



FELIPE FERREIRA BOCCA

**PRODUTIVIDADE DE CANA-DE-AÇÚCAR: CARACTERIZAÇÃO  
DOS CONTEXTOS DE DECISÃO E UTILIZAÇÃO DE TÉCNICAS DE  
MINERAÇÃO DE DADOS PARA MODELAGEM**

CAMPINAS

2014





**UNIVERSIDADE ESTADUAL DE CAMPINAS**  
**Faculdade de Engenharia Agrícola**

FELIPE FERREIRA BOCCA

**PRODUTIVIDADE DE CANA-DE-AÇÚCAR: CARACTERIZAÇÃO  
DOS CONTEXOS DE DECISÃO E UTILIZAÇÃO DE TÉCNICAS DE  
MINERAÇÃO DE DADOS PARA MODELAGEM**

Dissertação apresentada à Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas como parte dos requisitos exigidos para obtenção do título de Mestre em Engenharia Agrícola, na área de Planejamento e Desenvolvimento Rural Sustentável.

Orientador: Prof. Dr. Luiz Henrique Antunes Rodrigues

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL  
DA DISSERTAÇÃO DEFENDIDA PELO ALUNO FELIPE  
FERREIRA BOCCA E ORIENTADO PELO PROF. DR.  
LUIZ HENRIQUE ANTUNES RODRIGUES.

Assinatura do Orientador

CAMPINAS

2014

Ficha catalográfica  
Universidade Estadual de Campinas  
Biblioteca da Área de Engenharia e Arquitetura  
Rose Meire da Silva - CRB 8/5974

B63p      Bocca, Felipe Ferreira, 1988-  
Produtividade de cana-de-açúcar : caracterização dos contextos de decisão e utilização de técnicas de mineração de dados para modelagem / Felipe Ferreira Bocca. – Campinas, SP : [s.n.], 2014.

Orientador: Luiz Henrique Antunes Rodrigues.  
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mineração de dados (Computação). 2. Cana-de-açúcar. 3. Modelagem. I. Rodrigues, Luiz Henrique Antunes, 1959-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

Informações para Biblioteca Digital

**Título em outro idioma:** Sugarcane yield : characteristics of decision contexts and data mining techniques application for modeling

**Palavras-chave em inglês:**

Data mining

Sugarcane

Modeling

**Área de concentração:** Planejamento e Desenvolvimento Rural Sustentável

**Titulação:** Mestre em Engenharia Agrícola

**Banca examinadora:**

Luiz Henrique Antunes Rodrigues [Orientador]

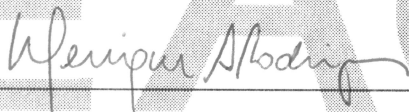
Marcelo Theophilo Folhes

Carlos Alberto Alves Meira

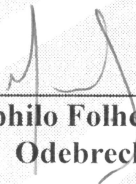
**Data de defesa:** 18-02-2014

**Programa de Pós-Graduação:** Engenharia Agrícola

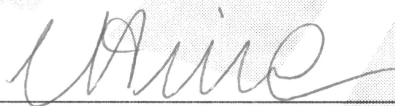
Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Felipe Ferreira Bocca**, aprovada pela Comissão Julgadora em 18 de fevereiro de 2014, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.



**Prof. Dr. Luiz Henrique Antunes Rodrigues – Presidente e Orientador**  
Feagri/Unicamp



**Dr. Marcelo Theophilo Folhes – Membro Titular**  
Odebrecht



**Dr. Carlos Alberto Alves Meira – Membro Titular**  
Embrapa/CNPq

Faculdade de  
**Engenharia Agrícola**  
Unicamp



## RESUMO

A tomada de decisão e o planejamento de uma usina de cana-de-açúcar têm como principal variável a produtividade dos cultivos, que em conjunto com a área permite estimar a produção. A cana-de-açúcar, uma cultura semi-perene, nas condições brasileiras, possui um ciclo inicial que pode variar de 12 a 18 meses e, após a primeira colheita, é colhida anualmente até que o decréscimo de produtividade leve ao replantio da área. Considerando o tamanho das áreas de cultivo, e o horizonte temporal, projeções de produtividade são fornecidas em diferentes contextos de decisão para cultivos que se encontram em diferentes momentos do ciclo de crescimento. Foi conduzida uma pesquisa exploratória junto a uma usina com intuito de contextualizar as principais decisões que são influenciadas pela perspectiva de produtividade futura, bem como a forma que essas previsões afetam o planejamento. Tomando por base o resultado de entrevistas semiestruturadas e acompanhamento de atividades, foi possível identificar decisões chave e suas características, que foram relacionadas a soluções propostas pela comunidade científica e enquadradas dentro de uma proposta de framework para tomada de decisão e planejamento. Entre as decisões, chamou atenção as que são tomadas nos elos iniciais da cadeia de valor, que terão efeitos em todos os processos posteriores e que são tomadas na maior situação de incerteza, sendo consideradas pontos críticos no planejamento. No framework, baseado no uso de modelos empíricos de produtividade, é possível explorar o potencial das informações climáticas para projeção da produtividade e também explorar o potencial dos dados acumulados pelo setor. Para tal, foram desenvolvidos modelos empíricos de produtividade utilizando diferentes técnicas de mineração de dados. Os modelos de produtividade possuíam como atributos preditores os dados referentes aos talhões e seu manejo, em conjunto com os dados do clima ocorrido. Foi possível reduzir a magnitude de erro para menos da metade do encontrado em uma abordagem anterior. Entre as técnicas utilizadas, a SVM e a *Random Forest* obtiveram os melhores desempenhos, embora o modelo utilizando SVM tenha utilizado significativamente menos atributos. A estratégia de modelagem baseada em dados permitiu a criação de modelos específicos para o contexto produtivo da própria unidade, na escala da menor unidade de gestão, os talhões. Os modelos de produtividade criados possuem potencial para projeção de produtividade se utilizados em conjunto com projeções de clima.

## **ABSTRACT**

Decision making and planning of sugarcane production have as main variable the crop yield, which in conjunction with the field area allows us to estimate production. Sugarcane, a semi-perennial crop, in Brazilian conditions, has an initial cycle that varies from 12 to 18 months and after the first harvest, is harvested annually until yield reduction lead to replanting the area. Considering the size of cultivated areas, and the time horizon, yield projections are provided in different contexts of decision for crops that are in different stages of the growth cycle. An exploratory study was conducted within a sugarcane mill to contextualize the main decisions that are influenced by the perspective of future yield, as well as how those predictions affect planning. Based on the result of semi-structured interviews and activities follow-up, it was possible to identify key decisions and their characteristics, which were related to the solutions proposed by the scientific community and framed within a proposed framework for decision making and planning. Decisions made in the first echelons of the value chain demanded early predictions and have effects in the whole value chain, being considered a critical point for planning. In the framework, based on the use of empirical models of yield, it is possible to exploit the potential of climate information to forecast yield and also explore the potential of data accumulated by the sector. Empirical yield models were developed using different data mining techniques. The models used and data from the blocks and their management, coupled with the climatic data as predictive variables. Error magnitude was reduced by half from a previous approach. Among the techniques used, SVM and Random Forest got the best performance, although the SVM model has significantly fewer attributes. The modeling strategy based on data enabled the creation of specific models for the production context of the mill, on the scale of the smallest management unit. The yield models created have potential for yield forecast if used in conjunction with weather forecasts.



## SUMÁRIO

1	INTRODUÇÃO.....	1
2	REVISÃO BIBLIOGRÁFICA.....	4
2.1	Produtividade de cana-de-açúcar e sua modelagem .....	4
2.2	Mineração de dados .....	8
2.3	Avaliação de modelos de regressão .....	14
3	CONTEXTUALIZAÇÃO DA TOMADA DE DECISÃO .....	18
3.1	Resumo .....	18
3.2	Introdução .....	18
3.3	Metodologia .....	26
3.4	A realização de projeções e estimativas na usina .....	28
3.5	A tomada de decisão e o planejamento.....	31
3.6	Proposta de Framework .....	36
3.7	Conclusões .....	41
3.8	Recomendações e trabalhos futuros.....	42
3.9	Referências.....	43
4	TÉCNICAS DE MINERAÇÃO DE DADOS APLICADAS À MODELAGEM DA PRODUTIVIDADE DE CANA-DE-AÇÚCAR .....	46
4.1	Resumo .....	46
4.2	Introdução .....	47
4.3	Material e métodos.....	53
4.4	Resultados e Discussão .....	64
4.5	Conclusion .....	72
	Referências Bibliográficas.....	73
5	CONSIDERAÇÕES FINAIS .....	77
	REFERÊNCIAS .....	79



## **AGRADECIMENTOS**

Aos meus pais, Marcos e Rositale, pelo apoio, não só durante o mestrado, e por sua influência na pessoa que me tornei. O mestrado é mais uma entre as coisas que hoje se apoiam na base que vocês construíram. Aos demais familiares, estendo esse agradecimento, pois se meus pais construíram a base, foi com a ajuda de vocês.

Ao meu orientador, prof. Luiz Henrique, por sua orientação impecável e que hoje considero um amigo. Obrigado pela confiança, paciência, conversas e convivência.

À Monique e ao Víctor, pela amizade e paciência. Obrigado por estarem presentes na minha vida, dentro e fora da universidade.

Ao Paulo e ao prof. Nilson, pela amizade e colaboração no desenvolvimento deste trabalho.

Aos membros da banca, Carlos e Marcelo pelas valiosas contribuições ao manuscrito final.

À Feagri e sua comunidade, por terem me acolhido e dado suporte em mais esta fase da minha formação.

Aos meus irmãos, Matheus, Martin e Ayla, aos amigos Cesar, Enrico, Victor, Caio, Guilherme e William, entre outros, que enriqueceram esse período da minha vida.

À equipe da unidade Alcídia da Odebrecht Agroindustrial, pela colaboração e apoio.

À FAPESP/Odebrecht Agroindustrial, pelo apoio financeiro ao projeto (Processo FAPESP n.º 12/50049-3).

Por fim, agradeço a todos os outros que tornaram esse trabalho possível, direta ou indiretamente, e cujos nomes não seria possível listar na totalidade.



# 1 INTRODUÇÃO

Com uma produção de cana-de-açúcar de 650 milhões de toneladas estimada para a safra de 2013 (CONAB, 2013), o Brasil é o maior produtor desta cultura (FAO, 2013). As projeções para a cultura no ano de 2013 são de que a área plantada será de 9,5 milhões de hectares, correspondente a 14 % dos 67,25 milhões de hectares cultivados no Brasil (IBGE, 2013), possuindo um valor bruto da produção de R\$ 48 Bilhões, o que corresponde a 17 % do VPB da Agricultura que foi de R\$ 276 Bilhões (MAPA, 2013). Como é natural na produção agrícola, além da condução da atividade fim, tem-se associado a essa produção as atividades de planejamento.

Para realização do planejamento no setor sucroenergético, é necessário estimar a produtividade da cana-de-açúcar dos talhões que fornecerão matéria-prima para a unidade industrial. Com o conhecimento da produtividade e da área do talhão, é possível estimar sua produção de cana-de-açúcar. Do ponto de vista gerencial, a produtividade dos talhões é uma informação essencial para realização de um planejamento agrícola que permita que se atinjam as metas empresariais em termos de matéria-prima (BRUGNARO; SBRAGIA, 1982). Em um planejamento ideal, a produtividade da cana-de-açúcar seria estabelecida em função do tipo de solo, tratos culturais, condições climáticas, variedade plantada, entre outros fatores (MARGARIDO, 2006). Para realização de um planejamento que contemple as variáveis pertinentes e as possíveis soluções, Argenton et al. (2010) destacam a necessidade do desenvolvimento e adoção de ferramentas de gestão da produção, que necessariamente utilizem modelos de previsão da produtividade, para planejamento e dimensionamento da produção agrícola.

O método utilizado amplamente para previsão da produtividade é a estimativa de especialistas. O procedimento de estimativa, em linhas gerais, consiste na inspeção dos canaviais por técnicos que consideram fatores como o tipo de solo, fertilidade, comportamento da variedade, estágio da cultura, épocas de corte e colheita, condições do clima, manejo, ocorrência de pragas ou doenças, produtividades anteriores, histórico da área, entre outros. Assim, o especialista agrega seu conhecimento e experiência para realizar a previsão da produtividade.

Variações desse procedimento podem incluir recursos como a análise da biometria da cana-de-açúcar, imagens de satélite e uso dos dados históricos. Ao longo da safra, essas estimativas são corrigidas em função de novos eventos ou do comportamento de outros talhões ou, ainda, aprimoradas por considerar novas informações. Cabe destacar que esse procedimento é subjetivo e depende do conhecimento implícito do especialista, e que em situações onde o especialista percorre somente o entorno do canavial, o desconhecimento das condições do interior dos talhões leva a erros na estimativa. Como alternativa, podem ser empregadas técnicas de modelagem para realização de previsão.

O uso de modelos de predição para tomada de decisão na agricultura possui diversas características ligadas à natureza do modelo utilizado. Um aspecto relevante é que, dada a importância do clima na modelagem agrícola, além do uso do modelo, é necessário utilizar também uma projeção de clima, conforme pode ser visto no trabalho de Hoozeboom (2000), onde o autor destaca a importância da agrometeorologia na modelagem agrícola. Além disso, Meinke e Stone (2005), ao discutirem o uso das projeções de clima na tomada de decisão intermediada por modelos, apontam a importância dos modelos, a despeito da estratégia de modelagem, em traduzir a projeção climática em uma grandeza relevante para a tomada de decisão. Assim, não basta ao tomador de decisão a simples projeção climática, é necessário um modelo para “traduzir” esse clima futuro em, por exemplo, produtividade agrícola. Os autores apontam a importância de uma abordagem interdisciplinar nesse processo, onde é necessário que seja apontado onde, quando e como utilizar as projeções de clima. Embora os autores dediquem seu trabalho a contextualizar o uso de diferentes projeções de clima para diferentes tomadas de decisão, se modelos serão empregados, é necessário que esses modelos também sejam contextualizados. São apontadas aplicações com horizonte de decisão variando de semanas a décadas, na escala de talhões à nacional, empregadas por tomadores de decisão de diferentes setores e níveis organizacionais.

Uma característica relevante da produção de cana-de-açúcar é a presença de diversos tomadores de decisão, uma vez que está inserida nas cadeias de valor do complexo da cana-de-açúcar (HIGGINS et al., 2007). O presente trabalho buscou contextualizar a tomada de decisão dentro de uma usina através de uma pesquisa exploratória descritiva. Para um contexto de

aplicação, foram exploradas técnicas de mineração de dados para criação de um modelo de produtividade, que foram então avaliadas. Nas demais partes deste trabalho, tem-se um capítulo de revisão da bibliografia (Capítulo 2), dividido em seções de relativas à produção da cana-de-açúcar (Seção 2.1) e sua modelagem, abordagens de modelagem dentro do contexto de mineração de dados (Seção 2.2) e quanto à avaliação de modelos de regressão (Seção 2.3). Na sequência, na forma de artigo, temos a contextualização da tomada de decisão no setor sucroenergético (Capítulo 3) e uma proposta de modelagem com potencial uso em um dos contextos identificados (Capítulo 4). O último capítulo, Considerações Finais, discute os resultados dos dois artigos.

## **2 REVISÃO BIBLIOGRÁFICA**

### **2.1 Produtividade de cana-de-açúcar e sua modelagem**

As variedades modernas de cana-de-açúcar são híbridos de espécies do gênero *Saccharum*, plantas que crescem em forma de touceira, sendo a parte aérea formada por colmos, o caule típico das gramíneas, folhas e inflorescências, com uma parte subterrânea formada por raízes e rizomas (MOZAMBANI et al., 2006). O desenvolvimento da cultura segue uma sequência de eventos, que caracterizam o crescimento da cana-de-açúcar. Essa sequência de eventos, ou estádios fenológicos, são fruto da interação entre as características genéticas da variedade em questão e o ambiente.

É possível dividir o crescimento da cana-de-açúcar em quatro etapas com características diferenciadas. De acordo com Silva et al. (2010), no estágio inicial, brotação e estabelecimento, tem-se a propagação da planta a partir de uma ou mais gemas contidas em um rebolo (pedaço do colmo). Em condições favoráveis, a gema brota e tem-se a emergência do primeiro perfilho, sendo uma fase dependente das reservas energéticas do rebolo. Na etapa posterior, perfilhamento e estabelecimento da cultura, são emitidos mais perfilhos e suas raízes se desenvolvem, sendo esta etapa considerada a principal no estabelecimento da produção através do fornecimento do número de colmos adequados à produção para aquela touceira. A terceira etapa é de crescimento e desenvolvimento da parte aérea, sendo caracterizada por processos fortemente dependentes da disponibilidade de água para a planta. O desenvolvimento das folhas que já ocorria na fase de perfilhamento agora acompanha uma fase de alongamento dos colmos, sendo uma etapa onde é acumulada 75 % de toda a matéria seca. Após essa etapa, os colmos começam a amadurecer gradualmente, em um processo onde um colmo apresenta entrenós mais maduros em sua base e menos maduros na região com folhas verdes, uma diferença que diminui com a evolução do processo de amadurecimento. O processo de amadurecimento depende de condições desfavoráveis ao desenvolvimento vegetativo, assim, baixas temperaturas e a presença de um moderado déficit hídrico intensificam a maturação.

Após o corte, permanecem no solo as raízes e rizomas da planta que passam a fornecer água e nutrientes para as gemas que irão brotar, dando origem a um novo ciclo de crescimento. A



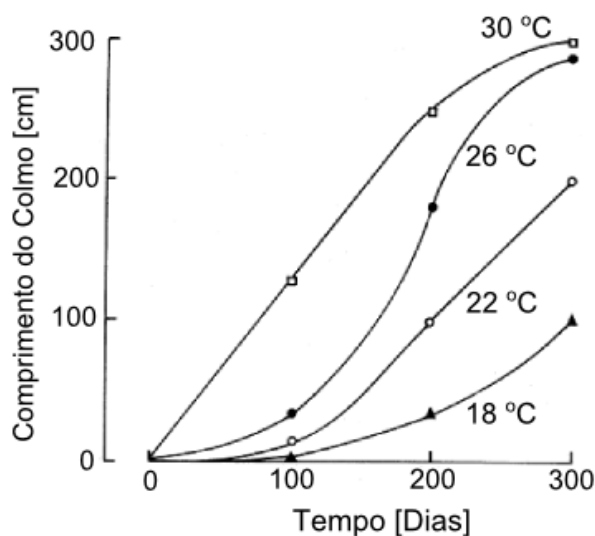
planta originária de um colmo plantado é conhecida como cana-planta, enquanto a planta originada da rebrota é denominada cana-soca. Ao longo dos sucessivos cortes, há um decréscimo na produtividade de forma que é necessário o replantio quando a produção deixa de ser economicamente atrativa.

