	*max value*	-78
11	*TMIN_max1*	4,2	*RAD_max1*	-2	*TMIN_avg1*	-8,1	*TMIN_avg1*	-2,8		
12	*ndvi6*	2,1	*ndvi4*	-5	*ndvi5*	-9,2	*ndvi3*	-4,3		
13	*ndvi4*	1,9	*ndvi5*	-8	*ndvi0*	-12	*left derivative*	-16		
14	*RRR_max1*	1,1	*ndvi0*	-14	*left derivative*	-14	*max value*	-72		
15	*ndvi5*	-0,4	*left derivative*	-19	*max value*	-76				
16	*ndvi3*	-4,4	*max value*	-75						
17	*left derivative*	-16								
18	*max value*	-71								

	11		12	
	Var	I	Var	I
1	*ndvi7*	10,1	*ndvi3*	-2,3
2	*ndvi4*	10	*ndvi0*	-2,5
3	*ndvi0*	4	*left derivative*	-11
4	*ndvi3*	-2	*ndvi5*	-15
5	*ndvi5*	-6	*max value*	-76
6	*left derivative*	-10		
7	*max value*	-71		

Na Tabela 13 são apresentados os resultados de desempenho da MLP para cada etapa do *wrapper*. A etapa zero (0) representa o banco de dados inicial, contendo as 12 variáveis espectrais e as 16 variáveis meteorológicas.

Tabela 13: Resultados da seleção de variáveis aplicando o *wrapper* de seleção sequencial para trás sobre o banco de dados espectral-meteorológico reduzido.

Etapa	Treinamento		Validação		Teste		Todos		Variáveis excluídas
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	
0	5,9	0,60	8,5	0,30	9,0	0,26	7,0	0,48	Sem exclusão
1	6,2	0,57	8,2	0,29	8,6	0,27	7,0	0,46	*TMIN_min2*
2	6,0	0,57	8,1	0,33	8,5	0,26	6,9	0,47	*sos, RRR_avg1*
3	6,2	0,55	7,6	0,38	8,0	0,31	6,8	0,47	*RAD_min1, amplitude, ndvi1, RAD_avg2*
4	6,1	0,57	7,6	0,36	7,9	0,33	6,7	0,49	*RAD_min2, ndvi2*
5	6,0	0,58	7,7	0,36	7,9	0,32	6,7	0,49	*TMIN_min1*
6	6,2	0,56	7,1	0,42	7,6	0,35	6,6	0,50	*RAD_max2, RRR_avg2*
7	6,1	0,57	7,3	0,38	7,6	0,38	6,5	0,51	*TMIN_max2*
8	6,1	0,56	7,0	0,43	7,5	0,36	6,6	0,50	*TMIN_avg2*
9	6,1	0,57	6,9	0,41	7,4	0,38	6,5	0,51	*RRR_max2, RRR_max1, RAD_avg1, TMIN_max1*
10	6,5	0,52	7,2	0,43	7,3	0,36	6,9	0,47	*RAD_max1, ndvi6, TMIN_avg1*
11	6,7	0,48	7,0	0,41	**7,3**	**0,38**	6,9	0,45	*ndvi7, ndvi4*

Na etapa zero (0) do *wrapper* (sem exclusão de variáveis) o desempenho de uma MLP simples na parcela de teste do banco de dados foi de um RMSE de 9,0 t ha^{-1} e de um R^2 de 0,26. Após 11 etapas, 23 variáveis foram excluídas e o desempenho para a parcela de teste melhorou, ocorrendo uma diminuição do RMSE, de 9,0 para 7,3 t ha^{-1} e aumento do R^2, de 0,26 para 0,38. O processo de exclusão de variáveis foi interrompido na 11ª etapa do *wrapper*, pois a partir desta etapa a exclusão de variáveis prejudicaria o desempenho da MLP, conforme mostrado na etapa 12 da Tabela 12, em que a exclusão de qualquer variável acarretaria I negativo. Da mesma maneira que ocorreu na seleção de variáveis do tópico 4.3.1, todas as variáveis meteorológicas presentes no banco de dados reduzido foram excluídas.

Verifica-se na Tabela 13 que, conforme as etapas do *wrapper* foram realizadas, a exclusão de variáveis irrelevantes / redundantes minimizou o *overfitting* ocorrido na etapa 0 (zero), em que ainda não se havia excluído variáveis.

As cinco variáveis finais selecionadas estão apresentadas na Tabela 14 em ordem decrescente de importância.

Tabela 14: Variáveis selecionadas para o modelo reduzido municipal, em ordem decrescente de importância.

	Variáveis selecionadas
1	*max_value*
2	*ndvi5*
3	*left_derivative*
4	*ndvi0*
5	*ndvi3*

4.4.2. Predição por comitês

Feita a seleção de variáveis, foram avaliados para o modelo reduzido municipal se o uso de comitês de modelos melhoraria o desempenho de uma MLP simples. As cinco variáveis espectrais apresentadas na Tabela 14, acrescentadas da variável "*rend*", foram os dados de entrada nos modelos de predição por comitês. A seguir são apresentados os resultados para os comitês *Bagging* (a) e *Stacking* (b).

a) *Bagging*

Na Tabela 15 estão apresentados os resultados de predição usando *Bagging* de MLP no modelo reduzido municipal, para cada tamanho de comitê avaliado.

Tabela 15: Resultados de RMSE [t ha^{-1}] e R^2 para diferentes quantidades de membros no *Bagging* do modelo reduzido municipal.

QM	Treinamento		Validação		Teste		Todos	
	RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
2	7,5	0,32	7,5	0,32	7,6	0,33	7,6	0,32
3	7,7	0,29	8,0	0,31	7,8	0,30	7,8	0,29
5	7,8	0,30	7,8	0,31	7,8	0,32	7,8	0,31
10	7,8	0,34	7,7	0,32	8,0	0,35	7,8	0,34
20	7,9	0,38	7,9	0,36	7,7	0,37	7,9	0,37
30	7,9	0,36	8,1	0,39	8,1	0,35	8,0	0,36
50	7,9	0,39	8,3	0,40	8,0	0,40	8,0	0,39
100	8,0	0,41	7,7	0,44	8,0	0,44	8,0	0,42
200	8,0	0,45	7,9	0,45	7,7	0,44	7,9	0,44
300	8,0	0,39	7,9	0,39	7,8	0,40	8,0	0,38

* Quantidade de membros.

Observa-se na Tabela 15 que novamente comitês obtidos por *Bagging* de MLP não foram capazes de melhorar os resultados obtidos com a MLP simples usada no *wrapper*.

b) *Stacking*

Na Tabela 16 são apresentados os resultados de predição usando *Stacking* composto de membros MLP e rede neural MLP como algoritmo agregador para o modelo reduzido municipal.

Tabela 16: Resultados de RMSE [t ha^{-1}] e R^2 para diferentes quantidades de membros no *Stacking* do modelo reduzido municipal.

QM*	Treinamento		Validação		Teste		Todos	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
2	6,9	0,43	7,4	0,41	7,2	0,43	7,0	0,43
3	6,8	0,44	6,5	0,45	7,2	0,41	6,9	0,43
5	6,6	0,46	6,7	0,43	**6,7**	**0,43**	6,7	0,45
10	6,8	0,45	7,2	0,40	7,2	0,39	6,9	0,43
20	6,7	0,47	7,5	0,36	9,8	0,33	7,9	0,42
30	6,8	0,46	7,4	0,32	8,0	0,31	7,2	0,40
50	6,6	0,49	7,7	0,34	8,0	0,26	7,0	0,42
100	8,0	0,31	8,8	0,23	9,6	0,18	8,5	0,27
200	7,8	0,35	9,7	0,09	11,0	0,11	9,0	0,24
300	7,4	0,45	10,3	0,12	10,6	0,10	8,5	0,28

* Quantidade de membros.

Verifica-se na Tabela 16 que o *Stacking* composto de cinco membros foi adequado para a predição da variável "*rend*", melhorando os resultados de uma simples MLP. O *Stacking* composto de cinco membros apresentou um RMSE de 6,7 t ha^{-1} e R^2 de 0,43. O desempenho do *Stacking* piorou conforme se aumentou a quantidade de membros no comitê, ocorrendo *overfitting*.

Na Figura 26 é apresentado o gráfico de dispersão entre os dados de produtividade observados (dados oficiais do IBGE) e os dados de produtividade preditos pelo *Stacking* composto de cinco membros MLP e rede neural MLP como algoritmo agregador, para o modelo reduzido municipal. Os pontos no gráfico representam todas as instâncias do banco de dados (treinamento, validação e teste).

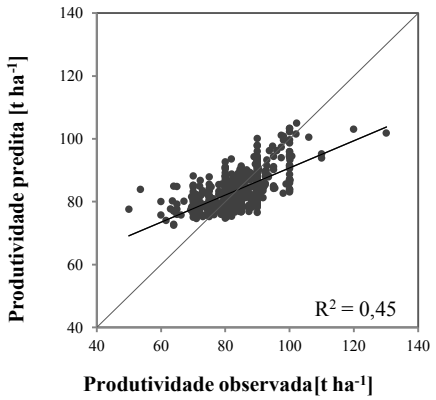


Figura 26: Gráfico de dispersão entre dados observados e dados preditos de produtividade para o *Stacking* de cinco membros no modelo reduzido municipal.

Comparando a Figura 26 com a Figura 25d, nota-se queda na qualidade do ajuste entre variáveis observadas e preditas ao se desconsiderar as variáveis dos últimos três meses. O RMSE no modelo espectral municipal foi de 5,7 t ha^{-1} e R^2 de 0,61, para a parcela de teste. Ao se realizar o mesmo procedimento (seleção de variáveis e predição por *Stacking*), porém desconsiderando as variáveis espectrais que dependiam dos últimos três meses da safra, o RMSE aumentou, de 5,7 para 6,7 t ha^{-1}, e o R^2 diminuiu, de 0,61 para 0,43 na parcela de teste. Nota-se assim a importância das variáveis "*integral*", "*lenght*" e "*righ_derivative*" para a predição da variável "*rend*" usando MLP, especialmente das variáveis "*integral*" e "*lenght*", que aparecem em segundo e terceiro lugar na ordem de importância da Tabela 6. Verifica-se também que a variável "*amplitude*", que fora selecionada no modelo espectral municipal, e que acarretaria em queda de desempenho de 34,7% ao ser removida (etapa 5 da Tabela 4), foi excluída já na segunda etapa do modelo reduzido municipal (etapa 3 da Tabela 12). Isso mostra uma vantagem em se usar o método *wrapper* aplicado neste estudo para a seleção de variáveis, pois o uso do próprio algoritmo preditor na seleção de variáveis possibilitou avaliar importância de cada variável para o desempenho do algoritmo quando combinada com outras variáveis, e não somente a importância de cada variável isoladamente.

4.5. Modelo estadual

Com o objetivo de se comparar os resultados de predição obtidos neste estudo com as estimativas oficiais de produtividade fornecidas pela Conab, que são realizadas somente em escala estadual, foi obtido o modelo estadual.

A Conab estima a produtividade da cana-de-açúcar em três diferentes datas: abril/maio (1º levantamento), agosto/setembro (2º levantamento) e novembro/dezembro (3º levantamento) (Conab, 2015). Foi então calculada a produtividade média entre os 60 municípios a partir dos resultados do modelo espectral municipal, denominado modelo espectral estadual (MEE), e a partir dos dados do modelo reduzido municipal, denominado modelo reduzido estadual (MRE). Obteve-se assim uma produtividade média estadual baseada em dados espectrais obtidos ao longo de toda a safra (MEE) e com três meses de antecedência ao final da safra (MRE).

Foram calculados o RMSE e R^2 para os resultados do MEE, do MRE e para os três levantamentos da Conab, tendo como referência os dados oficiais de produtividade estadual fornecidos pelo IBGE. Os resultados de produtividade estimada, RMSE e R^2 para o MEE, MRE e para os três levantamentos da Conab são mostrados na Tabela 17.

Tabela 17: Resultados da produtividade estimada [t ha^{-1}], RMSE [t ha^{-1}] e R^2 para o MEE, MRE e para os três levantamentos da Conab.

Estimativas	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	RMSE	R^2
Conab – 1º	-*	-	81,2	87,2	89,5	-	86,6	87,5	79,5	73,0	4,0	0,58
Conab - 2º	-	-	83,5	86,1	87,2	84,1	86,9	84,5	72,3	73,1	4,0	0,82
Conab - 3º	-	-	84,4	86,1	86,7	89,0	86,4	82,5	70,5	74,7	4,2	0,92
MRE	82,4	82,5	85,0	82,9	85,7	85,3	85,9	81,2	81,6	-	1,4	0,61
MEE	82,2	82,0	84,9	83,3	86,9	85,9	87,4	82,2	79,8	-	1,3	0,80

* Não disponível.

Observa-se na Tabela 17 que os levantamentos da Conab apresentaram um RMSE em torno de 4 t ha^{-1}. O valor do R^2 foi de 0,58 (abril/maio), 0,82 (agosto/setembro) e 0,92 (novembro/dezembro). O MRE apresentou um RMSE de 1,4 t ha^{-1} e R^2 de 0,61, podendo ser obtido a partir de agosto, o que equivale ao segundo levantamento da Conab. O MEE apresentou um RMSE de 1,3 t ha^{-1} e R^2 de 0,80,

podendo ser obtido a partir de novembro. Verifica-se que os modelos estaduais propostos apresentaram RMSE bastante inferior aos levantamentos da Conab, além de valores de R^2 compatíveis.

4.6. Discussões gerais

A importância das variáveis espectrais para a predição da produtividade da cana-de-açúcar, apontada em estudos como os de Fortes *et al*. (2009), Simões *et al*. (2009) e Fernandes *et al*. (2011), pode ser confirmada nos resultados de seleção de variáveis deste estudo, apresentados nas Tabelas 6, 10 e 13. A variável espectral "*max_value*", que se refere ao valor máximo alcançado pela série temporal do NDVI ao longo da safra, foi a variável que apresentou maior importância dentre as variáveis selecionadas. Esta também foi a variável de maior correlação com a variável alvo (*rend*), que se refere à produtividade média municipal. Os resultados confirmam a adequabilidade dos dados de NDVI do sensor MODIS na predição da produtividade de culturas agrícolas (Ren *et al.*, 2008; Mkhabela *et al*., 2011; Kogan *et al*., 2013; Mulianga *et al*., 2013; Sakamoto *et al*., 2014).

A não seleção de variáveis meteorológicas pode estar associada a diferentes fatores, podendo inclusive haver uma combinação deles. Primeiramente, é importante observar que os dados meteorológicos são provenientes de um modelo de simulação, que possui incertezas.