O ciclo da cana-planta varia conforme sua época de plantio, sendo diferenciada em cana-de-ano, cana-de-ano e meio e cana de inverno. Assim, conforme Segato et al. (2006), para região centro-sul do Brasil, a cana-de-ano é plantada de setembro a novembro, possui um ciclo médio de 12 meses, passando por um desenvolvimento máximo de novembro a abril, e de acordo com a variedade pode ser colhida a partir de julho, fazendo com que a cultura tenha no geral até 8 meses de desenvolvimento e 4 meses para maturação. A cana-de-ano e meio, que é plantada no período que vai de janeiro ao início de abril ou maio, passa por um desenvolvimento inicial, que é limitado a partir de abril e é retomado com o retorno das chuvas tipicamente em setembro e vegeta novamente até abril. Dessa forma, a cana-de-ano e meio passa por um período inicial de 3 meses de desenvolvimento, 5 meses em condição de limitada disponibilidade de água no solo, e mais 7 meses de desenvolvimento, seguido do amadurecimento no inverno, o que leva a um ciclo que varia de 14 a 21 meses, conforme a data de plantio e a época de maturação da variedade utilizada. No geral, a produtividade do primeiro corte da cana-de-ano e meio é superior devido ao seu total de 10 meses de desenvolvimento, em relação aos 8 meses na cana-de-ano. Por último, os cultivos plantados nos meses de junho a agosto são conhecidos como cana de inverno ou de 15 meses e depende fortemente de estratégias que possam fornecer água para o desenvolvimento inicial da cultura (brotação e início do perfilhamento). Após a colheita da cana-planta, o ciclo da cana-soca é em média de 12 meses.

Uma consequência direta da época de plantio para a cana-planta, ou da colheita para a soqueira, é o clima ao qual estará submetida para o crescimento. Conforme pode ser visto no trabalho de Inman-Bamber e Smith (2005), é essencial que água esteja disponível para a fase de brotação e, após o início do perfilhamento, a planta se mostra resiliente a períodos de falta de água. Essa resiliência se dá pela capacidade de compensar o período de baixa disponibilidade com um crescimento acelerado quando o solo estiver úmido novamente. Os autores exemplificam com um experimento onde após a irrigação para brotação, a área não foi irrigada por 5 meses, e

obteve 50 % da área foliar, porém a produtividade final não foi afetada. Essa capacidade não ocorre para a fase de crescimento rápido, período onde a cana-de-açúcar apresenta maior sensibilidade ao déficit hídrico. Ao entrar na fase de maturação, a necessidade de água diminui. Cabe destacar que o déficit hídrico na fase de maturação aumenta a qualidade industrial da cana-de-açúcar.

Além do regime hídrico, o crescimento da planta é influenciado pela temperatura. De forma geral, a cana-de-açúcar possui uma temperatura ótima de crescimento de 30 °C e necessita temperaturas superiores a 18 °C para seu crescimento, apresentando limitações de crescimento para temperaturas superiores a 45° C. Cabe destacar que esses valores variam entre as diferentes cultivares e que o desenvolvimento a temperaturas superiores a 30 °C não sofre grandes limitações caso haja disponibilidade de água (EBRAHIM et al., 1998; GLASZIOU et al., 1965). É possível visualizar o efeito da temperatura no crescimento na Figura 1.



**Figura 1. Desenvolvimento do colmo em função do tempo e temperatura, adaptado de Glasziou et al. (1965).**

Uma das formas de estudar o efeito do clima nas diferentes fases de desenvolvimento é a caracterização do clima para cada subperíodo. Essa abordagem foi utilizada por Binbol et al. (2006) para estudar o efeito da variabilidade do clima na produtividade. Outros autores caracterizaram o clima para intervalos de tempo pré-estabelecidos. Greenland (2005) agrupou diferentes períodos trimestrais, enquanto Jiménez et al. (2008) utilizaram os 4 primeiros e 4 últimos meses do ciclo de desenvolvimento. Essa abordagem se mostra mais coerente quando

comparada à utilização da precipitação total durante o desenvolvimento da cultura, como realizado por Alvarez et al. (1982) e Brüggemann et al. (2001). Por não oferecerem justificativas para os períodos utilizados, a separação do clima se mostra arbitrária na abordagem de Jiménez et al. (2008) e Greenland (2005), enquanto a abordagem de Binbol et al. (2006) depende do conhecimento do comportamento geral da fenologia da cultura, sujeito a variações de acordo com as condições de desenvolvimento e cultivar em questão. Além da própria temperatura, um conceito relevante na modelagem de culturas são os graus-dias. Essa medida é uma forma de caracterizar a variação da temperatura ao longo do dia, utilizando as temperaturas máxima e mínima e computando, a partir da temperatura de base, o acúmulo do tempo térmico. Já foi demonstrado que o acúmulo de tempo térmico governa fenômenos importantes no desenvolvimento da cana-de-açúcar, como a taxa de aparição de folhas (BONNETT, 1998), perfilhamento (INMAN-BAMBER, 1994) e no desenvolvimento do dossel (INMAN-BAMBER, 1994; ROBERTSON et al., 1998).

De acordo com Prado et al. (2010) o ambiente de produção da cana-de-açúcar é definido em função das condições físicas, hídricas, morfológicas, químicas e mineralógicas do solo, sob manejo adequado da camada arável, associada com as condições de subsuperfície do solo e o clima regional. Essa definição é sintetizada pelos autores nos componentes: água, textura, profundidade e fertilidade. Em primeiro lugar os autores destacam a disponibilidade água, que pode reduzir drasticamente a produtividade se sua quantidade for limitante, sendo a disponibilidade de água dependente de diversos fatores como o regime pluviométrico, textura, estrutura, teor de matéria orgânica, mineralogia, gradiente textural, impedimentos físicos e profundidade. Os outros dois componentes, textura e profundidade, são relacionados a disponibilidade de água, desenvolvimento das raízes e volume explorado pelas raízes. O último componente, a fertilidade, é considerada fundamental pelos autores para que se obtenha produtividades altas. Assim, a camada arável deve ser manejada para que seja quimicamente favorável ao desenvolvimento das plantas.

Cabe destacar que enquanto as características químicas da camada arável se correlacionam com as produtividades no primeiro e segundo corte, as condições de subsuperfície apresentam melhor correlação com os cortes posteriores (LANDELL et al., 2003). A diversidade

de solos existentes associada à grande amplitude de condições climáticas nas quais a cana-de-açúcar é cultivada gera uma diversidade de ambientes de produção que vão interagir com as características genéticas da planta para determinar a produtividade agrícola do cultivo, que ainda é condicionada a outros fatores como ocorrência de pragas e doenças (LANDELL; BRESSIANI, 2010). Ao considerar ainda que o cultivo da cana-de-açúcar tem como objetivo maximizar a produção de açúcar e/ou álcool por unidade de área, os autores ainda consideram os efeitos da época de colheita na qualidade da matéria-prima ao longo do período de colheita. Para a safra da região centro-sul, que tipicamente vai de março a novembro, os autores agrupam os meses de março a junho como safra de outono, de julho a setembro como safra de inverno e de outubro a novembro para a safra de primavera. A interação das três épocas de colheita e dos tipos do solo, agrupados em favoráveis, médios e desfavoráveis, é representada em uma matriz de ambientes, com diversas implicações no manejo varietal, que de acordo com Landell e Bressiani (2010) é a estratégia que procura explorar os ganhos gerados pela interação entre genótipo e ambiente.

Assim, além das variáveis climáticas, nos estudos realizados por Alvarez et al. (1982) e Brüggemann et al.(2001), onde foram propostos modelos empíricos para a produtividade de cana-de-açúcar, também foram incorporadas variáveis relacionadas ao manejo e ao ambiente de produção. Essa caracterização se deu através da classificação do solo, manejo de adubação e caracterização da granulometria do solo. O uso de diferentes variáveis para o desenvolvimento de modelos empíricos é uma forma de relacionar a produtividade aos diferentes fatores que influenciaram o desenvolvimento daquele cultivo. Um modelo robusto que estabeleça essa relação empírica é uma forma de realizar previsões e embasar a tomada de decisão, porém não se deve esperar que esse tipo de modelo tenha aplicações para fora do seu escopo de criação (PASSIOURA, 1996).

## **2.2 Mineração de dados**

Com um volume cada vez maior de dados sendo gerados e armazenados, tem aumentado a diferença entre a quantidade gerada e nosso entendimento desses dados. Esses conjuntos de dados podem conter informações potencialmente úteis, e devido ao tamanho desses conjuntos de dados, a avaliação por analistas é difícil, fazendo com que as informações contidas nesses dados sejam raramente explicitadas ou exploradas (WITTEN; FRANK, 2005). Nesse contexto, de acordo com

os autores, as técnicas de mineração de dados são uma opção para encontrar essas informações a partir do grande volume de dados gerados. Mineração de dados é definida então como o processo de descoberta de padrões em dados, sendo o processo automatizado ou semi-automatizado nos quais devem-se encontrar padrões úteis.

As tarefas de mineração de dados podem ser divididas em descrição ou predição. Em descrição, temos as tarefas de agrupamento, onde se busca produzir grupos nos quais a diferença entre seus componentes é minimizada, enquanto a diferença entre os grupos é maximizada. Nas tarefas de associação, o objetivo é encontrar atributos que ocorrem simultaneamente com certa frequência. Na aplicação mais comum das tarefas de associação, são analisados os históricos de compras e busca-se identificar os produtos que são comprados em conjunto frequentemente. Diferente das tarefas de descrição, as tarefas de predição buscam prever o valor de um determinado atributo, chamado atributo meta, em função dos outros atributos do conjunto de dados. Quando esse atributo meta é categórico, a tarefa é de classificação, enquanto casos onde o atributo meta é um valor numérico contínuo, a tarefa é de regressão (WITTEN; FRANK, 2005).

O processo de descoberta de conhecimento reúne uma série de atividades e apresenta um caráter iterativo. De acordo com a metodologia CRISP-DM (CHAPMAN et al., 2000), o processo é dividido em 6 etapas. Na Figura 2 pode ser vista a sucessão das etapas e a representação da natureza cíclica do modelo. As etapas são explicadas na sequência.

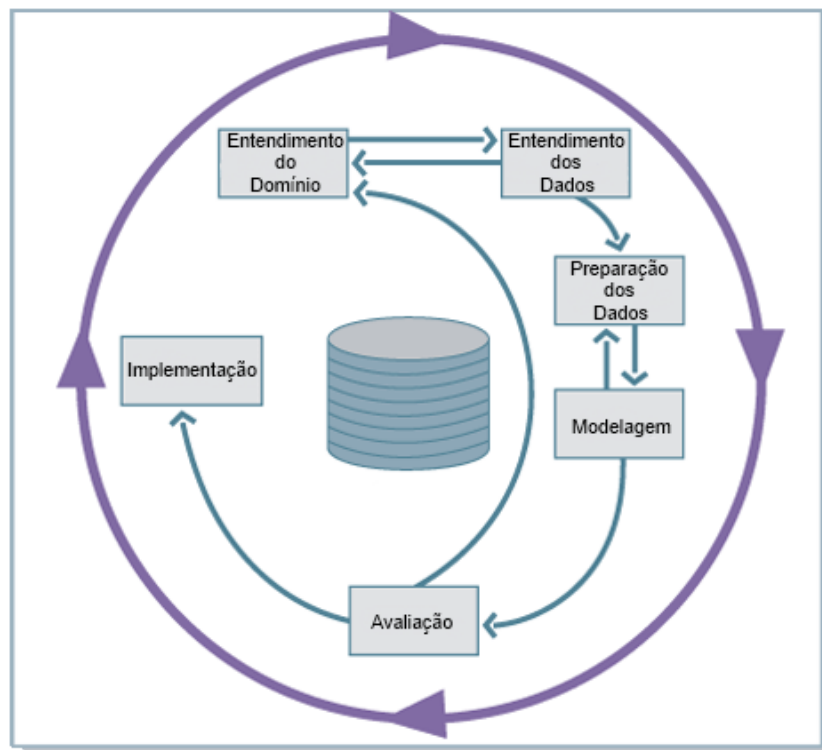


Figura 2. Etapas do ciclo CRISP - DM (Adaptado de CHAPMAN et al. (2000)).

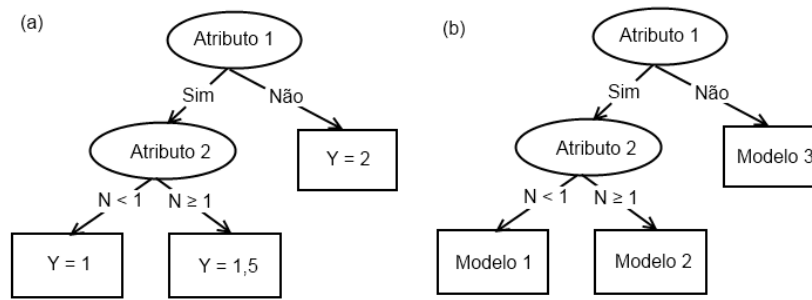
1. **Entendimento do domínio** - Esta etapa consiste no entendimento da área de conhecimento onde serão empregadas as técnicas de mineração de dados para descoberta de conhecimento. Com o avanço das outras etapas, pode ser necessário retomar essa etapa de entendimento.
2. **Entendimento dos dados** - Após o entendimento do domínio, os dados devem então ser estudados e analisados. Os dados devem ser descritos, em termos do que representam, unidades utilizadas, além de identificar possíveis problemas como dados faltantes, valores incoerentes, erros de preenchimento, etc.
3. **Preparação dos dados** - De acordo com a técnica de mineração de dados que será utilizada, pode ser necessário realizar transformações nos dados. Além disso, em certos casos, a transformação dos dados pode não ser um requisito, mas pode colaborar para um melhor resultado.
4. **Modelagem** - Aplicação das técnicas de mineração de dados e calibração dos seus parâmetros. Uma vez que várias técnicas podem ser aplicadas para diferentes problemas e essas técnicas necessitam de diferentes tratamentos nos dados de entrada, pode ser

necessário uma retomada da etapa anterior. Nesta etapa, são ajustados os parâmetros das diferentes técnicas, o que é uma tarefa iterativa com o objetivo de otimizar a performance do modelo.

5. **Avaliação** - Após a etapa de modelagem, os resultados obtidos devem ser avaliados, assim como as etapas desenvolvidas em sua elaboração. Resultados insatisfatórios nesta etapa podem levar à retomada de etapas anteriores.
6. **Implementação** - O conhecimento descoberto é utilizado, de forma direta ou na forma de implementação de um modelo.

O crescimento das aplicações de mineração de dados na agricultura pode ser visto nas revisões realizadas por Mucherino et al. (2009) e Mucherino e Ruß (2011). Tendo em vista os casos de sucesso na área agrícola e o volume de dados acumulado pelo setor sucroenergético, têm-se um potencial a ser explorado. Em estudos com dados de menor escala, com intuito de estudar a variabilidade da produtividade de cana-de-açúcar, técnicas de mineração de dados têm se mostrado superiores a técnicas estatísticas convencionais (ANDERSON et al., 1999; FERRARO et al., 2009). Em ambos os estudos, a técnica de mineração de dados utilizada foi a árvore de regressão (*Classification and Regression Trees*, CART). Essa técnica, junto com redes neurais artificiais (*Artificial Neural Network*, ANN) e máquinas de vetor de suporte (*Support Vector Machines*, SVM) foram utilizadas na predição de produtividade de trigo em função de dados de reflectância na cor vermelha, condutividade elétrica do solo e dados sobre as estratégias de adubação (RUß, 2009). O autor aponta que a técnica SVM se mostrou superior às demais, sendo recomendada para tarefas desse tipo.

Árvores de regressão são construídas da mesma forma que árvores de decisão. O conjunto de dados é particionado sucessivamente, até que no último nível da árvore, as folhas, tenha-se o valor médio do atributo meta dos registros daquela folha, o que caracteriza a árvore de regressão. Outra opção é uma árvore de modelos, que apresenta uma equação de regressão em cada folha (WITTEN; FRANK, 2005). A Figura 3 apresenta uma ilustração de uma árvore de regressão (a) e árvore de modelos (b). Uma vantagem dessa técnica é a construção de um modelo que pode ser melhor entendido, o que não acontece com técnicas do tipo "caixa-preta" como as ANN ou SVM. Por outro lado, técnicas "caixa-preta" tendem a apresentar melhores capacidades preditivas.



**Figura 3. Árvore de Regressão e Árvore de Modelo**

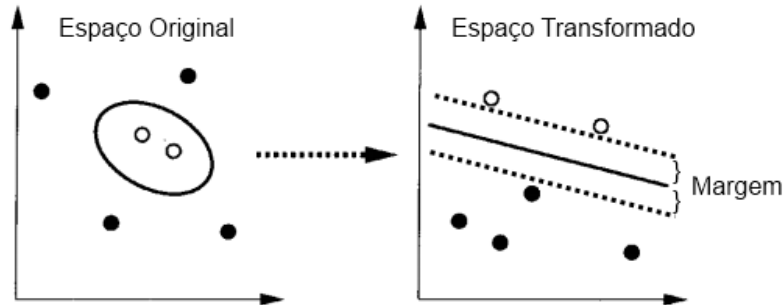
Uma importante evolução dos modelos baseados em árvores são as *Random Forest*. Breiman (2001) demonstrou que para um modelo composto por diversos modelos (*ensemble*), a taxa de erro é dependente da capacidade preditiva de cada submodelo e inversamente proporcional à correlação entre os mesmos. A solução encontrada pelo autor consiste em induzir árvores a partir de um subconjunto dos atributos originais e escolhendo aleatoriamente o atributo utilizado para divisão dos nós. Com a geração de diversos modelos, o resultado final é gerado a partir de uma composição dos resultados, podendo ser uma votação para tarefas de classificação ou a média para tarefas de regressão.