Os dados de produtividade usados neste estudo se referem à média municipal. No entanto, dentro de um mesmo município ocorrem diferentes variedades genéticas de cana-de-açúcar, que apresentam ciclos de desenvolvimento de diferentes durações, podendo estas ser colhidas com cerca de seis meses (precoce), nove meses (médio), doze meses (tardio) ou após um ano e meio do início da safra (cana de ano e meio). E ciclos fisiológicos de diferentes durações fazem com que ocorrências climáticas de um determinado período influenciem cada variedade de forma diferenciada. Porém, quando se analisa uma média municipal, se torna difícil relacionar efeitos climáticos no desenvolvimento da cultura.

Outro fator que pode ter contribuído para a não seleção de variáveis meteorológicas é o fato de se ter aplicado neste estudo variáveis meteorológicas diretamente ao invés de se construir variáveis agrometeorológicas, que trazem embutidas a relação solo-planta-atmosfera, conforme feito em estudos de produtividade da cana-de-açúcar como em Monteiro e Sentelhas (2014).

Um quarto fator que pode ter influenciado a não seleção das variáveis meteorológicas usadas neste estudo é a existência de correlação entre variáveis meteorológicas e variáveis espectrais. Verifica-se, por exemplo, na matriz de correlação da Figura 24 a existência de correlação positiva entre a variável meteorológica *RAD_max3* (que significa valor máximo de radiação fase final do desenvolvimento da cana) e as variáveis espectrais *lenght* (duração da safra) e *integral* (integral da série temporal de NDVI). Sabe-se que a ocorrência de altos valores de radiação na segunda metade do ciclo de desenvolvimento da cana-de-açúcar pode ser benéfica à produtividade da cultura (Lisson *et al.*, 2005; Park *et al.*, 2005a). Sabe-se também que existe relação entre o Índice de Área Foliar (IAF), que mede a capacidade da planta em absorver radiação solar (Nassif *et al.*, 2012), e o NDVI extraído de imagens de satélite (Simões *et al.*, 2005a). Dessa forma, variáveis espectrais selecionadas podem trazer embutidas "informações meteorológicas".

Bagging de redes neurais MLP não foi adequado para melhorar o desempenho de uma MLP simples na tarefa de predição da produtividade municipal média da cana-de-açúcar. A média aritmética usada como método de agregação das saídas dos membros do comitê fez com que os resultados do *Bagging* se aproximassem do valor médio da produtividade. Este problema não ocorreu com o método de comitê *Stacking* usado neste estudo, uma vez que as saídas dos membros do comitê foram entradas em um novo espaço de busca, no qual, por meio do algoritmo de aprendizagem MLP, foi possível predizer a produtividade de forma mais ajustada com os dados observados, melhorando os resultados de uma MLP simples.

O uso de comitês de modelos obtidos pelo método *Stacking* composto de poucos membros possibilitou aumentar a capacidade de generalização de uma MLP simples. Isso se verifica na diferença existente entre resultado na parcela de treinamento e na parcela de teste. Para o modelo espectral municipal, a diferença entre o RMSE da parcela de treinamento e da parcela de teste para uma MLP simples (*wrapper*) foi de 0,6 t ha^{-1}, e a diferença entre o R^2 para estas duas parcelas do banco de dados foi de 0,08. Após a aplicação do *Stacking* houve, além de melhora no desempenho, diminuição na diferença do RMSE entre as parcelas de treinamento e de teste, que passou a ser de 0,3 t ha^{-1} e na diferença do R^2, que passou a ser de 0,04. O mesmo efeito se verificou para o modelo reduzido municipal. Isso demonstra a maior capacidade do comitê *Stacking*, quando comparado a uma MLP simples, em predizer registros que não participaram da

etapa de treinamento. No entanto, em comitês *Stacking* compostos de muitos membros houve diminuição na capacidade de generalização, ocorrendo *overfitting*.

Os resultados de predição da produtividade da cana-de-açúcar encontrados neste estudo são compatíveis com outros estudos na literatura que usaram diferentes métodos. Na Tabela 18 são apresentados os resultados de RMSE e R^2 para cinco estudos de diferentes autores para predição da produtividade de biomassa da cana-de-açúcar. Cuadra *et al*. (2012) simularam a produtividade média da cana-de-açúcar para quatro mesorregiões (agrupamento de municípios) no estado de São Paulo para um período de estudo de 16 anos usando um modelo biofísico de crescimento da cana-de-açúcar. Greenland (2005) modelou a produtividade anual baseado em quatro variáveis climáticas críticas selecionadas previamente e um modelo de regressão linear (DSSAT/CANEGRO) para Louisiana, EUA, para o período entre 1963 e 2002. Nassif *et al*. (2012) calibraram o modelo DSSAT/CANEGRO para algumas variedades genéticas brasileiras de cana-de-açúcar usando áreas canavieiras localizadas em dois municípios de São Paulo e durante uma safra. Monteiro e Sentelhas (2014) calibraram um modelo agrícola da FAO para estimar a produtividade da cana-de-açúcar para o estado de São Paulo. E Picoli *et al*. (2009) combinaram o modelo agrícola da FAO com dados de NDVI do sensor MODIS para estimar a produtividade da cana-de-açúcar para algumas áreas canavieiras de São Paulo e durante duas safras.

Tabela 18: Resultados de cinco estudos de predição de produtividade de biomassa da cana-de-açúcar.

Estudo	**RMSE** (t ha^{-1})	**R^2**
Cuadra *et al*. (2012)	3,6	-*
Greenland (2005)	5,1	-*
Nassif *et al*. (2012)	30,0	0,51
Monteiro e Sentelhas (2014)	5,0	0,56
Picoli *et al*. (2009)	20,0	0,30

* Não disponível.

Considerando algumas limitações envolvidas neste estudo, como a resolução espacial moderada das imagens MODIS, o uso de uma média municipal para a produtividade da cana-de-açúcar, ausência de informações locais sobre as condições de desenvolvimento da cultura e o fato de que apenas variáveis espectrais foram

selecionadas, os resultados podem ser considerados promissores, pois foram compatíveis com outros estudos que fizeram uso de informações mais detalhadas.

As limitações citadas neste estudo são fontes de incertezas do modelo, uma vez que, devido à escala espacial do estudo - que fornece um valor de produtividade para todo um município ou estado - não é possível obter variáveis detalhadas sobre o sistema produtivo.

No modelo estadual foi possível obter uma estimativa com menor erro que as estimativas feitas pela Conab. Trata-se de um método de baixo custo e de rápida execução, uma vez que as imagens do sensor MODIS podem ser obtidas gratuitamente e com pouco atraso de tempo.

Capítulo 5 – Conclusões e Recomendações

O método proposto se mostrou adequado à predição da produtividade média municipal da cana-de-açúcar nos principais municípios produtores do estado de São Paulo. Variáveis espectrais construídas por meio de métricas extraídas da série temporal do NDVI, obtida a partir do uso conjunto das plataformas Terra e Aqua do sensor MODIS, possibilitaram modelar a produtividade média municipal da cana-de-açúcar. A produtividade foi modelada usando um comitê *Stacking* com redes MLP, apresentando desempenho similar, e em alguns casos melhor, que outros estudos da literatura.

A seleção de variáveis usando um *wrapper* com o algoritmo MLP e seleção sequencial para trás permitiu eliminar variáveis irrelevantes e/ou redundantes do banco de dados inicial, diminuindo consideravelmente o número de variáveis necessárias ao modelo, reduzindo o *overfitting* e melhorando o desempenho da MLP. O uso do próprio algoritmo preditor na seleção de variáveis possibilitou avaliar importância do conjunto de variáveis para o desempenho do algoritmo, e não somente a importância de cada variável isoladamente.

O uso de um comitê de modelos baseado em um *Stacking* composto por membros MLP e algoritmo agregador MLP melhoraram o desempenho de uma MLP simples, aumentando a capacidade de generalização do modelo e se mostrando adequado à predição da produtividade média municipal da cana-de-açúcar. Essa melhoria ocorreu com comitês *Stacking* compostos por poucos membros. Comitês *Stacking* compostos por muitos membros não apresentaram melhoria de desempenho, ocorrendo *overfitting*. *Bagging* não foi adequado para melhorar o desempenho de uma MLP simples na tarefa de predição da produtividade média municipal da cana-de-açúcar.

As variáveis meteorológicas construídas neste estudo não foram adequadas ao uso conjunto com variáveis espectrais na tarefa de predição da produtividade média municipal da cana-de-açúcar. É necessário que se avalie o uso de variáveis agrometeorológicas em conjunto com as variáveis espectrais construídas neste estudo. Variáveis agrometeorológicas podem apresentar maior relevância ao modelo e contribuir para a melhora do desempenho.

A cana-de-açúcar é uma cultura complexa, que apresenta características que dificultam a modelagem, como diferentes variedades genéticas com diferentes durações de ciclo de desenvolvimento e diversos cortes ao longo do ciclo de reforma, que apresentam diferentes potenciais produtivos. Portanto, considerando as dificuldades de modelagem desta cultura e as limitações presentes neste estudo, o método proposto apresentou resultados satisfatórios para a predição da produtividade da cana-de-açúcar em escala municipal e estadual. É interessante que se avalie o desempenho do método proposto na predição da produtividade da cana-de-açúcar em escala local, ou seja, em nível de talhões, o que possibilitaria considerar dados importantes, como variedade genética, corte, aplicação de insumos, além de dados de produtividade por talhão, diminuindo assim as incertezas. Isso possibilitaria também obter uma série temporal do NDVI para cada variedade genética, o que permitiria representar de forma mais precisa o ciclo de desenvolvimento de cada variedade e assim obter variáveis climáticas mais ajustadas ao ciclo de desenvolvimento.

O uso de comitês de modelos proposto neste estudo se adequa perfeitamente a ambientes de processamento intensivo de dados, onde milhares de modelos podem ser treinados simultaneamente. Além disso, estes ambientes fornecem ferramentas específicas para lidar com fluxo intensivo de dados online (*data streams*). Atualmente, a Conab coleta informações subjetivas sobre o sistema produtivo em diversas localidades produtoras de cana-de-açúcar. É possível pensar em um modelo objetivo de predição de produtividade em que colaboradores locais e sensores instalados nas áreas produtoras pudessem fornecer dados sobre o ambiente produtivo que alimentassem modelos locais de predição de produtividade. Isso permitiria o uso de diferentes algoritmos de predição simultâneos, como redes neurais, máquinas de vetores de suporte e árvores de regressão. E, posteriormente, os resultados dos modelos locais simples pudessem ser entrada em um modelo agregador para a predição final da produtividade.

Outra possibilidade para trabalhos futuros de predição da produtividade municipal da cana-de-açúcar por meio de comitês de modelos é o uso de membros no comitê que fossem especializados em grupos climáticos. Para isso, a partir dos dados meteorológicos, poderia ser feito um agrupamento sobre diferentes grupos climáticos existentes. Então, para cada grupo climático, haveria um membro no comitê especializado para a respectiva condição climática. Ao final, os membros especializados forneceriam resultados que seriam entrada em um segundo modelo agregador. Isso

poderia trazer ganhos em relação ao uso de variáveis meteorológicas e/ou agrometeorológicas em modelos de predição da produtividade.

Porém, considerando os resultados obtidos neste estudo, já é possível se pensar em melhoria do sistema atual de estimativa da safra da cana-de-açúcar. O modelo estadual proposto neste estudo apresentou menor erro de predição que o método de estimativa de safra usado pela Conab. Trata-se de um método objetivo e de baixo custo, baseado em dados espectrais gratuitos e em escala municipal e que pode fornecer estimativas de produtividade mais acuradas e detalhadas que o método oficial de estimativa da safra de cana-de-açúcar realizado pela Conab.

Referências Bibliográficas

Almeida, T. I. R.; De Souza Filho, C. R.; Rossetto R. ASTER and Landsat ETM+ images applied to sugarcane yield forecast. International Journal of Remote Sensing, v.27 n.19, p.4057-4069, 2006.

Alvarez, R. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. European Journal of Agronomy, v.30, p.70–77, 2009.

Arel, I.; Rose, D. C.; Karnowski, T. P. Deep Machine Learning - A New Frontier in Artificial Intelligence Research. IEEE Computational Intelligence Magazine, 2010. doi:10.1109/MCI.2010.938364.

Banfield, R. E.; Hall, L. O.; Bowyer, K. W.; Kegelmeyer, W. P. A comparison of decision tree ensemble creation techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence, v.29, n.1, p.173-180, 2007.

BNDES. Bioetanol de cana-de-açúcar : energia para o desenvolvimento sustentável / organização BNDES e CGEE. – Rio de Janeiro : BNDES, 2008. 316 p.

Bolton, D. K.; Friedl, M. A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology, v.173, p.74-84, 2013.

Breiman, L. Bagging Predictors. Machine Learning, v.24, n.3, p.123-140, 1996.

Breiman, L. Random forests. Machine Learning, v.45, n.1, p.5-32, 2001.

Brown, G. *et al.* Diversity creation methods: a survey and categorisation. Information Fusion, v.6, p.5-20, 2005.

CEPAGRI. Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura. Disponível em: <http://www.cepagri.unicamp.br/outras-informacoes/clima-dos-municipios-paulistas.html>. Acesso em: 25 nov. 2014.

Conab. Levantamentos sobre o acompanhamento da safra brasileira de cana-de-açúcar. Disponível em: <http://www.conab.gov.br/conteudos.php?a=1253&t=2>. Acesso em: 10 jun. 2015.

Cuadra, S. V.; Costa, M. H.; Kucharik, C. J.; Da Rocha, H. R.; Tatsch, J. D.; Inman-Bamber, N. G.; Leite, C. C.; Cabral, O.M.R. A biophysical model of sugarcane growth. Global Change Biology Bioenergy v.4, p.36-48, 2012.

Dietterich, T.G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning v.40, n.2, p.139–157, 2000.

Duveiller, G.; López-Lozano, R.; Baruth, B. Enhanced Processing of 1-km Spatial Resolution fAPAR Time Series for Sugarcane Yield Forecasting and Monitoring. Remote Sensing, v.5, p.1091-1116, 2013.

ECMWF. European Centre for Medium-Range Weather Forecasts. Disponível em: <http://www.ecmwf.int/>. Acesso em: 07 dez. 2012.

EPE. Empresa de Pesquisa Energética. Balanço Energético Nacional 2013: Ano base 2012. Disponível em: <https://ben.epe.gov.br/downloads/Relatorio_Final_BEN_2015.pdf>. Acesso em: 10 jul. 2015.

Espinosa, C. H.; Sospedra, J. T.; Redondo, M. F. Improving the Combination Module with a Neural Network. In: Huang, D. S.; Li, K.; Irwin G.W. (Org.): ICIC 2006, LNCS 4113, 2006. p. 146-155.