As ANN surgiram no campo da neurologia, quando pesquisadores da área buscaram simular neurônios. Uma ANN é composta basicamente por uma rede com unidades de entrada e saída interconectadas, com diferentes pesos para as conexões. Durante o treinamento da ANN os pesos das conexões e outros parâmetros do modelo são calibrados, para que a rede seja capaz de produzir uma saída próxima à variável que está tentando prever. O erro entre a predição e o valor real é utilizado na calibração, o que caracteriza o algoritmo de *Back Propagation* (HAN et al., 2012).

As SVM foram desenvolvidas com base em teorias de aprendizado estatístico nos anos 1990, sendo inicialmente concebidos para tarefas de classificação e posteriormente generalizados para tarefas de regressão (SCHÖLKOPF et al., 1997). Pela forma como são formulados em um problema de otimização, o vetor suporte encontrado é um ótimo global, diferente dos resultados encontrados para as RNA, que são ótimos regionais. Ainda de acordo com os autores, o algoritmo encontra o vetor que separa as classes com maior margem possível. Caso os dados não sejam linearmente separáveis, é feita a transformação para um espaço de dimensão maior, onde podem



ser separados linearmente. Na Figura 4, a seguir, pode ser vista transformação e a determinação do vetor suporte com a maior margem possível.



**Figura 4.** Transformação dos dados e determinação do vetor suporte, adaptado de Schölkopf et al. (1997).

De forma análoga, em uma tarefa de regressão, os vetores suporte são utilizados para representar uma função, que através do mapeamento em um espaço de dimensão maior, pode não ser linear no espaço original. Para permitir a generalização do algoritmo de classificação para a regressão, o algoritmo busca o vetor que melhor aproxima a função dentro de uma margem  $\epsilon$  pré-estabelecida, caracterizando o algoritmo de regressão por vetores suporte ( $\epsilon$ -SVR), de acordo com Schölkopf et al. (2000). Quanto menor o valor de  $\epsilon$ , maior a complexidade do modelo gerado, ou seja, maior o número de vetores suporte criados para representar a função. Para mediar o *trade-off* entre a complexidade do modelo e a margem de erro tolerável, a variável  $C$  penaliza na função objetivo os pontos fora da margem  $\epsilon$ . Para o caso bidimensional e linear, a representação das variáveis pode ser vista na Figura 5. Outro algoritmo, o  $\nu$ -SVR, através de uma reformulação da função que é otimizada para determinação dos vetores suporte, minimiza  $\epsilon$  em função da variável  $C$  e do *trade-off* com a complexidade do modelo.

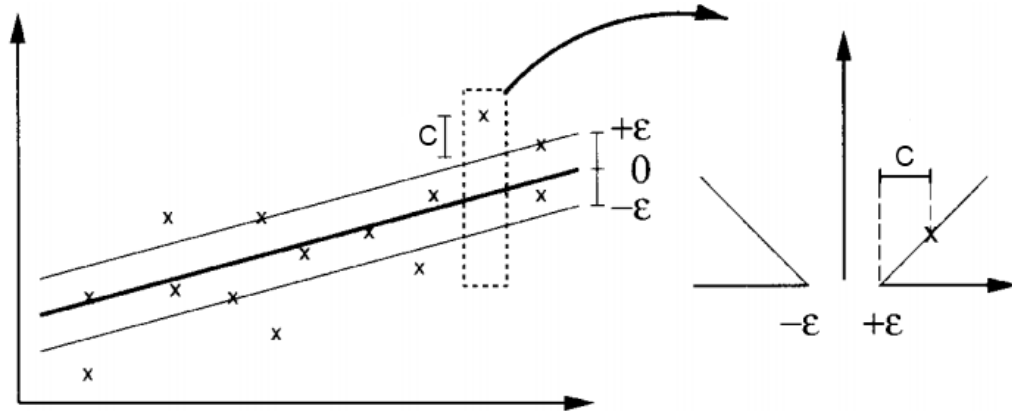


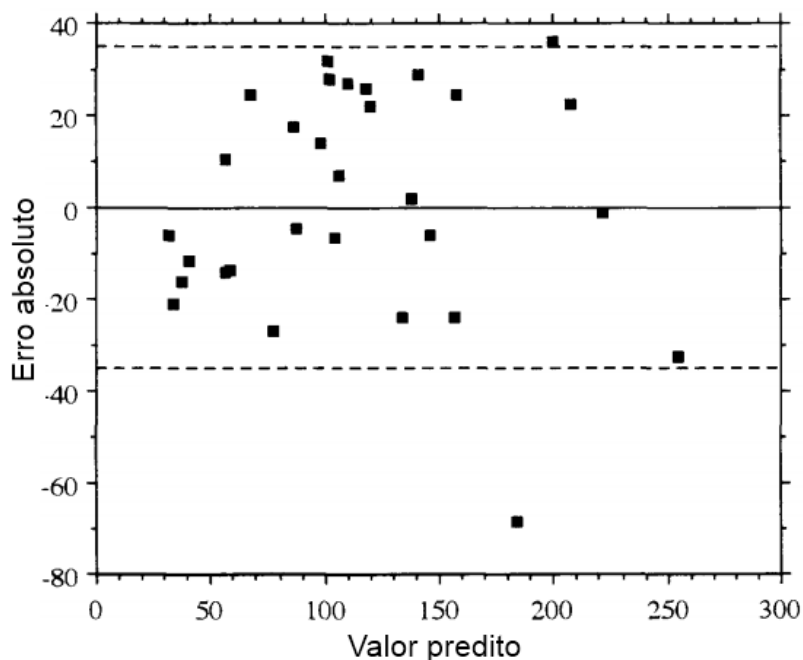
Figura 5. Parâmetros das SVM e seu significado na regressão. Adaptada de Schölkopf et al. (2000).

### 2.3 Avaliação de modelos de regressão

A avaliação de modelos de regressão, independente da técnica utilizada para sua criação, é realizada a partir de métricas que quantifiquem a adequação do seu resultado em relação aos valores reais. Entre as métricas disponíveis, temos o erro quadrático médio (*Mean Square Error*, MSE), raiz do erro quadrático médio (*Root Mean Square Error*, RMSE) e erro absoluto médio (*Mean Absolute Error*, MAE), medidas que também podem ser relativizadas. Outra métrica frequentemente utilizada é o coeficiente de correlação (WITTEN; FRANK, 2005).

O uso do coeficiente de regressão como métrica de avaliação do modelo é feito através do cálculo do mesmo na comparação entre os valores preditos e os observados, o que de acordo com Mitchell (1997) e Harrison (1990) é um uso impróprio do coeficiente. De acordo com Mitchell (1997), a regressão linear não foi concebida para esse tipo de tarefa, e sim determinar a equação da reta para uma variável resposta em função da variável independente. As críticas do autor se estendem quanto à ambiguidade da hipótese nula dos testes de confiança realizados sobre os coeficientes da regressão e à incompatibilidade entre o tipo de teste sendo conduzido e a maturidade pressuposta para um modelo em fase de validação. Por fim, é destacado pelo autor que os dados raramente respeitam as premissas teóricas para regressão. A preocupação quanto às premissas teóricas, só que em relação aos testes estatísticos conduzidos sobre os coeficientes da regressão, também é abordada por Harrison (1990). Uma proposta é avaliar o erro das predições ao longo dos valores preditos, comparando os erros com uma margem de erro tolerável pré-estabelecida, o que pode ser visto na Figura 6. Por outro lado, a comparação entre valores

observados e preditos, é vista por Bi e Bennet (2003) como uma forma de identificar tendências do modelo. Cabe destacar a crítica de Mitchell (1997) de que embora a regressão da reta dos valores preditos e observados possa estar próxima da reta 1:1, esse não é um motivo suficiente para constatar a qualidade do modelo.

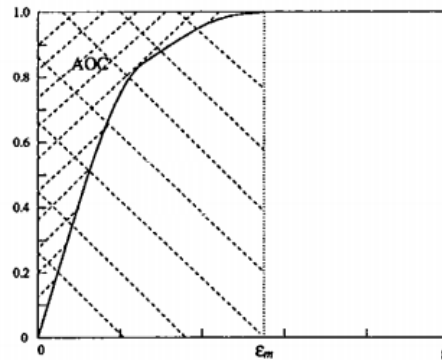


**Figura 6.** Inspeção do erro absoluto ao longo dos valores preditos. Adaptado de Mitchell (1997).

Ao comparar as métricas RMSE e MAE, Willmott e Matsuura (2005) concluem que o uso do MAE apresenta vantagens sobre o RMSE. O MAE é mais fácil de calcular e interpretar, pois tem um significado direto. Enquanto ao calcular o RMSE, o valor encontrado é uma função da variância do erro, do número de pontos avaliados e do resíduo do modelo, fazendo com que não seja possível interpretar diretamente o valor ao comparar modelos, pois há ambiguidade na medida. Outro aspecto relevante do RMSE é que não há uma forma de interpretar o valor do erro quando calculado dessa forma. Os autores ainda destacam uma terceira métrica em sua discussão, a tendência média de erro (*Mean Bias Error*, MBE), que é a média da soma dos valores preditos (P) menos os valores observados (O). Essa métrica permite inferir se o modelo tende a superestimar ( $P > O$ ,  $MBE > 0$ ) ou subestimar ( $P < O$ ,  $MBE < 0$ ).

Uma proposta para avaliar os resultados de regressão são as curvas características dos erros de regressão (*Regression Error Characteristic*, REC) descritas por

BI e BENNETT (2003). As curvas REC representam a função de distribuição acumulada dos erros, sendo a abcissa a magnitude do erro e o eixo ordenado a acurácia, entendida como a probabilidade  $p_i$  - do erro  $\epsilon$  ser menor que  $\epsilon_i$ . Os autores demonstram como a análise de curvas REC pode fornecer ao usuário dos modelos de regressão informações de forma visual, facilitando a interpretação por usuários não especialistas em estatística e permitindo a inferência de parâmetros como erro médio e a estatística KS (Kolmogorov- Smirnov) a partir de propriedades geométricas das curvas. No *framework* proposto pelos autores, é possível comparar a performance de diversos modelos de forma simultânea. As curvas REC, ainda de acordo com os autores, são análogas às curvas ROC e possuem uma série de suas características positivas. Caso o valor de erro utilizado na construção da curva REC seja o erro absoluto, a área sobre a curva (*Area Over Curve*, AOC) representa uma sub-estimativa do MAE. Na Figura 7, pode ser visto um exemplo de uma curva REC com AOC delimitada.



**Figura 7. Exemplo de curva REC com delimitação da AOC.**

Além da determinação da métrica de avaliação, o procedimento para validação do modelo apresenta diferentes possibilidades. A primeira alternativa é dividir o conjunto em treino/teste, onde é realizada uma divisão aleatória do conjunto de dados onde uma parte do conjunto é utilizada para treinamento do modelo, enquanto o restante é utilizado para teste. Esse procedimento pode ser repetido várias vezes para uso da média do parâmetro utilizado na avaliação. Outro procedimento é a validação cruzada, ou *cross-validation* (CV). No caso do procedimento escolhido ser a CV, o conjunto de dados é particionado em N subconjuntos de tamanho aproximadamente igual. O conjunto é treinado em N-1 subconjuntos e avaliado no subconjunto restante (HAN et al., 2012). Esse procedimento é realizado N vezes, de forma que o modelo seja treinado e avaliado em todo o conjunto de dados. Os subconjuntos são chamados de

*folds* e normalmente a CV é realizada com 10 conjuntos, ou 10-folds. Diferente do procedimento anterior, na CV os subconjuntos são divididos inicialmente e treinados/testados de maneira excludente.

O uso das estratégias de validação de modelos destacadas acima tem como objetivo avaliar e evitar o *overfitting*, que é caracterizado por um modelo ser ajustado ao conjunto de dados onde é treinado de maneira tão específica que passa a não apresentar boas performances quando avaliado em outros conjuntos. Em linhas gerais, árvores com muitas folhas estão mais propensas ao *overfit*, o que pode ser controlado nos parâmetros de geração do modelo. As SVM são menos propensas ao *overfit* quando comparadas com a maioria das técnicas (HAN et al., 2012).

### **3 CONTEXTUALIZAÇÃO DA TOMADA DE DECISÃO**

Título do Artigo:

**Diferentes demandas de informação de produtividade no planejamento de uma usina de cana-de-açúcar**

#### **3.1 Resumo**

---

Na tomada de decisão e planejamento de uma usina de cana-de-açúcar, é constante a necessidade de quantificar o resultado da decisão na produtividade ou produção de cana-de-açúcar e seu efeito no planejamento. Foi conduzida uma pesquisa exploratória junto a uma usina com intuito de mapear as principais decisões que são influenciadas pela perspectiva de produtividade futura, bem como a forma que essas previsões afetam o planejamento. Após uma série de entrevistas e acompanhamento de atividades, foi possível identificar decisões chave e suas características, que foram relacionadas a soluções propostas pela comunidade científica e enquadradas dentro de uma proposta de framework unificado para tomada de decisão e planejamento. Entre as decisões identificadas, as projeções de produtividade utilizadas para subsidiar a elaboração de orçamento e plano de safra se mostraram críticas, dado que as ações tomadas à partir dessas decisões irá atravessar toda a cadeia de valor, fazendo com que seja necessário um framework de tomada de decisão que mostre os efeitos das decisões nos processos posteriores. A proposta apresentada se apoia na integração dos diferentes planos sobre uma base de informações comuns, uso de modelos empíricos de produtividade para subsidiar a tomada de decisão com uso conjunto de modelos de planejamento.

---

#### **3.2 Introdução**

Devido ao longo ciclo de crescimento da cana-de-açúcar, variando de 12 a 18 meses em condições típicas, surgem diversas necessidades de prever a produtividade e o resultado agrícola. Cabe destacar que o interesse não é apenas sobre a produtividade (expressa em toneladas por hectare – TCH), mas também sobre a quantidade de açúcar acumulada nos colmos, o que permite inferir o resultado da produção em termos de açúcar e etanol. O ciclo extenso em

conjunto com uma safra que ocorre entre março e novembro para a região centro-sul do Brasil, que concentra 90 % da produção (CONAB, 2013), faz com que as áreas cultivadas estejam em diversas fases de crescimento ao longo do ano. Essa diversidade de condições faz com que, ao longo do tempo, sejam necessárias novas previsões de produtividade, para refletir o resultado das condições ocorridas e para embasar novas decisões que precisam ser tomadas no manejo agrícola e na operação de usinas.

A cada momento, algumas áreas estão próximas ao período de colheita, nas quais é mais coerente falar em estimativa da produtividade, em um sentido de estimativa como uma medida aproximada. Enquanto para áreas com a colheita distante é mais coerente falar de uma projeção da produção, uma vez que diversos eventos futuros, notadamente o clima futuro e o surgimento de condições fitossanitárias adversas que podem interferir na produtividade. Destacada a diferenciação adotada entre estimativa e projeção, o termo previsão será utilizado para antecipação de um resultado futuro.

Com o avanço do ciclo de crescimento e aproximação da época de colheita, há um aumento das informações disponíveis para realizar uma melhor previsão da produtividade, até o momento em que se deixa de realizar projeções e é possível estimar a produtividade através de levantamentos de campo. É importante destacar uma prática do setor, a de realização de estimativas por especialistas. Assim, especialistas que atuam na produção de cana-de-açúcar são capazes de estimar a produtividade baseada em inspeção visual, ou projetar a produtividade a partir do conhecimento sobre o histórico da região, desempenho de cultivares, características das áreas, clima típico e ocorrência de pragas e doenças. Em um contexto mais amplo, as previsões de produtividade embasam o planejamento da produção de cana-de-açúcar, enquanto novas previsões permitem o replanejamento da produção.

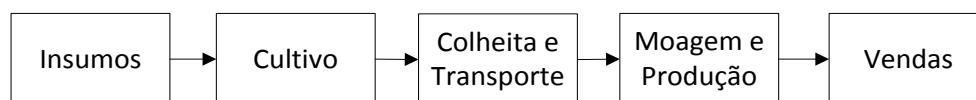
Investigando o uso de projeções na performance organizacional, Danese e Kalchschmidt (2011) mostraram que, para companhias de manufatura, a adoção de um processo estruturado para realização de projeções orientado a tomada de decisão tem impacto positivo direto nos custos e performance das entregas. Os resultados dos autores são surpreendentes ao demonstrarem que um menor erro não leva necessariamente a um melhor desempenho da companhia, nem a adoção de técnicas leva necessariamente a um menor erro. De acordo com os autores, a melhor performance organizacional está ligada à variável “coleta de informações a

partir de diversas fontes” e à variável “papel das projeções na tomada de decisão”. Eles destacam que as projeções devem ter credibilidade, estarem disponíveis na época correta e permitirem o uso na tomada de decisão, de acordo com as necessidades gerenciais. Higgins et al. (2007) apontam que as principais diferenças das cadeias de valor de manufatura em relação às cadeias de valor na agricultura são a maior variabilidade dos sistemas de produção. Entre os principais fatores, os autores apontam as incertezas e a variabilidade associada ao clima e às características biofísicas. Na cadeia de valor do açúcar, no contexto australiano, ainda é destacada a presença de diversos agentes tomadores de decisão e a necessidade de decisões em diversas escalas, variando de talhões individuais até a área total de uma usina. Assim, enquanto nos sistemas manufatureiros o sistema de produção é mais previsível e existe uma necessidade de projeção da demanda, em sistemas agrícolas, é necessário também realizar projeções da produção, resultado da interação entre fatores controlados e não controlados.