Esquerdo, J. C. D. M.; Zullo Júnior, J.; Antunes, J. F. G. Use of NDVI/AVHRR time-series profiles for soybean crop monitoring in Brasil. International Journal of Remote Sensing, v.32, p.3711-3727, 2011.

Everingham, Y. L.; Smyth, C. W.; Inman-Bamber, N. G. Ensemble data mining approaches to forecast regional sugarcane crop production. Agricultural and Forest Meteorology, v.149, p.689 - 696, 2009.

FAOSTAT. Food and Agricultural Organization of the United Nations. Disponível em <http://faostat3.fao.org/home/index.html#VISUALIZE>. Acesso em: 10 ago. 2015.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. PiatetskyShapiro, P. Smyth, and R. Uthurusamy, 1–30. Menlo Park, Calif.: AAAI Press. 1996.

Fernandes, J. L.; Rocha, J. V.; Lamparelli, R. A. C. Sugarcane yield estimates using time series analysis of spot vegetation images. Scientia Agricola, v.68, n.2, p.139-146, 2011.

Ferraro, D. O.; Rivero, D. E.; Ghersa, C. M. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. Field Crops Research, v.112, p.149–157, 2009.

Fortes, C. Discriminação varietal e estimativa de produtividade agroindustrial de cana-de-açúcar pelo sensor orbital ETM+/Landsat **7**. 131p. Dissertação (Mestrado em Solos e Nutrição de Plantas) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba. 2003.

Fortes, C.; Demattê, A. M. Discrimination of sugarcane varieties using Landsat 7 ETM+ spectral data. International Journal of Remote Sensing, v. 27, n. 7, p. 1395-1412, 2006.

Fortes, C.; Demattê, J. A. M.; Genú, A. M. Estimativa de produtividade agroindustrial de cana-de-açúcar por dados espectrais orbitais ETM+/LANDSAT 7. Ambiência, v.5 n.3 p.489 – 504, 2009.

Freitas, R. M. *et al.* Virtual laboratory of remote sensing time series: visualization of MODIS EVI2 data set over South America. Journal of Computational Interdisciplinary Sciences, v.2, n.1, p.57-68, 2011.

Freund, Y.; Schapire, R. E. Experiments with a New Boosting Algorithm. In: Proceedings… ICML 1996, pp. 148-156.

Gasca, E.; Sanchez, J. S.; Alonso, R. Eliminating redundancy and irrelevance using a new MLP-based feature selection method. Pattern Recognition, v.39, p.313–315, 2006.

Galvão, L. S.; Formaggio, A. R.; Tisot D. A. The influence of spectral resolution on discriminating Brazilian sugarcane varieties. International Journal of Remote Sensing, v.27, n.3-4, p.769-777, 2006.

Greenland, D. Climate variability and sugarcane yield in Louisiana. Journal of Applied Meteorology, v.44, p.1655–1666, 2005.

Guan, S.; Liu, J.; Qi, Y. An incremental approach to contribution-based feature selection. Journal of Intelligence Systems, v.13, n.1, 2004.

Han, J.; Kamber, M. Data Mining: Concepts and Techniques. 2. ed. Estados Unidos da América: Elsevier, 2006. 770p.

IBGE. Pesquisas Agropecuárias. Série Relatórios Metodológicos. Disponível em: <http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/PesquisasAgropecuarias2002.pdf>. Acesso em: 10 jun. 2015.

Inman-Bamber, N.G. A growth model for sugarcane based on a simple carbon balance and the CERES-Maize water balance. South African Journal of Plant and Soil, v.8, n.2, p.93-99, 1991.

Inman-Bamber, N.G. e Smith, D.M. Water relations in sugarcane and response to water deficits. Field Crops Research, v.92, p.185-202, 2005.

INMET. Instituto Nacional de Meteorologia. Banco de Dados Meteorológicos para Ensino e Pesquisa. Disponível em: <http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>. Acesso em: 04 dez. 2012.

INPE-CPTEC. Instituto Nacional de Pesquisas Espaciais - Centro de Previsão de Tempo e Estudos Climáticos: Dados Observacionais. Disponível em: <http://bancodedados.cptec.inpe.br/>. Acesso em: 05 dez. 2012.

Jame, Y. W.; Cutforth, H. W. Crop growth models for decision support systems. Canadian Journal of Plant Science, n.76, p.9-19, 1996.

Jones, M.; Singels, A. DSSAT v4.5- CANEGRO Sugarcane Plant Module: Scientific documentation. South African Sugarcane Research Inst. Mount Edgecombe, South Africa, 58p, 2008.

Jönsson, P. e Eklundh, L. TIMESAT - a program for analysing time-series of satellite sensor data, Computers and Geosciences v.30, p.833-845, 2004.

JRC. Joint Research Centre. Data Distribution. Disponível em: <http://mars.jrc.ec.europa.eu/mars/About-us/FOODSEC/Data-Distribution>. Acesso em: 14 mar. 2014.

Kabir, M. M.; Islam, M. M.; Murase, K. A new wrapper feature selection approach using neural network. Neurocomputing, v.73, p. 3273–3283, 2010.

Kaul, M.; Hill, R. L.; Walthall, C. Artificial neural networks for corn and soybean yield prediction. Agricultural Systems, v.85. p.1-18, 2005.

Keating, B. A. *et al.* Modelling sugarcane production systems I. Development and performance of the sugarcane module. Field Crops Research, v.61, p.253-271, 1999.

Kim, Y. Boosting and measuring the performance of ensembles for a successful database marketing. Expert Systems with Applications, v.36 p. 2161–2176, 2009.

Kim, M. J.; Kang, D. K. Ensemble with neural networks for bankruptcy prediction. Expert Systems with Applications, v.37, p. 3373–3379, 2010.

King, M. A.; Abrahams, A. S.; Ragsdale, C. T. Ensemble methods for advanced skier days prediction. Expert Systems with Applications. 2013. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2013.08.002>. Acesso em: 15 set. 2013.

Kogan, F. *et al*. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models. International Journal of Applied Earth Observation and Geoinformation, v.23, p.192–203, 2013.

Kotsiantis, S. B.; Pintelas, P. E. Combining Bagging and Boosting. International Journal of Computacional Intelligence, v.1, n.4, 2004.

Lisson, S. N.; Inman-Bamber, N. G.; Robertson, M. J.; Keating, B. A. The historical and future contribution of crop physiology and modelling research to sugarcane production systems. Field Crops Research, v.92, p.321-335, 2005.

Liu, D. L.; Bull, T. A. Simulation of biomass and sugar accumulation in sugarcane using a process-based model. Ecological Modelling, v.144, p.181-211, 2001.

Liu, G.; Yang, X.; Li, M. An Artificial Neural Network Model for Crop Yield Responding to Soil Parameters. LNCS, Advances in Neural Network, v.3498, p.1017–1021, 2005.

Lobato, D. H.; Munoz, G. M.; Suarez, A. Empirical analysis and evaluation of approximate techniques for pruning regression bagging ensembles. Neurocomputing, v.74, p. 2250–2264, 2011.

LP-DAAC. Land Processes Distributed Active Archive Center. Disponível em: <https://lpdaac.usgs.gov/>. Acesso em: 03 set. 2015.

Maldonado, S.; Weber, R. A wrapper method for feature selection using Support Vector Machines, Information Sciences, v.179, n.13, p. 2208–2217, 2009.