Quando se analisa a tomada de decisão na cadeia de valor, é possível analisar também os impactos das decisões nas cadeias subsequentes, fazendo com que o uso de projeções e modelagem possa melhorar a tomada de decisão para um contexto além da produção agrícola *per se*. Ahumada e Villalobos (2009) caracterizam as cadeias de valor agrícola como compostas pelos processos de produção, colheita, e armazenagem e distribuição. Por se tratar de uma cadeia de valor agroindustrial, Higgins et al. (2007) adicionam à cadeia de valor do açúcar os processos relacionados à usina, ou seja, moagem e produção, transporte do açúcar e armazenagem, cabendo destaque também para produção de outros produtos (etanol e eletricidade a partir de biomassa). Nas condições brasileiras, as áreas do cultivo agrícola destinadas à produção de açúcar são controladas pelas usinas, fazendo com que a tomada de decisão, a partir de projeções, possa ser feita considerando todos os processos da cadeia. Além disso, grupos que controlam mais de uma usina podem se beneficiar de uma tomada de decisão que considere suas diversas unidades. A existência de diferentes agentes ao longo da cadeia é uma das dificuldades apontadas por Higgins et al. (2007) para adoção de um framework de tomada de decisão unificado. Tomando por base o descrito por Higgins et al. (2007) e Ahumada e Villalobos (2009) e adequando ao contexto brasileiro, a cadeia de valor do açúcar é ilustrada na Figura 8. Embora a literatura revisada e as considerações deste artigo se refiram à cadeia de valor do açúcar, elas podem ser extrapoladas



para a cadeia de valor do etanol ou eletricidade, ou para outros produtos do complexo da cana-de-açúcar.



**Figura 8. Cadeia do valor do açúcar.**

Ao descrever as decisões que podem ser beneficiadas por previsões e uso de modelagem, diferentes autores (AHUMADA; VILLALOBOS, 2009; EVERINGHAM et al., 2002; MEINKE; STONE, 2005) descrevem decisões em diferentes escalas, tomadas em diferentes momentos de antecipação, realizadas por diferentes tomadores de decisão, o que ilustra a dificuldade apresentada por Higgins et al. (2007) na cadeia de valor do açúcar. Ahumada e Villalobos (2009) revisaram os diferentes sistemas de suporte à decisão para cadeias de valor agrícolas quanto ao nível organizacional da tomada de decisão (operacional, tático ou estratégico) e quanto à área funcional que é afetada (cultivo, colheita, distribuição e armazenagem) e diferenciando-os quanto ao tipo de modelagem utilizada (estocástica ou determinística). Em outra abordagem, Meinke e Stone (2005) exemplificam decisões que podem se beneficiar do uso de projeções climáticas e destacam o uso de modelos para gerar a previsão de produtividade. Essas decisões variam da logística interna de colheita a políticas de uso da terra e diferentes frequências (intra-safra a décadas) e janelas temporais (meses a anos). A relação entre a época da tomada de decisão e sua janela temporal pauta quais informações sobre o clima estão disponíveis e quais outras deverão ser projetadas. Uma relação de decisões na cadeia do açúcar tomada a partir de informações climáticas é mostrada por Everingham et al. (2002), que utilizaram projeções de clima para tomada de decisão na cadeia de valor do açúcar. Entre as diversas decisões listadas pelos autores, que dividem a cadeia em cultivo, colheita e transporte, moagem e produção, e vendas, eles apresentam decisões chave da indústria e utilizam a projeção de clima para melhorar a tomada de decisão para quatro delas: a projeção da produtividade e seu efeito na venda de açúcar no mercado futuro, o uso da projeção climática na tomada de decisão no manejo de irrigação, a determinação do início e fim da safra, e manejo da colheita.

A previsão de produtividade pode ser realizada de diversas formas, de acordo com as informações disponíveis. Uma forma de prever a produtividade são os modelos de crescimento, que, em conjunto com informações de manejo e projeções de clima, são capazes de descrever o

crescimento da planta e fornecer projeções da produtividade. Aplicações deste tipo foram realizadas para África do Sul e Austrália, com a proposta de prever a produção da safra seguinte (BEZUIDENHOUT; SINGELS, 2007; EVERINGHAM et al., 2002). Uma proposta para contornar a demanda de informações para essas aplicações, na Austrália, é apresentada por Everingham et al. (2009), que utilizou a combinação do resultado de diversos modelos (*ensemble*) representados através de diferentes condições de modelagem para a projeção regional da produtividade. Outra alternativa é agrupar áreas similares em blocos homogêneos, estratégia utilizada por Bezuidenhout e Singels (2007) e por Le Gal et al. (2009) para diminuir o número de simulações de crescimento realizadas.

Uma alternativa aos modelos de crescimento são os modelos empíricos que buscam relacionar condições de cultivo e de clima com a produtividade final. Lobell e Burke (2010) apresentam os modelos empíricos e de crescimento como alternativas complementares de modelagem de produtividade. Cabe destacar que as conclusões dos autores foram realizadas no contexto de estudo de mudanças climáticas. Essa discussão é expandida por Stone e Meinke (2005) enquanto discutem as abordagens de modelagem para predição de produtividade. Eles apresentam os modelos de crescimento e os modelos empíricos como ferramentas de estudo, tanto de mudança climática, quanto para variabilidade climática, diferenciando a mudança como as alterações de longo prazo do clima e a variabilidade como um reflexo da variação intrínseca do clima.

É importante destacar que se trata de uma cadeia de informações, onde os elementos básicos são os modelos, os dados que caracterizam a produção e as projeções de clima. Modelos de produtividade combinados com os dados de produção e projeção de clima permitem a projeção da produtividade. Com modelos do sistema produtivo, é possível utilizar essas projeções de produtividade para tomar decisões em um próximo nível. Assim, parte-se de decisões que podem ser tomadas a partir da projeção de clima (EVERINGHAM et al., 2002), para decisões que podem ser tomadas com o uso das projeções de clima e modelos de produtividade (MEINKE; STONE, 2005; STONE; MEINKE, 2005), e finalmente para as decisões que utilizam essas informações para um planejamento que varia do nível tático ao estratégico (AHUMADA; VILLALOBOS, 2009).

Com o avanço da safra, é possível incorporar novas projeções de clima e a parte do clima ocorrido nas simulações de produtividade, sendo esperado que o erro na projeção diminua (HOOGENBOOM, 2000). Além das informações de clima e de manejo para os modelos empíricos, outra possibilidade é a utilização de informações de sensoriamento remoto, que podem agregar mais informações na projeção da produtividade. Entre as fontes de sensoriamento remoto, têm-se as informações obtidas a partir de sensores em satélites, fotografias aéreas ou mesmo sensores terrestres, que através da medição da resposta radiométrica da superfície são capazes de inferir propriedades que variam do Índice de Área Foliar (IAF) das culturas, stress hídrico, deficiências de nutrientes e condições de fitossanidade à própria biomassa (ABDELRAHMAN; AHMED, 2008). Uma aplicação dessa abordagem pode ser vista no trabalho de Picoli (2006), que utilizou os dados de sensoriamento remoto (NDVI) em conjunto com dados de talhão para modelar empiricamente a produtividade de cana-de-açúcar. Essas informações podem gerar modelos nas escalas em que são criadas; assim, a predição de produtividade pode ser na escala de talhões ou na escala utilizada para agricultura de precisão.

Além de subsidiar a tomada de decisão, essas projeções de produtividade também podem integrar-se a modelos de otimização de sistemas agrícolas. Entre as aplicações para cana-de-açúcar, há o uso de modelos para realização do plano de safra, plano de colheita e roteirização. O plano de safra otimiza a entrega de matéria-prima para moagem com a maior quantidade de açúcar disponível, de acordo com as perspectivas de produção. Assim, a quantidade de cana-de-açúcar fornecida deve ser tal que a moagem possa operar ininterruptamente, durante a safra e sujeita ao limite de moagem, e minimiza a armazenagem da cana-de-açúcar, pois sua qualidade industrial diminui depois de colhida. Aplicações de otimização do plano de safra podem ser tanto para cadeias de valor onde a produção e a usina são controladas por diferentes agentes como na Austrália (LE GAL et al., 2009), África do Sul (JIAO et al., 2005) e Tailândia (PIEWTHONGNGAM et al., 2009), quanto para as condições da Venezuela (GRUNOW et al., 2007) e Brasil (JENA; POGGI, 2013), onde se têm um agente único.

Após a realização desse plano, tem-se o plano de colheita. Na etapa anterior (plano de safra), já são consideradas as datas de colheita de forma aproximada, porém neste momento o plano é refinado para refletir a evolução dos cultivos. Em geral, esse planejamento, tido como de nível

operacional, é feito em onda<sup>1</sup>, sendo escolhidos os talhões ou fazendas que devem ser colhidos na próxima semana ou mês, a partir do plano de safra. Ele também é realizado diversas vezes ao longo da safra e possui diferentes características dependendo do arranjo entre produtores e usinas (LE GAL et al., 2009; HIGGINS, 2002; PIEWTHONGNGAM et al., 2009) ou de acordo com as características das usinas (GRUNOW et al., 2007; JENA; POGGI, 2013). Para usinas, as restrições são operacionais com o objetivo de maximizar a qualidade da matéria-prima, enquanto em arranjos de produtores-usinas, acordos comerciais podem impor outras restrições.

Embora o plano de colheita determine quais talhões serão colhidos em determinado intervalo de tempo, é necessário também planejar a sequência de colheita das áreas, ou seja, roteirizar a colheita. Na roteirização é feito um balanceamento da distância em que as diferentes frentes se encontram e se busca minimizar o deslocamento total das frentes. É possível realizar o plano de colheita com restrições relativas à roteirização (LE GAL et al., 2009; GRUNOW et al., 2007; JENA; POGGI, 2013). Ainda cabe destacar a necessidade de replanejamento diante de condições críticas, como chuvas excessivas que inviabilizam a colheita ou tráfego, o florescimento de uma área que leva a queda de qualidade industrial ou mesmo incêndios acidentais ou criminosos. Tanto o florescimento quanto os incêndios fazem com que a área deva ser colhida o mais rápido possível. Enquanto em condições brasileiras não há necessidade de replanear a colheita devido à ocorrência de neve (BEZUIDENHOUT; SINGELS, 2007), a ocorrência de geadas também faz com que seja necessário alterar o planejamento.

Das aplicações revisadas anteriormente, apenas Piewthongngam et al. (2009) utilizaram um modelo para prever a produtividade (CANEGRO/DSSAT), enquanto as outras aplicações utilizaram médias históricas para realizar o planejamento. Le Gal et al. (2009), em sua conclusão, chamam a atenção para o uso do modelo como uma forma de melhorar a representação do sistema e apontam esse uso como um trabalho futuro. Nesse ponto, duas críticas podem ser feitas ao uso de médias históricas para o planejamento agrícola. Em primeiro lugar, Ahumada e Villalobos (2009) destacam a necessidade de que o planejamento utilize técnicas estocásticas, para refletir melhor a natureza dos sistemas. Higgins et al. (2007) também chamam atenção para

---

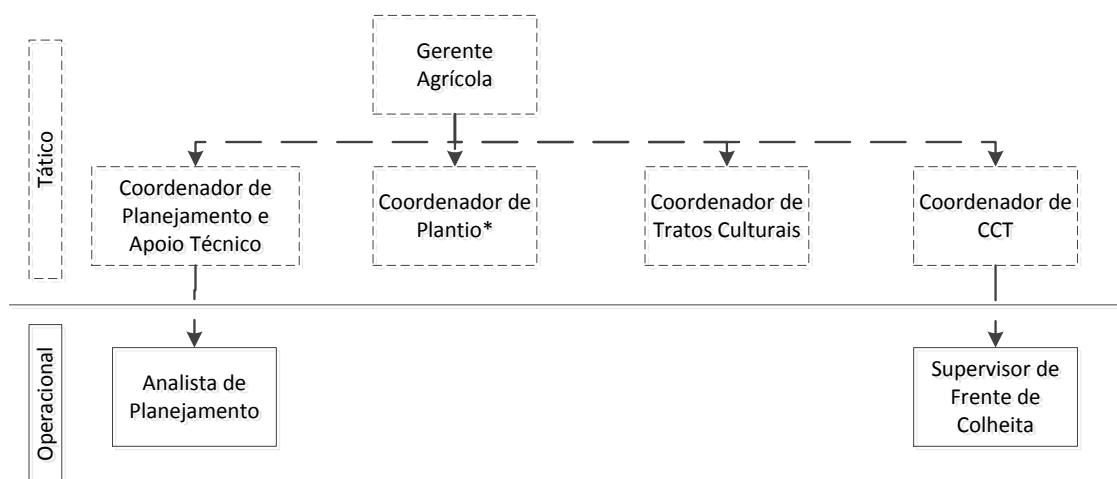
<sup>1</sup> O planejamento em onda é uma estratégia de planejamento onde é delineado um plano geral e feito o planejamento para o próximo período de tempo para o qual se tem segurança e clareza sobre os parâmetros do projeto (próxima semana ou alguns meses). Ao longo do tempo, novas etapas do plano são detalhadas de forma iterativa e progressiva.

variabilidade dos sistemas, destacada anteriormente. Além de negligenciar a característica estocástica, o uso de médias históricas assume um comportamento histórico médio do clima e deixa de utilizar informações meteorológicas de grande potencial para agricultura (HOOGENBOOM, 2000). Cabe destacar que o uso de modelos não garante uma resposta estocástica de produtividade, porém o uso de diversos cenários de clima permite estimar um comportamento médio e sua variação, o que pode ser visto nas aplicações descritas por Everingham et al. (2002) e Bezuidenhout e Singels (2007). No Brasil o uso DSSAT/CANEGRO pode ser visto nos trabalhos de Nassif et al. (2012) e Marin et al. (2011), referentes a parametrização do modelo, enquanto o uso em conjunto com cenários de clima futuro para quantificação do impacto da mudança climática na produtividade de cana-de-açúcar pode ser visto em Marin et al. (2013).

Dada a relevância das predições de produtividade no planejamento na cadeia de valor do açúcar, e potencialmente para as outras cadeias de valor do complexo da cana e o potencial do uso de projeções de produtividade na agricultura e em especial para a cana-de-açúcar, o presente trabalho se propõe a caracterizar os contextos de tomada de decisão, nos quais a previsão de produtividade é utilizada, e a partir dessa caracterização, propor um framework. Entende-se que esse resultado irá ampliar o entendimento da tomada de decisão na cadeia de valor do açúcar, de uma forma análoga ao apresentado por Everingham et al. (2002) para a cadeia do valor do açúcar, mas ultrapassando a tomada de decisão baseada em projeções climáticas e incorporando as decisões que potencialmente podem se beneficiar por modelos de produtividade e modelos de planejamento. Além disso, o presente trabalho fornecerá características importantes das decisões conforme Meinke e Stone (2005).

### 3.3 Metodologia

O presente trabalho se enquadra como uma pesquisa exploratória descritiva conduzida a partir de um estudo de caso baseado em entrevistas semiestruturadas e acompanhamento de atividades de planejamento em uma usina do setor sucroenergético. Os profissionais entrevistados foram o gerente agrícola, os coordenadores de Planejamento e Desenvolvimento Vegetal, Tratos Culturais e de Corte Carregamento e Transporte (CCT), considerados no nível tático. Também foi entrevistado o Supervisor de Frentes de colheita e um Analista de Planejamento, no nível operacional. O estudo de caso foi conduzido na unidade Alcídia da Odebrecht Agroindustrial, localizada no município de Teodoro Sampaio – SP. As práticas de planejamento da unidade podem ser extrapoladas para as demais unidades da empresa (9 no total), o que equivale a um processamento de 24 milhões de toneladas de cana-de-açúcar por ano em quatro estados brasileiros. Uma ilustração parcial do organograma pode ser vista na Figura 9 onde pode ser visualizado como se distribuem hierarquicamente os profissionais entrevistados



**Figura 9. Estrutura e nível organizacional dos profissionais entrevistados. (\* Não foi entrevistado)**

Para caracterização, buscou-se preencher um quadro que relacionava diferentes aspectos relevantes das decisões, de acordo com a Tabela 1, onde consta para cada decisão levantada (1), qual seu papel para organização (DANESE; KALCHSCHMIDT, 2011), representado como finalidade (2), a qual está associada à área funcional, uma característica também levantada por Ahumada e Villalobos (2009). Além desta característica apontada pelos autores, foi identificado o nível organizacional (3) da decisão. Outras características levantadas foram a escala da projeção/estimativa (4), e relações entre a época da realização da projeção (5) e horizonte de

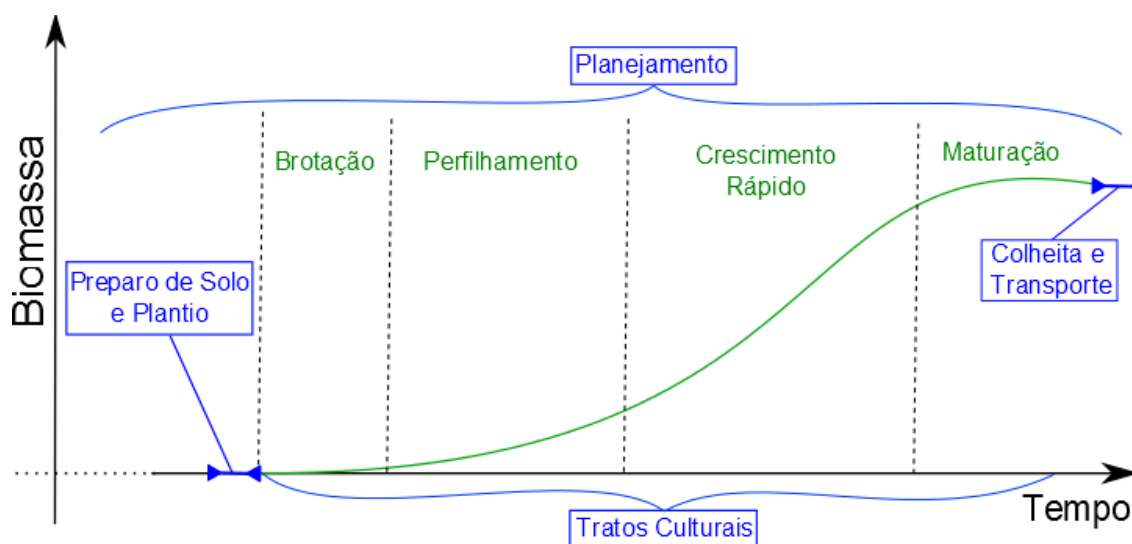
planejamento/decisão (6) conforme Meinke e Stone (2005). Por fim, buscou-se caracterizar o (7) uso de modelagem (de produtividade ou planejamento) utilizado em cada decisão. Cabe destacar que no contexto australiano, Everingham et al. (2002) apontaram as decisões em diferentes elos da cadeia do açúcar, o que no Brasil terá uma equivalência para as diferentes áreas funcionais da usina. Essa caracterização é entendida como uma aplicação dos 6 “W’s” - “*What?*/O que?”, “*When?*/Quando?”, “*Why?*/Por quê?”, “*Who?*/Quem?”, “*Where?*/Quando?” e “*How?*/Como?”. Além do preenchimento do quadro, buscou-se entender como as diferentes decisões se relacionam.

**Tabela 1. Cabeçalho da tabela utilizada nas entrevistas.**

<sup>(1)</sup> <b>Aplicação</b>	<sup>(2)</sup> <b>Finalidade</b>	<sup>(3)</sup> <b>Nível Organizacional</b>	<sup>(4)</sup> <b>Escala</b>	<sup>(5)</sup> <b>Quando</b>	<sup>(6)</sup> <b>Janela temporal</b>	<sup>(7)</sup> <b>Modelo</b>
What?	Why?/Who?	Who?	Where?	When?		How?

### 3.4 A realização de projeções e estimativas na usina

Diferentes profissionais interagem com a cultura em diferentes estádios de crescimento. De forma simplificada, essa interação é ilustrada na Figura 10. Profissionais da área de planejamento determinam a implantação da cultura, i.e., plantio, porém interagem com as áreas durante todo o ciclo de crescimento. Após a brotação da cultura, temos os profissionais da área de tratos culturais, que atuam nas diversas operações durante o crescimento da cultura. No último estágio, participam os profissionais da área de Corte, Carregamento e Transporte (CCT), responsáveis pela colheita e logística de transporte até a unidade industrial.



**Figura 10.** O ciclo de crescimento da cana-de-açúcar no primeiro ciclo e sua relação com as diferentes áreas de atuação na usina. Nos ciclos posteriores, com a rebrota da cana, não há interação dos profissionais de preparo de solo e plantio.

A realização de predições na usina é de responsabilidade do setor de planejamento agrícola, com destaque para o coordenador de planejamento. Esse profissional, considerado um especialista desta atividade, realiza as predições de acordo com seu conhecimento tácito quanto ao crescimento da cana-de-açúcar e dos fatores que afetam a produtividade. Na entrevista, na qual esse especialista foi questionado sobre a forma como realiza projeções, foi constatado que ela é análoga à hierarquia de fatores limitantes da produtividade utilizada em diversos modelos de crescimento de plantas, conforme descrito, por exemplo, por van Ittersum et al. (2003), na qual a produtividade potencial é minorada por uma sequência de fatores. De acordo com o especialista, sua primeira projeção, para um cultivo no seu 1º corte, leva em consideração a interação entre Cultivar x Ambiente de Produção x Época da Colheita. Naturalmente, produtividades mais altas



são esperadas para os ambientes mais favoráveis. Para época da colheita, a produtividade esperada é maior para as áreas colhidas no começo do ano e menor para as áreas colhidas no fim do ano. Esse resultado é apontado como consequência da disponibilidade hídrica em diferentes estádios de crescimento da cana. É possível relacionar esse resultado com o trabalho de Inman-Bamber e Smith (2005) e o regime de chuvas da região. Até meados de abril, a cana tem maior disponibilidade de água na sua fase de crescimento rápido, que é crítica para a produtividade final, uma vez que essa fase permite o aproveitamento das chuvas do verão local. Nas áreas de colheita tardia (de agosto a dezembro), os maiores volumes de água são disponibilizados no início do crescimento, fase que não demanda um volume elevado devido a sua baixa evapotranspiração, enquanto sua fase de crescimento intenso se dará no inverno do ano seguinte, caracterizado por menor disponibilidade hídrica, o que impacta negativamente na produtividade. Ao incluir a variedade nesta análise, o conhecimento implícito do especialista se mostra alinhado com as estratégias usadas por melhoristas e pesquisadores para melhorar e entender os fatores que condicionam a produtividade (RAMBURAN et al., 2011; SPIERTZ, 2013). O especialista relaciona então diferentes produtividades para os diferentes ciclos iniciais, diferenciando canas-planta em diferentes épocas e, portanto, submetidas a diferentes regimes hídricos (canas de 18 e 12 meses e cana de inverno). Essa projeção inicial incorpora os dois primeiros níveis hierárquicos na escala de fatores que afetam a produtividade, seguindo novamente a mesma estrutura dos modelos de crescimento (VAN ITTERSUM et al., 2003). A partir deste ponto, o especialista minora a produtividade baseado em seu conhecimento do histórico da área, relacionando também a ocorrência de pragas e plantas invasoras.

Do segundo corte em diante, a produtividade é estimada baseada na produtividade inicial e na ocorrência de fatores e/ou adversidades fitossanitárias, manejo inadequado como pisoteio e arranquio, assim como as condições da lavoura e condições de clima que afetarão a produtividade futura, como veranicos, geadas ou diminuição do fotoperíodo. Para cada ocorrência de situação adversa ou interação de fatores, o especialista estima a variação absoluta da TCH.

Conforme constatado na entrevista com os coordenadores das áreas de tratos culturais e CCT, a interação destes profissionais com os canaviais em diferentes estádios permite a eles adquirirem conhecimento tácito em relação ao crescimento da cana-de-açúcar e formação da produtividade final, se tornando capazes de realizar projeções ou estimativas quanto à

produtividade, a depender do estágio em que atuam. Assim, profissionais ligados ao CCT, por interagirem com áreas no momento da colheita, não precisam realizar uma projeção do resultado e são capazes de estimar a produtividade de talhões. Esses profissionais destacaram que suas estimativas se baseiam principalmente em características como o diâmetro, altura e número de plantas por metro. Assim, enquanto a produtividade está correlacionada com o diâmetro dos colmos, canaviais mais densos (*“fechados”*), possuem mais plantas e, portanto, maior produtividade. De forma análoga, a presença de falhas nas linhas de cana é considerada para minorar a produtividade. Cabe destacar que essas inspeções visuais são feitas no contorno dos talhões, sendo possível que a heterogeneidade das áreas e efeitos de borda interfiram nessas estimativas. Para os profissionais ligados aos tratos culturais, por se situarem temporalmente nos estágios intermediários de crescimento, é possível considerarem as condições da cultura e extrapolar os resultados para o futuro de acordo com uma data estimada para a época de colheita. A troca de informações entre os profissionais de planejamento, tratos culturais e CCT permite que divergências entre as previsões anteriores sejam confrontadas entre si e com outras informações. Uma vez que essas estimativas e projeções são dependentes do conhecimento tácito dos profissionais, não é possível extrapolar esses resultados para outros contextos.

### **3.5 A tomada de decisão e o planejamento**

#### **3.5.1 Orçamento, Plano de Safra e Estratégia Comercial**

Antes do fim da safra e condicionado por objetivos estratégicos, em outubro é realizada uma primeira projeção para o Orçamento, o Plano de Safra e a Estratégia Comercial da safra seguinte. Esse esforço mobiliza dentro da cadeia de valor os processos relacionados à produção agrícola e industrial, em conjunto com a área comercial, condicionados pela área financeira. Para o plano de safra são atribuídas diferentes produtividades às diferentes áreas da usina em função do manejo e das condições dos canaviais e uma perspectiva de mês de colheita. O plano de safra tem por objetivo garantir as condições de fornecimento de matéria-prima de acordo com capacidade de processamento industrial. Para isso, devem ser mobilizados recursos (Orçamento) compatíveis à Estratégia Comercial, uma vez que já é possível estimar a produção de açúcar e etanol. A elaboração conjunta do plano de safra, orçamento e estratégias de venda é iterativa, sendo necessária uma compatibilização entre as despesas, produção e receitas. Cabe destacar que essa decisão é tomada na maior condição de incerteza, pois algumas áreas sequer foram colhidas e é necessário estimar sua próxima produtividade. Para o Plano de Safra, novas projeções são realizadas em Janeiro. Esse planejamento, que pode ser considerado o de maior escopo dentro da usina. A realização do planejamento da próxima safra antes do fim da safra vigente é análoga à constatada nas condições da África do Sul e Austrália (BEZUIDENHOUT; SINGELS, 2007; EVERINGHAM et al., 2002). O fato da informação sobre a produtividade permear toda a cadeia é evidente, dado que a produtividade determina a matéria-prima disponível, que deverá ser manipulada, transportada e processada. Everingham et al. (2002) destacou o questionamento quanto ao volume de produção ao longo de toda a cadeia. Um sistema de otimização da colheita é utilizado como ferramenta de suporte a decisão. A estratégia de modelagem é de programação por restrições combinada a programação linear ou inteira, de acordo com a decisão, e utiliza o conjunto de dados da própria usina para projetar a produtividade e qualidade industrial médias para interações de números de corte, ambiente de produção, variedade e época de colheita.

Chama atenção o uso de um modelo de programação por restrições, o que não foi contemplado no estudo de Ahumada e Villalobos (2009). Lustig e Puget (2001) já apontavam para um uso crescente da programação por restrições na pesquisa operacional, em especial para problemas de sequenciamento e agendamento (*scheduling*, no original). Cabe destacar também o

uso de médias no planejamento em uma ferramenta de modelagem determinística. De acordo com Higgins et al. (2007) e Ahumada e Villalobos (2009), a representação estocástica dos sistemas agroindustriais é mais fiel ao comportamento dos mesmos.

Ao longo do ano, são realizadas duas novas projeções, sendo uma em maio e outra em setembro. Nesses momentos, com o avanço dos cultivos, novas inspeções de campo produzem resultados melhores, além de serem agregadas informações de outras áreas que já foram colhidas, informações de clima intermediadas por um especialista na área e inventários de biomassa obtidos através de sensoriamento remoto. Esses dados auxiliam para refinar as projeções e incorporarem mudanças necessárias no planejamento. Devido às necessidades impostas pela empresa, as diferentes usinas do grupo devem realizar o plano de safra na mesma época e contam principalmente com o conhecimento tácito desses diferentes profissionais, cada um ligado à sua unidade e área de atuação. O uso das médias de produção, apresentam as limitações já destacadas quanto a sua negligência de informações de clima futuro e por assumirem implicitamente a ocorrência de um clima de comportamento médio.

### **3.5.2 Plano de Colheita e Roteirização**

Embora o plano de safra apresente as épocas de colheita dos talhões, essas épocas são refinadas no plano de colheita. Esse novo plano, realizado em onda e com horizonte de uma semana, é feito semanalmente, e visa corrigir diferenças entre as expectativas e as condições constatadas em campo, além de contornar situações adversas como, por exemplo, geada, chuvas excessivas, florescimento ou queimadas. Essa tomada de decisão é descrita para outra usina brasileira por Jena e Poggi (2013). Com a presença de diferentes agentes para produção, colheita e transporte e moagem, Jiao et al. (2005) e Higgins (2002) descrevem o plano de colheita para Austrália, agregando as peculiaridades ligadas aos objetivos dos diferentes agentes ao longo da cadeia. Com a definição do plano de colheita da semana, são tomadas decisões diárias quanto à roteirização das frentes de colheitas que devem minimizar o deslocamento e manter o fluxo de matéria-prima ao longo do dia. O plano de colheita está contemplado na ferramenta utilizada para otimização do plano de safra.

### **3.5.3 Renovação e Contratos para Expansão**

Com o decréscimo da produtividade da cana-de-açúcar, é necessário replantar os canaviais. Uma vez que a produtividade é geralmente inferior a cada safra, essa decisão é tomada

quando uma área não apresenta perspectiva de produtividade rentável. No geral, as áreas são replantadas a cada cinco safras, ocorrendo variações. As decisões quanto à renovação são agregadas ao plano de safra, pois está ligada ao fornecimento de matéria-prima.

Em um contexto de aumento da produção agroindustrial, uma determinação estratégica, é necessário incorporar novas áreas à produção agrícola. Após a decisão da expansão em nível estratégico, diferentes áreas com potencial para incorporação são avaliadas pelo nível tático em relação a perspectiva de produção para, no mínimo, os próximos cinco ciclos, para subsidiar a mediação dos contratos. Além da produtividade, são consideradas também a distância, condição e uso atual da área, assim como o tempo de contrato. Cabe destacar que o aumento da oferta de matéria-prima também pode ser obtido com a intensificação dos cultivos, cabendo analisar as diferentes linhas de ação. Para este contexto de tomada de decisão, também está disponível um sistema de suporte a tomada de decisão que otimiza a alocação de variedades para o ambiente de produção e avalia a necessidade de renovação, dada as necessidades do plano de safra e a parametrização financeira disponível. O sistema de suporte considera as produtividades médias parametrizadas em função do número de cortes, variedade, ambiente de produção e época de colheita.

#### **3.5.4 Manejo**

Outras decisões identificadas que são pautadas pela projeção de produtividade, são o manejo fitossanitário (controle de nematóides e migdolus) e a aplicação de maturador, as quais são tomadas sob demanda, pelos responsáveis da área de planejamento e desenvolvimento vegetal, não estando ligadas ao planejamento de safra, mas sim às práticas agronômicas consolidadas na usina. Outros manejos fitossanitários seguem práticas agronômicas de controle de infestação baseado no nível de infestação. Não foi identificado entre as decisões o questionamento da produtividade futura para subsidiar a tomada de decisão quanto às estratégias de adubação. Everingham et al. (2002) apresentam a decisão quanto à adubação como uma das decisões que deve ser tomada no processo de produção, enquanto Hoogenboom (2000) aponta-a como uma das principais decisões táticas que podem se beneficiar do uso de modelos de predição de produtividade utilizados em conjunto com projeções de clima. Limitações na disponibilidade de nitrogênio e água são os primeiros fatores a minorar a produtividade e o uso dos dois pela planta depende da quantidade disponível de ambos, uma vez que a planta responde à adubação de

nitrogênio de acordo com a disponibilidade de água (WIEDENFELD, 1995), fazendo com que estes devam estar disponíveis em quantidade compatível para determinado patamar de produção. Quando questionados sobre a estratégia de adubação, os entrevistados responderam que a mesma segue um conjunto de práticas pré-estabelecidas, na qual a aplicação é proporcional à produtividade esperada. Uma vez que a adubação é realizada no plantio e após a colheita de cada talhão, esta decisão considera a predição de produtividade mais recente.

Entre os questionamentos relacionados ao manejo apresentados pelos autores (EVERINGHAM et al., 2002) para o processo de produção, nenhum foi mencionado nas entrevistas. Esse fato pode ser explicado pela motivação dos autores que foi pautada pela influência do clima e da precipitação e não necessariamente ligados à produtividade. As decisões relacionadas à etapa de cultivo, influenciadas pela produtividade estão relacionadas à aplicação de maturador e controle de pragas, que devido ao custo, são realizadas em áreas de maior produtividade e, portanto, de maior retorno econômico. De acordo com Hoogenboom (2000), ao discutir o uso de modelos de crescimento para previsão de produtividade em culturas de sequeiro, além da adubação poucas decisões podem ser tomadas na etapa de produção depois de decididas as condições de implantação da cultura. Porém, um melhor conhecimento da produtividade terá efeito nas etapas seguintes da cadeia, fazendo com que essa informação, agregada de talhão em talhão possa beneficiar a usina.

### **3.5.5 Síntese**

O resultado obtido no levantamento junto à usina foi sintetizado na Tabela 2. A coluna relacionada à modelagem foi omitida, dado que a projeção de produtividade é realizada pelos especialistas de planejamento. Entende-se que o modelo utilizado é um modelo tácito, dependente do conhecimento e experiência dos profissionais envolvidos. Dados históricos são utilizados para projetar a produtividade no sistema de planejamento disponível para auxílio a tomada de decisão.

**Tabela 2. Caracterização da tomada de decisão.**

<b>Aplicação</b>	<b>Finalidade</b>	<b>Nível Organizacional</b>	<b>Escala</b>	<b>Quando</b>	<b>Janela temporal</b>
<b>Orçamento</b>	Quantificar e disponibilizar recursos	Tático	Geral	Out- Jan da safra anterior	Safr (~ 9meses)
<b>Estratégia Comercial</b>	Vender e Estocar				
<b>Plano de safra</b>	Início e fim de safra e época de colheita dos diferentes talhões.				
<b>Renovação e Contratos de Expansão</b>	Fornecimento rentável de matéria prima.	Tático	Fazendas	Durante a safra	5 anos*
<b>Plano de colheita</b>	Determinar a data em que áreas serão colhidas.	Operacional	Geral	Semanal	Semanal
<b>Roteirização</b>	Organizar logística de colheita e ajuste diante de imprevistos	Operacional	Fazendas	Diário	Diário
<b>Aplicação de maturador</b>	Aumentar qualidade industrial dos talhões	Operacional	Talhão	Início da safra	20 a 40 dias
<b>Aplicação de defensivos</b>	Controle químico de Migdolus Controle químico de Nematóides	Operacional	Talhão	Verão Colheita/ Renovação	Próxima safra

\*No caso de expansão, a decisão estratégica é tomada em vista de intervalos mais longos, porém a análise de das áreas considera, em geral, o ciclo típico de 5 anos, sendo as áreas incorporadas ao longo do safra.

A distribuição das decisões nos níveis organizacionais e ao longo da cadeia de valor pode ser vista na Figura 11. Cabe notar a concentração de decisões na etapa de cultivo, reflexo das entrevistas a profissionais da área agrícola. Um fato importante a ser notado é a ausência de decisões ligadas a Moagem e Produção, o que corresponderia a área industrial. Devido às características da indústria, suas decisões estão mais ligadas ao fluxo diário de matéria prima e à quantidade total que será processado na safra. Evidentemente, a quantidade de matéria fornecida é função da produtividade e área dos talhões colhidos, porém cabe aos profissionais de CCT mediar essa relação entre a produtividade dos talhões e o fluxo de matéria prima para usina. Raciocínio similar se aplica às estratégias de vendas, onde a produtividade de cada talhão contribui para uma projeção da quantidade total de matéria- prima disponível que será utilizada para produção de açúcar e etanol.

	Insumos	Cultivo	Colheita e Transporte	Moagem e Produção	Vendas
Tático	Orçamento	Contratos para Expansão Renovação Plano de Safra			Estratégia Comercial
Operacional		Controle Químico Aplicação de Maturador	Plano de Colheita Roteirização		

↑ Janela temporal

**Figura 11.** Distribuição das decisões nos diferentes níveis organizacionais (linha tracejada) ao longo da cadeia de valor (horizontal), estando as decisões de maior janela temporal no topo da tabela e as demais ordenadas de forma decrescente. Essa figura sintetiza como as decisões da Tabela 2 se relacionam com as Figuras 8 e 9.