Marin, F. R. *et al.* Parameterization and Evaluation of Predictions of DSSAT/ CANEGRO for Brazilian Sugarcane. Agronomy Journal, v.103, n.2, p.304-315, 2011.

Marin, F. R.; Carvalho, G. L. Spatio-temporal variability of sugarcane yield efficiency in the state of São Paulo, Brazil. Pesquisa Agropecuária Brasileira, v.47, n.2, p.149-156, 2012.

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, v.51, n.4, p.1897-1913, 2013.

METAMP. Methodology of the MARS Crop Yield Forecasting Systems. Disponível em: <http://mars.jrc.ec.europa.eu/mars/Projects/Methodology-of-the-MARS-Crop-Yeld-Forecasting-Systems-METAMP>. Acesso em: 27 out. 2011.

Miao, Y.; Mulla, D. J.; Robert, P. C. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. Precision Agriculture, v.7, p.117–135, 2006.

Mkhabela, M. S.; Bullock, P.; Raj, S.; Wang, S.; Yang, Y. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agricultural and Forest Meteorology, v.151, p.385–393, 2011.

Monteiro, L. A. e Sentelhas, P. C. Potential and Actual Sugarcane Yields in Southern Brazil as a Function of Climate Conditions and Crop Management. Sugar Tech, v.16, n.3, p.264-276, 2014.

Moraes, R. A.; Rocha, J. V.; Rolim, G. S.; Lamparelli, R. A. C.; Martins, M. M. Avaliação dos dados decendiais de precipitação e temperatura máxima e mínima do ar simulados pelo modelo ECMWF para o estado de São Paulo. Irriga, v.17, n.3, p.397 - 407, 2012.

Moutinho, A. M. Otimização de Sistemas de Detecção de Padrões em Imagens. Tese (Doutorado). Rio de Janeiro: UFRJ/COPPE - Programa de Engenharia Civil, 2011.

Mulianga, B; Bégué, A.; Simões, M.; Todoroff, P. Forecasting Regional Sugarcane Yield Based on Time Integral and Spatial Aggregation of MODIS NDVI. Remote Sensing, v.5, p.2184-2199, 2013.

Nassif, D. S. P. *et al.* Parametrização e avaliação do modelo DSSAT/Canegro para variedades brasileiras de cana-de-açúcar. Pesquisa Agropecuária Brasileira, v.47, n.3, p.311-318, 2012.

Oliveira, R. A.; dos Santos, R. S.; Ribeiro, A; Zolnier, S; Barbosa, M. H. P. Estimativa da produtividade da cana-de-açúcar para as principais regiões produtoras de Minas Gerais usando-se o método ZAE. Revista Brasileira de Engenharia Agrícola e Ambiental, v.16, n.5, p.549–557, 2012.

Opitz, D W.; Maclin, R. Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Research, v.11, p. 169–198, 1999.

Papageorgiou, E. I.; Markinos, A.; Gemptos, T. Application of fuzzy cognitive maps for cotton yield management in precision farming. Expert Systems with Applications, v.36, p.12399–12413, 2009.

Park, S. E.; Robertson, M.; Inman-Bamber, N. G. Decline in the growth of a sugarcane crop with age under high input conditions. Field Crops Research, v.92, p.305-320, 2005a.

Park, S. J.; Hwang, C. S.; Vlek, P. L. G. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agricultural Systems, v.85, p.59–81, 2005b.

Partalas, I; Tsoumakas, G.; Vlahavas, I. Pruning an ensemble of classifiers via reinforcement learning. Neurocomputing, v.72, p. 1900–1909, 2009.

Pellegrino, G.Q. Utilização de dados espectrais do satélite NOAA14/AVHRR como fonte de dados para modelos matemáticos de estimativa da fitomassa da cana-de-açucar. 116p. Tese (Doutorado) - Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas - SP. 2001.

Peng, H.; Long, F.; Ding, C. Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research, v.3, p. 1371–1382, 2003.

Picoli, M. C. A. *et al.* Índice de Vegetação do Sensor MODIS na Estimativa da Produtividade Agrícola da Cana-de-açúcar. Bragantia, v.68, n.3, p.789-795, 2009.

Quinlan, J. R. Bagging, boosting, and C4.5, In: Proceedings of the 13th National Conference on Artificial Intelligence, Cambridge, MA, 1996, pp. 725–730.

Ramme, F. L. P.; Lamparelli, R. A. C.; Rocha, J. V. Perfis temporais NDVI MODIS, na cana-soca, de maturação tardia. Engenharia Agrícola, v.30, n.3, p.480-494, 2010.

Rao, P. V. K; Rao, V. V.; Venkataratnam, L. Remote Sensing: A Technology for Assessment of Sugarcane Crop Acreage and Yield. Sugar Tech, v.4, p.97 – 101, 2002.

Reid, S.; Grudic, G. Regularized linear models in stacked generalization, In: Benediktsson, J. A.; Kittler, J.; Roli, F. (Org.), Multiple Classifier Systems, Lecture Notes in Computer Science, v. 5519, Springer, 2009, p. 112–121.

Ren, J. *et al.* Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. International Journal of Applied Earth Observation and Geoinformation, v.10, p.403-413, 2008.

Rodrigues, J. D. Fisiologia da cana-de-açúcar, Unesp, Botucatu, 1995. Disponível em <http://www.malavolta.com.br/pdf/1071.pdf> Acesso em: 01 mai. 2013.

Romero, E.; Sopena, J. M. Performing feature selection with multilayer perceptrons. IEEE Transactions on Neural Networks, v.19, n.3, p. 431–441, 2008.

Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the great plains with ERTS. In: Earth Resources Technology Satellite-1 Symposium, 3., Washington, 1973. Proceedings… Whashington: NASA, 1974, v.1, p.309-317, 1973.

Rudorff, B. F. T.; Aguiar, D. A.; Silva, W. F.; Sugawara, L. M.; Adami, M.; Moreira, M. A. Studies on the Rapid Expansion of Sugarcane for Ethanol Production in São Paulo State (Brazil) Using Landsat Data. Remote Sensing, v.2, n.4, p.1057-1076, 2010.

Sakamoto, T.; Gitelson, A. A.; Arkebauer, T. J. Near real-time prediction of U.S. corn yields based on time-series MODIS data Remote Sensing of Environment, v.147, p.219–231, 2014.

Savin, I. Yu. *et al.* Prediction of Crop Yields with the Use of Neural Networks. Russian Agricultural Sciences, v.33, n.6, p.361–363, 2007.

Scarpari, M. S.; Beauclair, E. G. F. Sugarcane maturity estimation through edaphic-climatic parameters. Scientia Agricola, v.61, n.5, p.486-491, 2004.

Scarpari, M. S.; Beauclair, E. G. F. Physiological model to estimate the maturity of sugarcane. Scientia Agricola, v.66, n.5, p.622-628, 2009.

SIDRA. Sistema IBGE de recuperação de dados automática. Disponível em: < http://www.sidra.ibge.gov.br/>. Acesso em: 15 ago. 2015.

Simões, M. S.; Rocha, J. V.; Lamparelli, R. A. C. Spectral variables, growth analysis and yield of sugarcane. Scientia Agricola, v.62, n.3, p.199-207, 2005a.

Simões, M. S.; Rocha, J. V.; Lamparelli, R. A. C. Growth indices and productivity in sugarcane. Scientia Agricola, v.62, n.1, p.23-30, 2005b.

Simões, M. S.; Rocha, J. V.; Lamparelli, R. A. C. Orbital spectral variables, growth analysis and sugarcane yield. Scientia Agricola, v.66, n.4, p.451-461, 2009.

Singels, A. *et al.* The effect of crop start date, crop class and cultivar on sugarcane canopy development and radiation interception. Field Crops Research, v.92, p.249-260, 2005.

Souza, Z. M. *et al.* Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. Ciência Rural, v.40, n.4, p.840-847, 2010.

USDA. Understanding USDA Crop Forecasts. Disponível em: <http://www.nass.usda.gov/>. Acesso em: 26 out. 2011.

Van den Berg, M.; Burrough, P. A.; Driessen, P. M. Uncertainties in the appraisal of water availability and consequences for simulated sugarcane yield potentials in São Paulo State, Brazil. Agriculture, Ecosystems and Environment, v.81, p.43-55, 2000.

Vicente, L. E.; Gomes, D.; Victoria, D. C.; Garçon, E. A. M.; Bolfe, E. L.; Andrade, R. G.; da Silva, G. B. S. NDVI temporal series from the SPOT Vegetation sensor and

SAM algorithm applied to sugarcane mapping. Pesquisa Agropecuária Brasileira, v.47, n.9, p.1337-1345, 2012.

Wardlow, B.D.; Kastens, J.H.; Egbert, S.L. Using USDA Crop Progress Data and MODIS Time-Series NDVI for Regional-Scale Evaluation of Greenup Onset Date. Photogrammetric Engineering and Remote Sensing, v.72, n.11, p.1225-1234, 2006.

Wiedenfeld, R. P. Water stress during different sugarcane growth periods on yield and response to N fertilization. Agricultural Water Management, v.43, p.173-182, 2000.

Wolpert, D. H. Stacked generalization, Neural Networks, v.5, n.2, p.241–259, 1992.

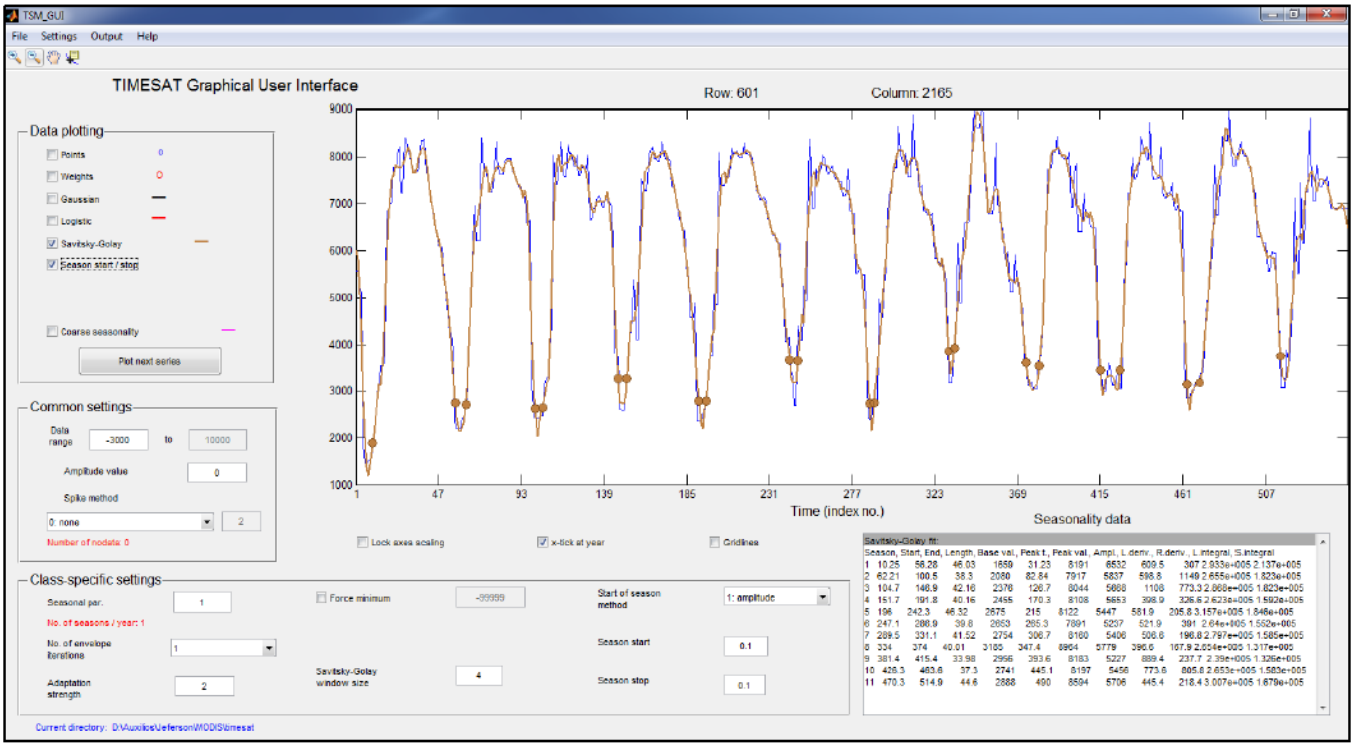
Xavier, A. C.; Rudorff, B. F. T.; Shimabukuro, Y. E.; Berka, L. M. S.; Moreira, M. A. Multi-temporal analysis of MODIS data to classify sugarcane crop. International Journal of Remote Sensing, v.27, n.4, p.755-768, 2006.

Zhang, B. *et al.* Predictive modelling of hill-pasture productivity: integration of a decision tree and a geographical information system. Agricultural Systems, v.87, p.1–17, 2006.

ANEXOS

Anexo 1

Graphical User Interface (GUI) do programa TIMESAT.