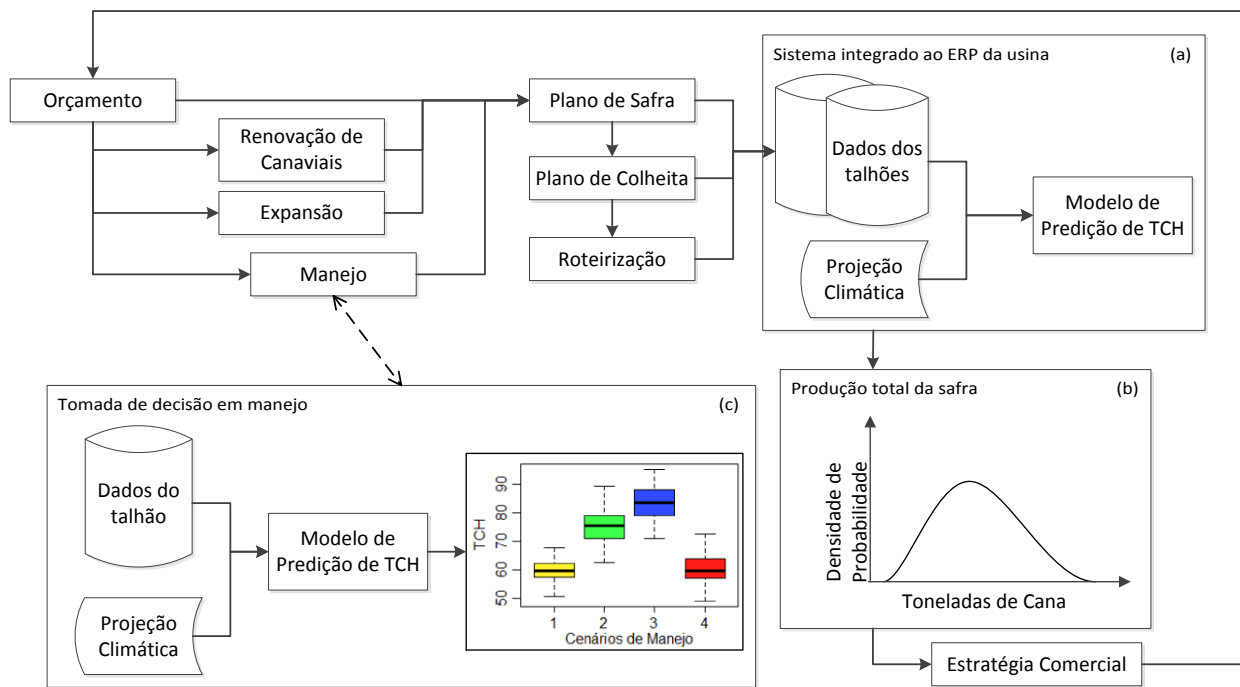
### 3.6 Proposta de Framework

A importância da predição da produtividade se dá pelo fato de que as variações da mesma atravessam os processos produtivos da cadeia de valor do açúcar, não se limitando ao processo ou etapa do processo onde a produtividade subsidia uma decisão. Dessa forma, a importância da predição da produtividade deixa de ser apenas o efeito na decisão em questão, mas também seus efeitos nos processos posteriores, criando a necessidade de que a análise da decisão quanto à produtividade também incorpore as outras etapas. Em síntese, o processo de tomada de decisão deve integrar o resultado das decisões na escala em que forem tomadas e se incorporando os resultados com as etapas posteriores. A incorporação dos modelos em uma ferramenta de tomada de decisão pode ser entendida como o desenvolvimento de um Sistema de Suporte à Decisão (SSD) baseado em modelos (POWER; SHARDA, 2007). Em uma etapa posterior, com a modelagem e adoção de técnicas de otimização, todas essas informações poderiam estar agregadas em um modelo de planejamento agrícola, análogo aos descritos por Ahumada e Villalobos (2009). Para subsidiar tanto a modelagem empírica quanto a modelagem para otimização do planejamento, a usina possui dados relativos à produção de cada talhão, sistematizados em um ERP<sup>2</sup>. O uso desses dados para tomada de decisões é defendido por Lawes e Lawn (2005), que mostram aplicações desses dados em tomadas de decisão, de forma complementar a outras iniciativas mais convencionais na pesquisa agronômica.

<sup>2</sup> *Enterprise Resource Planning* – Sistemas de suporte a decisão que integram as diversas áreas funcionais e permitem a gestão dos processos de uma empresa.

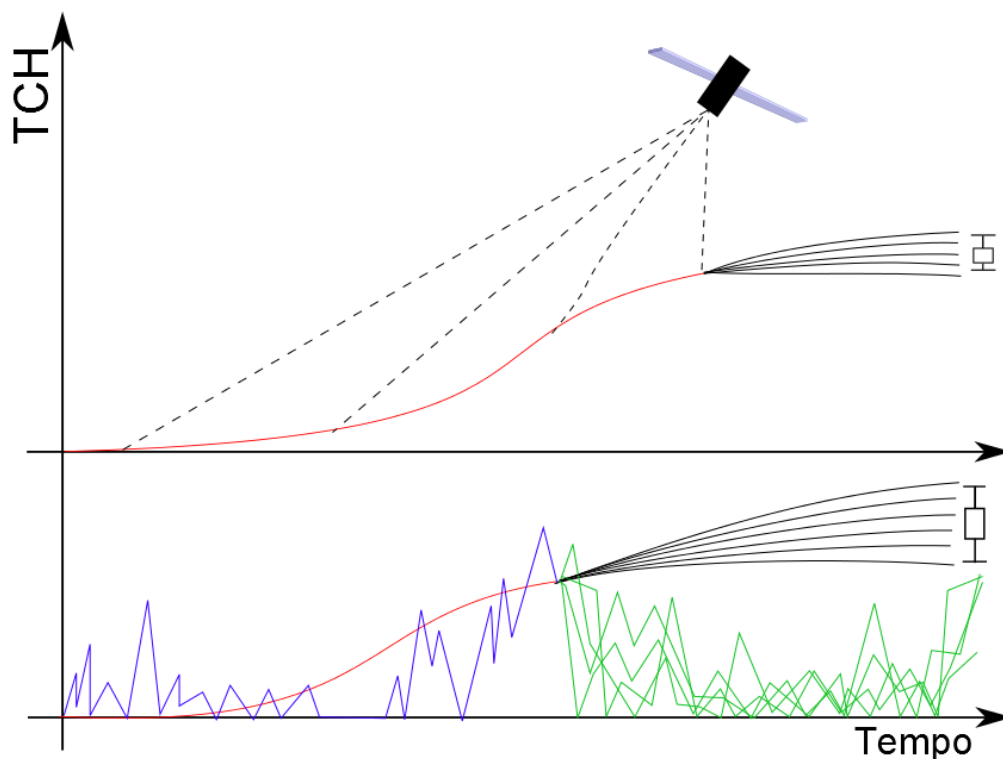


De forma geral, essa proposta se organizaria de acordo com o encadeamento apresentado na Figura 12. Devido à existência de um ERP, seria essencial que essa abordagem de modelagem fosse integrada ao sistema existente, funcionando como um módulo agregado e não como outro sistema, se comportando como um SSD baseado em ERP (STANEK et al., 2004). Essa integração ao sistema existente, junto a uma modelagem da cadeia facilitaria que ao realizar novas projeções ou estimativas, o impacto em outras áreas fosse visualizado. Antecipar essa concepção de sistema é uma forma de facilitar a convergência entre o SSD e o ERP, o que necessita de uma aproximação entre os fornecedores de ERP e dos criadores de SSD (SHARIAT; NWAKANMA, 2006). Um exemplo do que poderia ser proporcionado pela integração é: ao constatar que um bloco de talhões não vai atingir a produtividade esperada, a informação de menor oferta de matéria prima daquela área já fosse incorporada para o planejamento da colheita, uma vez que é necessário manter a moagem. Enquanto desvios de pequena escala possam ser contornados, sucessivos desvios consistentes podem interferir de forma agregada no resultado final, sendo necessária uma ferramenta que integre essas informações para as outras etapas da cadeia. Conforme observado durante o acompanhamento das atividades, áreas funcionais próximas como o planejamento agrícola e a área logística apresentam uma integração baseada na interação dos profissionais. Essa integração é menor com a área industrial e menor ainda com a área comercial. Essa proposta se alinha com a necessidade apontada por Bezuidenhout e Baier (2009) de melhor integração entre as diversas cadeias funcionais. Neste caso, o sistema promoveria uma melhor integração entre a cadeia de informações e a cadeia de agregação de valor, com ganhos na cadeia de transporte e manipulação de materiais (*material handling*).



**Figura 12. Encadeamento do Framework proposto para integração, automação e acompanhamento. Após a definição dos parâmetros dos planos, as informações dos talhões em conjunto com modelos de produtividade são utilizadas para projeção da produtividade (a). Os resultados, que devem ser expressados refletindo a incerteza da projeção podem ser representados através da probabilidade associada (b) ou boxplot (c), entre outros. O uso dessas informações por modelos de planejamento pode ser mediado pelo framework.**

A estratégia de modelagem deve ser tal que permita explorar os diferentes estádios de informação e crescimento dos diferentes talhões. Nessa proposta seria possível incorporar a substituição sucessiva da série de clima projetada (HOOGENBOOM, 2000) e inclusão dos dados de sensoriamento remoto, conforme ilustrado na Figura 13. Em uma etapa posterior, modelos de planejamento poderiam otimizar o planejamento. A integração necessária dos diferentes planos, seja o de safra, de colheita ou estratégias comerciais, faz com que quando novas informações estejam disponíveis, os planos estabelecidos devam ser reavaliados e readequados se necessário. Por gerar as informações e subsidiar a tomada de decisão na unidade fundamental de manejo, o talhão, o sistema permite incorporar o efeito de diversas decisões na predição da produtividade global.



**Figura 13.** Na parte superior, dados de sensoriamento em diversos momentos são utilizados para melhorar as projeções de produtividade. Na parte inferior, é ilustrada a projeção de produtividade em função do clima ocorrido para uma parte do ciclo e para as diversas perspectivas futuras, estando ilustrada a série de dados de precipitação (azul) e diferentes perspectivas de clima (verde).

Foi consenso entre os entrevistados a importância da informação de produtividade no planejamento e foi manifestada uma perspectiva de usufruir de uma melhor qualidade da mesma. Nesse sentido, diversas técnicas e estratégias foram criadas e utilizadas pela comunidade científica para prever ou melhorar a previsão da produtividade, assim como foram utilizadas diversas técnicas de modelagem em pesquisa operacional para melhorar a performance das operações, sendo diversas específicas para a cadeia de valor do açúcar, enquanto outras, aplicadas em outras cadeias possam também ser extrapoladas. Como exemplo, têm-se as relações e analogias entre os problemas logísticos em cana-de-açúcar e em sistemas florestais (WEINTRAUB; ROMERO, 2006). Embora diversas soluções estejam disponíveis, sua adoção mesmo por grandes grupos não pode ser entendida como um consenso, conforme os resultados desse estudo de caso.

De forma geral, o framework apresentado se baseia em três pontos principais: a integração dos diferentes planos sobre uma base de informações comuns, a adoção de técnicas de modelagem empírica para predição de produtividade e o uso de modelos de otimização para o planejamento agrícola. O uso de uma base de informações comum para projeções foi apresentado por Ahumada e Villalobos (2009) como uma das variáveis do processo de projeção que tem efeito positivo para companhias de manufatura e não há motivos para supor que isso seja diferente na cadeia de valor sucroenergética. Os autores também propõem que a coleta de informações venha de diferentes fontes, como consumidores, cenário econômico e fornecedores. Como na cadeia de valor do açúcar para condições brasileiras temos um agente único para a cadeia de valor da produção à venda, os clientes e fornecedores são as próprias áreas funcionais da usina, fazendo com que obter informações de clientes e fornecedores dentro da cadeia de valor seja na verdade o uso comum de informações de projeção e demanda. A obtenção de informações sobre a economia e mercado do açúcar, também serão importantes para se adequar a projeção de vendas, ponto onde a cadeia deixa de estar sob controle da usina. Ainda no sentido de obter informações de outras fontes, entendemos que isso também é atingido com o uso de imagens de sensoriamento remoto e projeções de clima. O uso da modelagem para estimar produtividade, além de já ser aplicado para algumas decisões (BEZUIDENHOUT; SINGELS, 2007; EVERINGHAM et al., 2002), é indicado como uma das formas de melhorar as informações para modelos de planejamento (LE GAL et al., 2009; PIEWTHONGNGAM et al., 2009).

### 3.7 Conclusões

Ao investigar o uso de predições para tomada de decisão e planejamento na cadeia de valor do açúcar para uma usina localizada no centro-sul brasileiro, foi possível traçar paralelos com a descrição destes processos para África do Sul, Austrália, Tailândia e Venezuela, assim como constatar que não são utilizadas ferramentas de modelagem para predição da produtividade. No encadeamento das diversas decisões, a tomada de decisão pautada por projeções de produtividade em Agosto para elaboração do orçamento, com consequência no plano de safra e estratégia comercial, é considerado um ponto crítico, dado o seu efeito sobre toda a cadeia e a elevada incerteza associada, consequência da incerteza climática. O refinamento do plano de safra em Janeiro permite melhorar a qualidade das projeções e suas consequência na estratégia comercial, porém ainda conta com a antecipação de resultados de até 10 meses

Diante das diversas soluções isoladas encontradas na literatura, foi apresentada uma proposta de framework para tomada de decisão e planejamento. Esse framework, baseado em modelagem empírica a partir dos dados de talhões, se beneficiaria do uso de modelos de produtividade e planejamento para tomada de decisão por diferentes gestores na cadeia. Ao mesmo tempo, também seria responsável por integrar as diferentes predições e automatizar a realização de novas predições assim que novas informações estejam disponíveis. Essa proposta pode ser vista como uma tentativa de orientar o desenvolvimento de soluções para a cadeia de valor do açúcar considerando os avanços em diversas áreas de pesquisa. O volume produzido pela empresa justifica a adoção dessas soluções. Cabe destacar que o volume produzido (24 milhões de toneladas) é superior ao de países que já implementaram soluções de predição de safra baseada em modelos (África do Sul e Austrália) (FAO, 2013).

### **3.8 Recomendações e trabalhos futuros**

Em trabalhos futuros, pode ser utilizada a lista de decisões apresentada neste trabalho na forma de um questionário a ser conduzido no setor. Além de medir a adesão desse framework de decisões em outras unidades, seria possível mensurar as variações desse processo e relacionar esse resultado com o desempenho operacional das usinas, em um trabalho análogo ao conduzido por Danese e Kalchschmidt (2011). Nesse trabalho, também poderia ser medido o uso de tecnologias de sensoriamento remoto, técnicas de modelagem, ferramentas e modelos de planejamento e uso de projeções climáticas.

### 3.9 Referências

- ABDEL-RAHMAN, E. M.; AHMED, F. B. The application of remote sensing techniques to sugarcane ( *Saccharum* spp. hybrid) production: a review of the literature. **International Journal of Remote Sensing**, v. 29, n. 13, p. 3753–3767. doi: 10.1080/01431160701874603, 2008.
- AHUMADA, O.; VILLALOBOS, J. R. Application of planning models in the agri-food supply chain: A review. **European Journal of Operational Research**, v. 195, n. 1, p. 1–20. doi: 10.1016/j.ejor.2008.02.014, 2009.
- BEZUIDENHOUT, C.; BAIER, T. A global review and synthesis of literature pertaining to integrated sugarcane production systems. Proceedings of the 82nd Annual Congress of the South African Sugar Technologists' Association, Durban, South Africa, 26-28 August 2009. **Anais...** p.93–101, 2009.
- BEZUIDENHOUT, C. N.; SINGELS, A. Operational forecasting of South African sugarcane 7production: Part 1 – System description. **Agricultural Systems**, v. 92, n. 1-3, p. 23–38. doi: 10.1016/j.agsy.2006.02.001, 2007.
- CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da Safra Brasileira: Cana-de-Açúcar**, 2º Levantamento. Brasília, 2013.
- DANESE, P.; KALCHSCHMIDT, M. The role of the forecasting process in improving forecast accuracy and operational performance. **International Journal of Production Economics**, v. 131, n. 1, p. 204–214. doi: 10.1016/j.ijpe.2010.09.006, 2011.
- EVERINGHAM, Y. .; MUCHOW, R. .; STONE, R. .; et al. Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. **Agricultural Systems**, v. 74, n. 3, p. 459–477. doi: 10.1016/S0308-521X(02)00050-1, 2002.
- 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, n. 3-4, p. 689–696. doi: 10.1016/j.agrformet.2008.10.018, 2009.
- FAO - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. FAOSTAT database. . Retrieved from [http://faostat3.fao.org/faostat-gateway/go/to/browse/rankings/countries\\_by\\_commodity/E](http://faostat3.fao.org/faostat-gateway/go/to/browse/rankings/countries_by_commodity/E), 2013.
- FRANCO, H.C.J., OTTO, R., FARONI, C.E., VITTI, A.C., ALMEIDA DE OLIVEIRA, E.C., TRIVELIN, P.C.O., 2011. Nitrogen in sugarcane derived from fertilizer under Brazilian field conditions. **Field Crops Research** 121, 29–41. doi:10.1016/j.fcr.2010.11.011
- LE GAL, P.-Y.; LE MASSON, J.; BEZUIDENHOUT, C. N.; LAGRANGE, L. F. Coupled modelling of sugarcane supply planning and logistics as a management tool. **Computers and Electronics in Agriculture**, v. 68, n. 2, p. 168–177. doi: 10.1016/j.compag.2009.05.006, 2009.
- GRUNOW, M.; GÜNTHER, H.-O.; WESTINNER, R. Supply optimization for the production of raw sugar. **International Journal of Production Economics**, v. 110, n. 1-2, p. 224–239. doi: 10.1016/j.ijpe.2007.02.019, 2007.

HIGGINS, A. J. Australian Sugar Mills Optimize Harvester Rosters to Improve Production. **Interfaces**, v. 32, n. 3, p. 15–25. doi: 10.1287/inte.32.3.15.41, 2002.

HIGGINS, A.; THORBURN, P.; ARCHER, A.; JAKKU, E. Opportunities for value chain research in sugar industries. **Agricultural Systems**, v. 94, n. 3, p. 611–621. doi: 10.1016/j.agsy.2007.02.011, 2007.

HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. **Agricultural and Forest Meteorology**, v. 103, n. 1-2, p. 137–157. doi: 10.1016/S0168-1923(00)00108-8, 2000.

INMAN-BAMBER, N. G.; SMITH, D. M. Water relations in sugarcane and response to water deficits. **Field Crops Research**, v. 92, n. 2-3, p. 185–202. doi: 10.1016/j.fcr.2005.01.023, 2005.

VAN ITTERSUM, M. .; LEFFELAAR, P. .; VAN KEULEN, H.; et al. On approaches and applications of the Wageningen crop models. **European Journal of Agronomy**, v. 18, n. 3-4, p. 201–234. doi: 10.1016/S1161-0301(02)00106-5, 2003.

JENA, S. D.; POGGI, M. Harvest planning in the Brazilian sugar cane industry via mixed integer programming. **European Journal of Operational Research**, v. 230, n. 2, p. 374–384. doi: 10.1016/j.ejor.2013.04.011, 2013.

JIAO, Z.; HIGGINS, A. J.; PRESTWIDGE, D. B. An integrated statistical and optimisation approach to increasing sugar production within a mill region. **Computers and Electronics in Agriculture**, v. 48, n. 2, p. 170–181. doi: 10.1016/j.compag.2005.03.004, 2005.

LAWES, R. A.; LAWN, R. J. Applications of industry information in sugarcane production systems. **Field Crops Research**, v. 92, n. 2–3, p. 353–363. doi: 10.1016/j.fcr.2005.01.033, 2005.

LOBELL, D. B.; BURKE, M. B. On the use of statistical models to predict crop yield responses to climate change. **Agricultural and Forest Meteorology**, v. 150, n. 11, p. 1443–1452. doi: 10.1016/j.agrformet.2010.07.008, 2010.

LUSTIG, I. J.; PUGET, J.-F. Program Does Not Equal Program: Constraint Programming and Its Relationship to Mathematical Programming. **Interfaces**, v. 31, n. 6, p. 29–53. doi: 10.1287/inte.31.6.29.9647, 2001.

MARIN, F. R.; JONES, J. W.; ROYCE, F.; 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.; JONES, J. W.; SINGELS, A.; et al. Climate change impacts on sugarcane attainable yield in southern Brazil. **Climatic Change**, v. 117, n. 1-2, p. 227–239. doi: 10.1007/s10584-012-0561-y, 2012.

MEINKE, H.; STONE, R. C. Seasonal and Inter-Annual Climate Forecasting: The New Tool for Increasing Preparedness to Climate Variability and Change In Agricultural Planning And Operations. **Climatic Change**, v. 70, n. 1-2, p. 221–253. doi: 10.1007/s10584-005-5948-6, 2005.



NASSIF, D. S. P.; MARIN, F. R.; PALLONE FILHO, W. J.; RESENDE, R. S.; PELLEGRINO, G. Q. Parametrização e avaliação do modelo DSSAT/Canegro para variedades brasileiras de cana-de-açúcar. **Pesq Agrop Brasileira**, v. 47, n. 3, p. 311–318, 2012.

PICOLI, M. C. A. **ESTIMATIVA DA PRODUTIVIDADE AGRÍCOLA DA CANA-DE-AÇÚCAR UTILIZANDO AGREGADOS DE REDES NEURAIS ARTIFICIAIS: ESTUDO DE CASO NA USINA CATANDUVA**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais - INPE, 2006.

PIEWTHONGNGAM, K.; PATHUMNAKUL, S.; SETTHANAN, K. Application of crop growth simulation and mathematical modeling to supply chain management in the Thai sugar industry. **Agricultural Systems**, v. 102, n. 1-3, p. 58–66. doi: 10.1016/j.agsy.2009.07.002, 2009.

POWER, D. J.; SHARDA, R. Model-driven decision support systems: Concepts and research directions. **Decision Support Systems**, v. 43, n. 3, p. 1044–1061. doi: 10.1016/j.dss.2005.05.030, 2007.

RAMBURAN, S.; ZHOU, M.; LABUSCHAGNE, M. Interpretation of genotype×environment interactions of sugarcane: Identifying significant environmental factors. **Field Crops Research**, v. 124, n. 3, p. 392–399. doi: 10.1016/j.fcr.2011.07.008, 2011.

SHARIAT, M.; NWAKANMA, H. Enterprise Resource Planning And Its Future Relationship To Decision Support System. **Journal of Business & Economics Research (JBER)**, v. 4, n. 12, 2006.

SPIERTZ, H. Challenges for Crop Production Research in Improving Land Use, Productivity and Sustainability. **Sustainability**, v. 5, n. 4, p. 1632–1644. doi: 10.3390/su5041632, 2013.

STANEK, S.; SROKA, H.; TWARDOWSKI, Z. Directions for an ERP-based DSS. Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8. 3 International Conference. **Anais...**, 2004.

STONE, R. C.; MEINKE, H. Operational seasonal forecasting of crop performance. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 360, n. 1463, p. 2109–2124. doi: 10.1098/rstb.2005.1753, 2005.

WEINTRAUB, A.; ROMERO, C. Operations Research Models and the Management of Agricultural and Forestry Resources: A Review and Comparison. **Interfaces**, v. 36, n. 5, p. 446–457. doi: 10.1287/inte.1060.0222, 2006.

WIEDENFELD, R. P. Effects of irrigation and N fertilizer application on sugarcane yield and quality. **Field Crops Research**, v. 43, n. 2-3, p. 101–108. doi: 10.1016/0378-4290(95)00043-P, 1995.

## 4 TÉCNICAS DE MINERAÇÃO DE DADOS APLICADAS À MODELAGEM DA PRODUTIVIDADE DE CANA-DE-AÇÚCAR

Título do Artigo:

**De planilhas a predições: Mineração de dados para modelagem da produtividade de cana-de-açúcar**

### 4.1 Resumo

---

Um modelo de produtividade é uma ferramenta capaz de auxiliar o planejamento agrícola para uma usina de cana-de-açúcar. Para explorar o volume de dados relacionados à produção, tipicamente disponível nas usinas de cana-de-açúcar, em conjunto com dados de clima, foram utilizadas técnicas de mineração de dados, como as redes neurais, que têm apresentado bons resultados para modelagem de produtividade, e técnicas mais avançadas como a SVM e a *Random Forest*. Estas foram aplicadas no conjunto de dados de produção da usina Alcídia, localizada no município de Teodoro Sampaio – SP. Entre as técnicas avaliadas, a *Random Forest* e a SVM obtiveram os melhores desempenhos e para SVM o modelo utilizou significativamente menos atributos, obtendo um erro absoluto médio de 4,61 [ton/ha] e  $R^2 = 0,79$ , obtendo intervalos de confiança de 95 % de 17,06, que é 54 % menor que o obtido em uma abordagem anterior similar. Durante a modelagem, a seleção de atributos mostrou que diferentes técnicas apresentaram diferentes capacidades de interagir com subconjuntos de atributos, ressaltando a necessidade da realização de um procedimento de seleção de atributos que evidencie os melhores atributos para cada modelo. Os modelos de desempenho inferior, obtiveram erros absolutos médios de aproximadamente 8 [ton/ha], desempenho considerado inadequado, dado que a predição à partir da média de produtividade pelo número de cortes possui erro médio de 9,58 [ton/ha]. A estratégia de modelagem baseada em dados permitiu a criação de modelos específicos para o contexto produtivo da própria unidade, na escala da menor unidade de gestão, os talhões. Os modelos de produtividade criados poderão realizar projeções de produtividade se utilizados em conjunto com projeções de clima.

---

## 4.2 Introdução

No contexto da indústria sucroenergética brasileira, tem-se frequentemente todas as etapas entre a produção e comercialização controladas por um único agente, responsável por plantar, cultivar, colher e transportar, e processar a cana-de-açúcar, gerando como produtos típicos o açúcar, etanol e eletricidade (HIGGINS et al., 2007). Além destes produtos, o setor também está expandindo seu escopo de produção com biopolímeros e a produção de etanol de 2ª geração. Com todas as etapas da cadeia de valor controladas por um único agente, a tomada de decisão pode ser feita de forma a obter o melhor desempenho para a cadeia como um todo, não sendo necessário conciliar os interesses de diferentes stakeholders. Por outro lado, também é necessário planejar todas as atividades da cadeia de forma sinérgica para os objetivos estabelecidos. No setor são tomadas decisões em diferentes escalas, por diversos gestores (HIGGINS et al., 2007), estando várias delas ligadas à produtividade do cultivo. Entre os fatores que afetam a produtividade da cana-de-açúcar, tem-se o número de cortes, as interações entre genótipo, ambiente e época da colheita (GILBERT et al., 2006). É possível desmembrar esses fatores em tipo de solo e suas frações granulométricas, capacidade de retenção de água, regime hídrico, variedades, acumulação de tempo térmico, chuva média, radiação recebida, duração do ciclo de crescimento, entre outros (RAMBURAN et al., 2009, 2011).

Ao estudar o uso de projeções de clima no planejamento agrícola da cadeia de cana-de-açúcar na Austrália, Everingham et al. (2002) listou as principais decisões afetadas pelo clima, apontando o impacto do clima na produtividade ao longo de toda a cadeia, o que é natural, dado que a produção depende da quantidade de cana disponível para processamento e sua qualidade industrial, dada pelo teor de sacarose acumulada nos colmos. Em resposta a essa demanda do setor, os autores utilizaram o modelo APSIM (KEATING et al., 2003) em conjunto com projeções climáticas para prever a produtividade na safra seguinte, permitindo estimar a produção total de cana-de-açúcar e melhorar a tomada de decisão nos contratos de venda futura de açúcar. Uma estrutura de predição da safra futura também é descrita por Bezuidenhout e Singels (2007) para África do Sul.

O uso de modelos de produtividade em conjunto com projeções de clima é uma forma direta de mostrar o impacto de uma projeção de clima na atividade agrícola, conforme pode ser visto no trabalho de Meinke e Stone (2005), onde os autores apontam diversas aplicações de

suporte à tomada de decisão. Modelos de produtividade se mostram como uma ferramenta para os tomadores de decisão usufruírem não só de projeções de clima, mas também para quantificar impactos de decisões e simular cenários. Entre as estratégias de modelagem disponíveis, temos os modelos baseados em processos ou mecanísticos, e os modelos empíricos. Enquanto modelos de processos são baseados na modelagem das interações solo-planta-atmosfera e no avanço do conhecimento em ecofisiologia (LISSON et al., 2005), modelos empíricos buscam relacionar a produtividade obtida a variáveis que representam os fatores condicionantes da produção agrícola. A modelagem empírica permite criar modelos específicos para os dados utilizados em sua criação, o que pode ser considerado uma vantagem, já que são modelos *ad hoc* para o problema analisado (AFFHOLDER et al., 2012), mas também uma limitação, pois seu escopo de aplicação é limitado ao escopo dos dados utilizados.

Uma característica favorável do setor sucroenergético é o acúmulo de informações relacionadas aos talhões e sua produtividade. O potencial dessas informações foi destacado por Lawes e Lawn (2005) e uma aplicação dos dados de talhões para modelagem da produtividade pode ser vista em Brüggemann et al. (2001), que utilizou dados de 146 talhões ao longo de 19 safras para modelagem, agregando os dados disponíveis de solo, manejo e clima. Embora o trabalho já apresente avanços em relação à abordagem de Alvarez et al. (1982), que também utilizou a regressão linear múltipla para predição da produtividade à partir de dados relacionados ao cultivo e a soma de precipitação no desenvolvimento da cultura, novas técnicas de modelagem têm gerado melhores resultados ao modelar a produtividade agrícola, entre elas as árvores de regressão (*Regression Tree*, RT) e redes neurais artificiais (*Artificial Neural Networks*, ANN). Essas técnicas foram utilizadas para modelar a produtividade de soja, milho e arroz (JI et al., 2007; KAUL et al., 2005; PARK et al., 2005; ZHENG et al., 2009), e tiveram um melhor desempenho que uma regressão linear múltipla, técnica de modelagem largamente utilizada. Foi destacada a capacidade de lidar com a colinearidade dos atributos, que não seguem a distribuição normal, interações não lineares entre fatores, efeitos a partir de determinados patamares e interação múltipla entre os fatores. Para comparar os modelos, Ji et al (2007) e Kaul et al. (2005) utilizaram a raiz do erro quadrático médio ( *root mean square error*, RMSE) e o coeficiente de regressão ( $R^2$ ) entre as predições dos modelos e os valores reais. Park et al. (2005) utilizaram apenas o  $R^2$ , e Zheng et al. (2009) utilizou a métrica de Redução Parcial do Erro, que calcula a

redução no erro pela inclusão de uma variável no modelo. A avaliação de modelos pelo coeficiente de regressão é desaconselhada por Mitchell (1997) e Harrison (1990), que apontam este uso como um desvio na função do coeficiente. Sobre o uso do RMSE, Willmott e Matsuura (2005) indicam que o erro absoluto médio (*Mean Absolute Error*, MAE) é preferível, dado o RMSE apresenta relação com a variância do erro, do número de pontos avaliados e do resíduo do modelo, fazendo com que não seja possível interpretar diretamente o valor ao comparar modelos, pois há ambiguidade na medida. Além disso, os autores destacam que essa ambiguidade é crítica na comparação de modelos gerados à partir de pontos diferentes. Além de não possuir essa ambiguidade, outra vantagem apontada pelos autores na avaliação do MAE é de que o parâmetro tem uma interpretação direta do seu significado.

Uma alternativa para avaliação de modelos de regressão são as curvas características de erro de regressão (*Regression Error Characteristic Curves*, REC) propostas por Bi e Bennet (2003). As curvas REC representam a função de distribuição acumulada empírica dos erros, sendo a abcissa a magnitude do erro e o eixo ordenado a acurácia, entendida como a probabilidade  $\rho_i$  - do erro  $\epsilon$  ser menor que  $\epsilon_i$ . As vantagens das curvas REC podem ser apontadas como a inspeção visual de métricas de erro, como a estatística KS (Kolmogorov- Smirnov) dada pela distância entre as curvas, e o valor médio da métrica de erro. Caso o valor de erro utilizado na construção da curva REC seja o erro absoluto, a área sobre a curva (*Area Over Curve*, AOC) representa uma sub-estimativa do MAE. Os autores ainda apontam que as curvas REC são análogas as curvas ROC (FAWCETT, 2006), e possuem boa parte de suas características positivas.

Uma aplicação de ANN para cana-de-açúcar pode ser vista no trabalho de Jiménez et al. (2008), porém os autores só utilizaram dados de clima e a duração do ciclo em sua modelagem. Por outro lado, uma aplicação que utilizou dados de produção na escala das fazendas é mostrada por Ferraro et al. (2009), que embora agregue variáveis relacionadas ao número de cortes, cultivar, época de colheita e atributos de clima, focaram sua análise em definir a categoria da produtividade avaliada em alto, média e baixa. Uma abordagem de modelagem empírica da produtividade de soqueiras utilizando dados de sensoriamento remoto (NDVI) e dados de talhões pode ser vista em Picoli (2006).

Não foram encontradas aplicações de técnicas de modelagem mais avançadas em dados de produção de cana-de-açúcar na escala de talhões, com objetivo de modelar numericamente a produtividade da cana considerando dados relativos a solo, manejo e clima em um contexto do cultivo não instalado. O uso de outras fontes de dados como variáveis climáticas, edáficas, bióticas e abióticas e de manejo dos talhões é indicada como uma forma de enriquecer as análises realizadas nesses conjuntos de dados (LAWES; LAWN, 2005). Cabe destacar que essas considerações, feitas para o contexto destes dados na Austrália divergem do encontrado em usinas no Brasil, onde uma série de informações edáficas e de manejo estão disponíveis junto com os dados de produtividade dos talhões. Ainda cabe destacar que Jiménez et al. (2008) apontam benefícios desta combinação de diferentes fontes de dados, mesmo que os dados apresentem ruídos, estejam incompletos ou sejam imprecisos. Com a inclusão de um grande número de atributos no conjunto de dados, é necessário realizar uma etapa de seleção de atributos, na qual são retirados atributos sem importância, implicando em um menor esforço computacional ou que representem ruído para os modelos. Por outro lado, o tratamento de dados faltantes, chamado de imputação no contexto de mineração de dados, tem como objetivo preservar o maior número de registros para a indução de modelos, sem diminuir a qualidade do conjunto de dados. Uma descrição detalhada das estratégias de imputação de dados e seleção de atributos, entre outras técnicas aplicadas ao pré-processamento dos dados, pode ser encontrada no livro-texto de Pyle (1999).

Enquanto as ANN e RT se mostram como alternativas a regressão múltipla, outras técnicas de modelagem têm sido utilizadas com sucesso na agricultura. Ruß (2009) utilizou a *Support Vector Machine* (SVM) para predição de produtividade em um contexto de agricultura de precisão, obtendo resultados melhores do que ANN. Em uma análise de diversos conjuntos de dados de referência (*benchmark*), Verikas et al. (2011) apontaram *Random Forest* (RF), SVM e ANN como técnicas que despontam em performance em tarefas de predição numérica, sendo necessário avaliar as técnicas disponíveis diante dos problemas encontrados, uma vez que não existe um algoritmo que apresente desempenho superior para qualquer problema.

De forma simplificada, ao utilizar a RF, são criadas diversas árvores de decisão, porém, para evitar que esses diferentes modelos sejam correlacionados, os atributos que serão utilizados para construir a árvore são escolhidos aleatoriamente. Ao gerar vários desses modelos aleatórios

e realizar a predição através do resultado médio das diferentes árvores, temos um modelo mais robusto, e em geral, com um desempenho superior à árvore de decisão ou regressão. As SVM foram desenvolvidas por Vapnik et al. (1996) para problemas de predição categórica (classificação) e posteriormente adaptadas para predição numérica (SCHÖLKOPF et al., 2000). O método consiste em encontrar um plano que melhor separe duas classes, otimizando a distância de separação. Caso uma separação linear não seja possível, a técnica é capaz de mapear os dados em um espaço de dimensão superior, onde é possível utilizar um hiperplano para separar os dados. Para predição numérica, a SVM busca maximizar o número de pontos dentro de uma margem, valendo destacar que a solução encontrada é um ponto ótimo global, diferente das ANN que podem convergir para um ótimo local. Além das SVM e RF que seriam consideradas técnicas mais avançadas de modelagem disponíveis, chama atenção o uso da RT sem que tenha sido considerado o uso das árvores de modelo, que podem ser interpretadas como um avanço. Em uma árvore de modelos, em cada folha, têm-se uma regressão múltipla representando os elementos daquela folha, e não uma representação pela média, como seria em uma RT. Dessa forma, a árvore particiona o conjunto de dados e gera regressões múltiplas para subconjuntos de dados, apresentando uma melhor capacidade preditiva (QUINLAN, 1993). Descrições completas da SVM e RF podem ser encontradas em Schölkopf et al. (2000) e Breiman (2001), respectivamente.

Foi identificado junto ao setor que um momento crítico no planejamento é a elaboração do orçamento em conjunto com o plano de safra, com consequente estabelecimento de contratos de venda futura de açúcar. Projeções iniciais são feitas para safra seguinte em Agosto, antes mesmo que a safra corrente termine para fins de orçamento. Novas estimativas são realizadas em Janeiro para o plano de safra que se inicia em Março. Na elaboração conjunta do plano de safra-orçamento-vendas, temos o planejamento de toda a cadeia, englobando a produção agrícola, processamento e comercialização. Além disso, temos o envolvimento da cadeia de manipulação de materiais (BEZUIDENHOUT; BAIER, 2009), pois o plano de colheita é dependente do plano de safra. Dada a diversidade de fatores que influenciam a produtividade da cana-de-açúcar (GILBERT et al., 2006; RAMBURAN et al., 2009, 2011), as interações complexas e não lineares que se dão entre elas e o volume de talhões para o qual uma usina deve estimar a produtividade

da cana de açúcar, que se encontram em diferentes estágios de crescimento, têm-se uma oportunidade para o uso de modelos de produtividade como ferramenta de suporte a decisão.

Nesse contexto de aplicação, Passioura (1996) destaca o uso de modelos empíricos com o único objetivo de fornecer uma resposta robusta para a tomada de decisão, sendo o uso desses modelos confinados ao escopo da criação dos mesmos. Assim, para o contexto apresentado, o objetivo desse trabalho é criar um modelo empírico de produtividade de cana-de-açúcar, a partir dos dados disponíveis no conjunto de dados de produção e dados climáticos relacionados ao período de desenvolvimento dos diferentes talhões. São considerados objetivos específicos a avaliação dos atributos mais importantes nos dados disponíveis e a avaliação e comparação do desempenho de técnicas utilizadas em abordagens anteriores (RT, ANN e Regressão Múltipla) e técnicas com potencial de melhor desempenho, dado seu desempenho em outros campos de aplicação (SVM, RF e Árvore de Modelos).



### 4.3 Material e métodos

Os dados utilizados foram fornecidos pela Odebrecht Agroindustrial, na forma de planilhas correspondentes aos dados de produção para as safras 2010/2011 e 2011/12 da unidade Alcídia, localizada no município de Teodoro Sampaio - SP. Além dos dados de produção, foram fornecidos dados de pluviometria diária de Janeiro de 2005 até Maio de 2013. Dados de temperatura foram obtidos de uma estação convencional próxima à unidade. Entende-se que a metodologia desenvolvida está de acordo com o descrito por Chapman et al. (2000), e embora o processo seja essencialmente iterativo, será descrito linearmente. Foi utilizado o software estatístico R (R. CORE TEAM, 2013) ao longo de todas as etapas do processo e pacotes adicionais utilizados serão citados oportunamente. Em linhas gerais, os dados foram recebidos no formato de planilhas e submetidos a uma etapa de limpeza e correção, conforme detalhado a seguir. Foi relacionado o clima ocorrido no desenvolvimento de acordo com as datas de plantio e colheita dos talhões de cana-planta e entre a colheita anterior e a colheita para os talhões de cana-soca. Após a composição do conjunto de dados, o mesmo foi dividido em 2/3 para treino e 1/3 para teste. No conjunto de dados de treino foi realizada a seleção de atributos e ajuste dos parâmetros dos algoritmos (*tunning*). Após o *tunning*, os modelos foram treinados no conjunto de treinamento e então utilizados para prever a produtividade do conjunto de teste. A avaliação dos modelos se deu através da comparação da predição realizada e valores reais da produtividade no conjunto de teste. O arranjo geral da metodologia pode ser visto na Figura 14, enquanto o detalhamento das etapas se dá no resto desta seção.

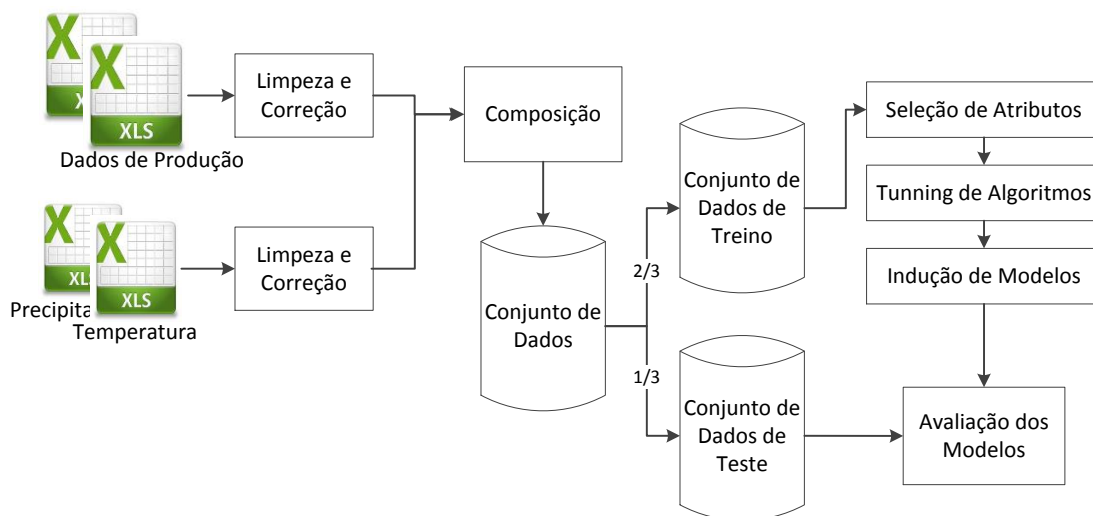


Figura 14. Arranjo geral da metodologia utilizada.

### **4.3.1 Dados de produção**

#### **4.3.1.1 Descrição Geral**

Foram recebidas duas planilhas contendo os dados de produção. Cada linha das planilhas continha informações relacionadas a um talhão e sua produtividade. Os atributos descreviam a química e as frações granulométricas do solo, assim como sua classificação. As informações de manejo disponíveis eram relativas à adubação, datas de plantio, colheita e colheita anterior, número de cortes, aplicação de torta de filtro e vinhaça, avaliações fitossanitárias, ocorrência de queima ou de geada. O conjunto de dados inicial possuía 65 colunas e 1303 linhas na planilha referente à safra do ano de 2010/11 e 1343 linhas para safra seguinte. Foram excluídos inicialmente os talhões sem informação de produtividade e talhões identificados como áreas em formação ou mudas, uma vez que não se aplicam para estudos da produtividade colhida. Além disso, foram excluídos talhões que não tivessem simultaneamente a data da colheita e a data da colheita anterior ou plantio, uma vez que nessa condição, não é possível delimitar o início e o fim do ciclo de desenvolvimento do talhão. As avaliações fitossanitárias disponíveis foram excluídas por se tratarem de informações obtidas na colheita, fazendo com que não estejam disponíveis com antecedência para predição. O conjunto de dados final corresponde a uma área de aproximadamente 25.000 ha e uma produção somada de 3 milhões de toneladas de cana-de-açúcar, referentes a 1102 talhões na safra de 2010/11 e 1153 para 2011/12, com 77 atributos, incluindo o atributo aleatório inserido e o atributo meta. Uma característica importante, identificada no conjunto de dados, é o fato de que em algumas situações, as colhedoras atravessam mais de um talhão para colheita, fazendo com que uma produtividade média seja atribuída a diferentes talhões. Não foi possível recuperar essa informação para exclusão dos registros nesta condição, interpretada como um ruído no atributo meta.

#### **4.3.1.2 Limpeza e Correção**

Os dados foram verificados de acordo com a coerência entre atributos, sendo:

- Soma das frações granulométricas deveria equivaler a 100 %

- Atributos de Solo como Soma de Bases, CTC, saturação de bases e saturação por alumínio foram calculados a partir dos valores disponíveis e comparados com os dados do conjunto.
- Além disso, foram verificadas as coerências entre as datas presentes nas planilhas fornecidas, como a colheita anterior para a safra de 2012 deveria coincidir com a colheita de 2011, a colheita deveria ser depois da colheita anterior e o número de cortes deveria ser coerente com a época de plantio e colheita.

Além das verificações de coerência, foram identificados através de inspeção dos histogramas e *boxplots* valores que poderiam ser considerados como *outliers* ou incoerentes. Esses valores foram identificados e submetidos à usina, que foi capaz de corrigir parcialmente o conjunto de dados ou validar a suspeita sobre os dados errôneos, sendo estes registros excluídos.

Foi necessário padronizar campos de entrada textual, devido à variação no uso de acentuação, letras maiúsculas e minúsculas ou espaços, *e.g.* “LV”, “lv”, “L V” ou “Lv” para o código do tipo de solo. Houve necessidade de compatibilizar as unidades referentes à lâmina de aplicação de vinhaça, torta de filtro e quantidade de adubo. No conjunto de dados, constavam lâminas de vinhaça em m<sup>3</sup> por hectare ou litros por hectare, que foram padronizadas para mm. Para torta de filtro e adubo, as quantidades foram padronizadas para kg.

Foi fornecida pela usina a classificação dos talhões dentro de áreas homogêneas, utilizada internamente para manejo e outro campo relativo a classificação do solo, no formato XXXX Y – Z, sendo XXXX a classificação do solo de acordo com o Sistema Brasileiro de Classificação de Solos ou SIBCS (SANTOS et al., 2006), Y um código numérico onde 1 corresponde a textura muito argilosa e varia até 6 para muito arenosa, estando o número 7 relacionado a textura siltosa e Z corresponde a uma gradação de fertilidade, sendo 1 para a categoria mais fértil e 7 a menos fértil. Estas informações foram desmembradas em diferentes colunas. O atributo de textura foi denominado Textura e o atributo relativo a fertilidade foi denominado Fertilidade, enquanto a classificação de solos seguiu a sigla do SiBCS no segundo nível hierárquico.

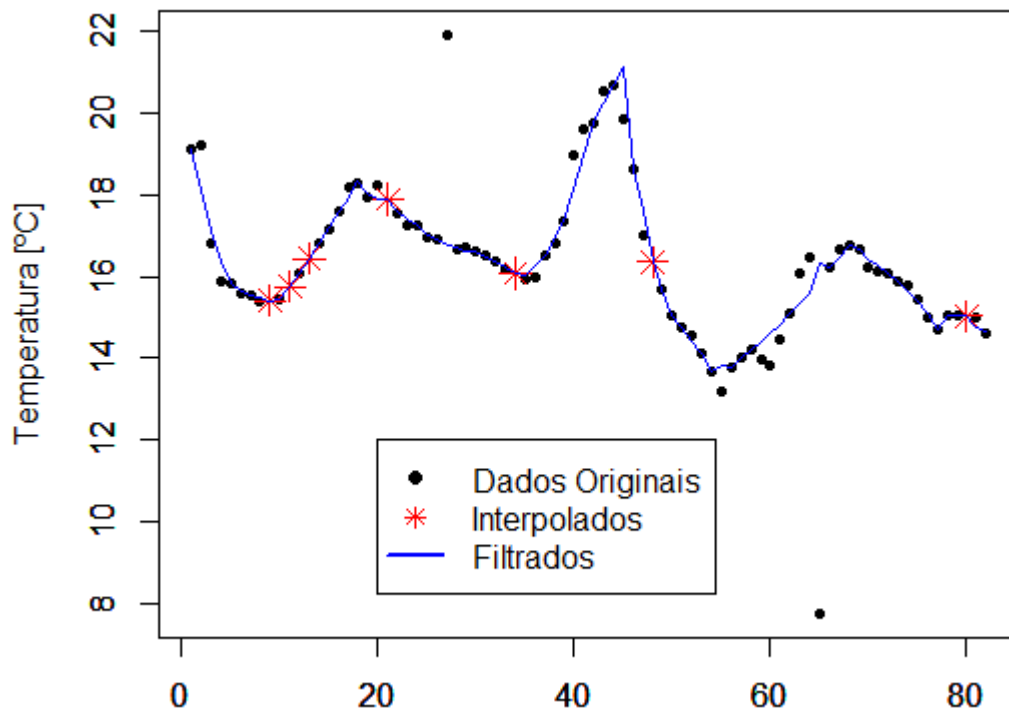
Entre os valores disponíveis para a química de solo, havia valores faltantes para o pH, matéria orgânica (MO), P, K, Ca, Mg, Acidez trocável (H+Al) e Al. Para opção quanto a imputação, foram avaliados nos registros sem dados faltantes a performance da imputação pela

média, pelo algoritmo kNN e por árvore de regressão em cross-validation de 10 folds. Os melhores resultados foram obtidos pela árvore de regressão, algoritmo que foi utilizado para imputação.

#### **4.3.2 Dados de clima**

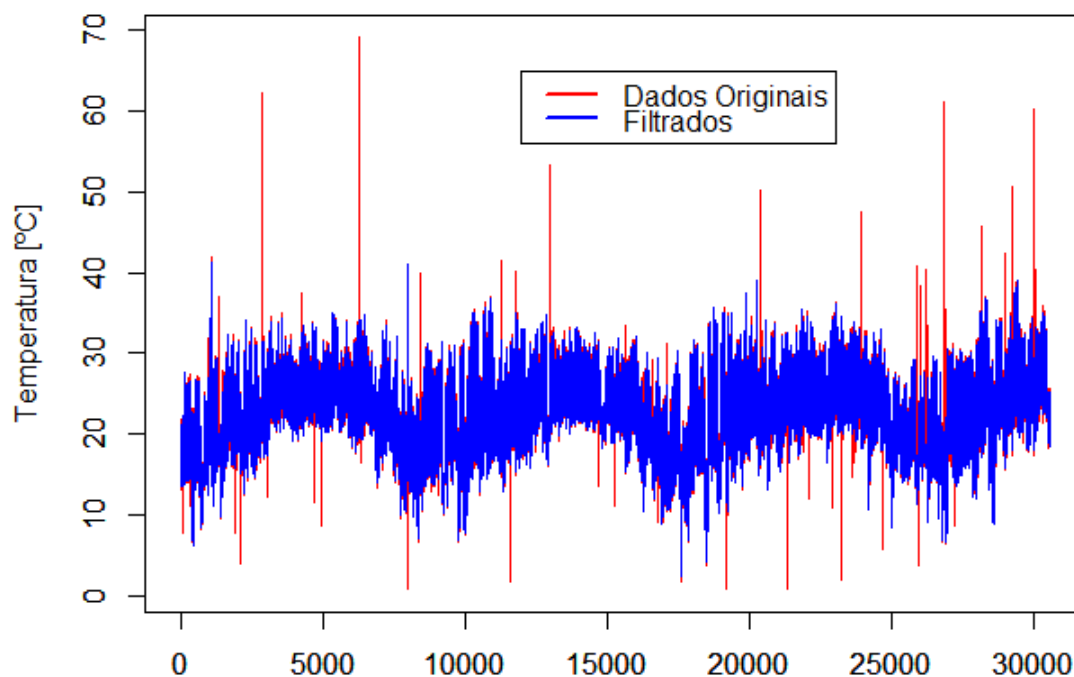
##### **4.3.2.1 Temperatura**

Os dados de temperatura foram recebidos em 9 planilhas e correspondiam a dados horários de temperatura da estação meteorológica do Parque Estadual do Morro do Diabo (Teodoro Sampaio – SP), das 19 horas do dia 6 de Abril de 2009 até as 23 horas do dia 25 de Junho de 2013. Foi selecionada a série entre os dias 23 de Junho de 2009 a 18 de Dezembro de 2012, que compreende o início (brotação ou plantio) do primeiro ciclo de crescimento nos dados de produção até o fim (colheita) do último ciclo. Nesta série de dados havia 2432 registros faltantes, eventualmente consecutivos. Optou-se por interpolar linearmente os dados faltantes em sequências de até 3 valores. A interpolação de sequências maiores implica perda de informações relacionadas a alteração da temperatura após os pontos de mínima ou máxima. Após a interpolação, foi aplicado um filtro para remoção de valores incoerentes. Os valores incoerentes correspondiam a falhas no registro da temperatura como pode ser visto na Figura 15. Foi utilizado o filtro *robust.filter* configurado para análise de 5 pontos consecutivos, disponível no pacote *robfilter*, implementado por Fried et al. (2012).



**Figura 15.** Seção da série de dados de temperatura com pontos interpolados destacados e série de dados após aplicação do filtro. Dois pontos foram removidos pelo filtro, um correspondente a um outlier de 22 °C em uma sequência de temperaturas entre 16 e 18 °C. Outro outlier corresponde ao valor de 8 °C.

Após a aplicação do filtro, a série temporal foi agregada para uma frequência diária, para a qual foram calculados os valores de temperatura mínima, média e máxima. Para a série diária, foram identificados os dias que ainda possuíam valores faltantes. Para esses dias, os valores de mínima, média e máxima foram interpolados linearmente com os dias imediatamente antes e depois. Na Figura 16 pode ser vista a série de dados completa após a aplicação do filtro.



**Figura 16.** Série de dados de temperatura antes e depois da aplicação do filtro.

#### **4.3.2.2 Precipitação**

Dados de precipitação, referente aos pluviômetros da unidade Alcídia, foram agrupados para representar a média de precipitação ocorrida na unidade. Não foi possível associar as medidas de precipitação a diferentes regiões ou talhões, pois não é feito registro das posições dos pluviômetros.

#### **4.3.2.3 Clima ocorrido**

O clima ocorrido no período que compreende o ciclo de crescimento nos talhões analisados está sintetizado na Figura 17.