Anexo 2

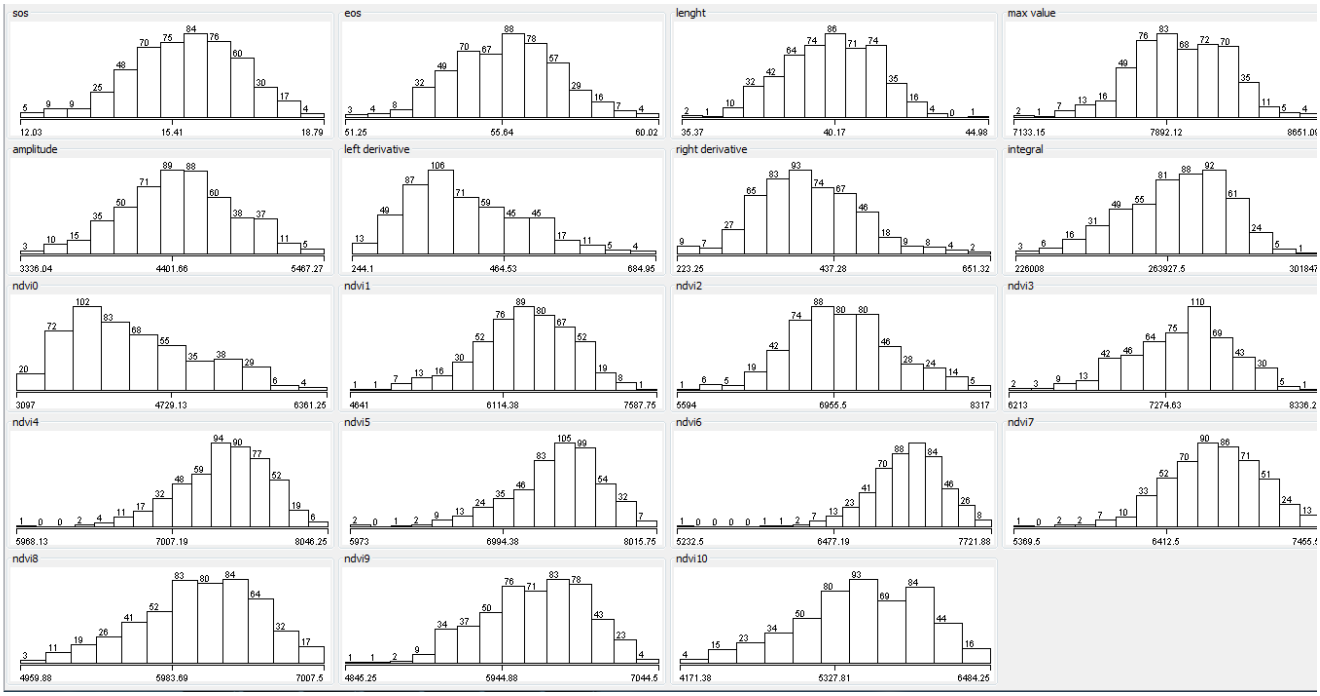
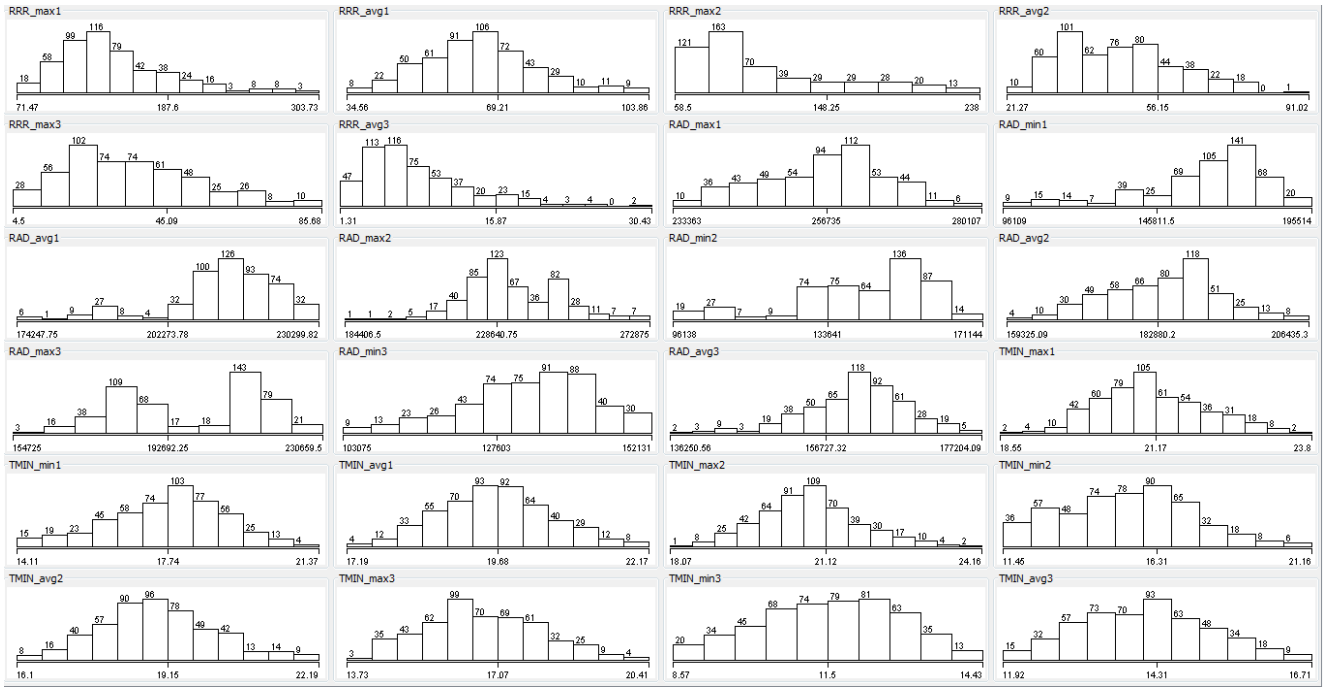
Estatísticas básicas das variáveis espectrais e meteorológicas construídas neste estudo e usadas como entrada nos modelos propostos.

Variáveis	Mínimo	Máximo	Média	Mediana	Desvio padrão
RRR_max1	71.5	303.7	146.0	136.2	42.2
RRR_avg1	34.6	103.9	65.2	64.4	13.0
RRR_max2	58.5	238.0	111.2	94.8	43.6
RRR_avg2	21.3	91.0	47.1	46.0	13.1
RRR_max3	4.5	85.7	35.7	33.6	17.5
RRR_avg3	1.3	30.4	8.5	7.2	4.9
RAD_max1	233363.0	280107.0	256305.0	257642.5	9348.6
RAD_min1	96109.0	195514.0	161265.6	165735.8	19997.2
RAD_avg1	174247.8	230299.8	212986.6	214184.8	10394.5
RAD_max2	184406.5	272875.0	233146.0	230160.0	13224.0
RAD_min2	96138.0	171144.0	142250.1	147010.0	16423.2
RAD_avg2	159325.1	206435.3	183775.7	184436.8	8721.6
RAD_max3	154725.0	230659.5	198566.4	203174.5	17512.3
RAD_min3	103075.0	152131.0	132576.0	134009.0	10454.7
RAD_avg3	136250.6	177204.1	160716.1	161002.0	6614.0
TMIN_max1	18.6	23.8	21.1	21.0	0.9
TMIN_min1	14.1	21.4	17.7	17.9	1.4
TMIN_avg1	17.2	22.2	19.6	19.6	0.9
TMIN_max2	18.1	24.2	20.8	20.8	1.0
TMIN_min2	11.5	21.2	15.4	15.4	2.0
TMIN_avg2	16.1	22.2	18.9	18.9	1.2
TMIN_max3	13.7	20.4	16.7	16.7	1.3
TMIN_min3	8.6	14.4	11.5	11.7	1.3
TMIN_avg3	11.9	16.7	14.1	14.1	1.0
sos	12.0	18.8	15.7	15.8	1.2
eos	51.2	60.0	55.7	55.8	1.5
lenght	35.4	45.0	40.0	40.1	1.5
max value	7133.2	8651.1	7962.0	7953.7	239.9
amplitude	3336.0	5467.3	4453.6	4455.0	384.5
left derivative	244.1	685.0	408.0	391.7	83.5
right derivative	223.2	651.3	403.2	396.6	70.8
integral	226008.0	301847.0	266402.9	267741.0	12903.2
ndvi0	3097.0	6361.3	4330.8	4218.5	680.2
ndvi1	4641.0	7587.8	6326.2	6337.4	464.9
ndvi2	5594.0	8317.0	7030.2	6992.9	460.8
ndvi3	6213.0	8336.3	7385.0	7429.8	342.2
ndvi4	5968.1	8046.3	7355.5	7380.0	299.1
ndvi5	5973.0	8015.8	7351.5	7400.3	306.5
ndvi6	5232.5	7721.9	7054.8	7082.0	293.9
ndvi7	5369.5	7455.5	6738.6	6746.3	316.1
ndvi8	4959.9	7007.5	6167.2	6191.0	403.6
ndvi9	4845.3	7044.5	6172.5	6192.0	368.8
ndvi10	4171.4	6484.3	5522.5	5557.1	462.3

Anexo 3

Histogramas das variáveis construídas

1) Variáveis espectrais:



2) Variáveis meteorológicas: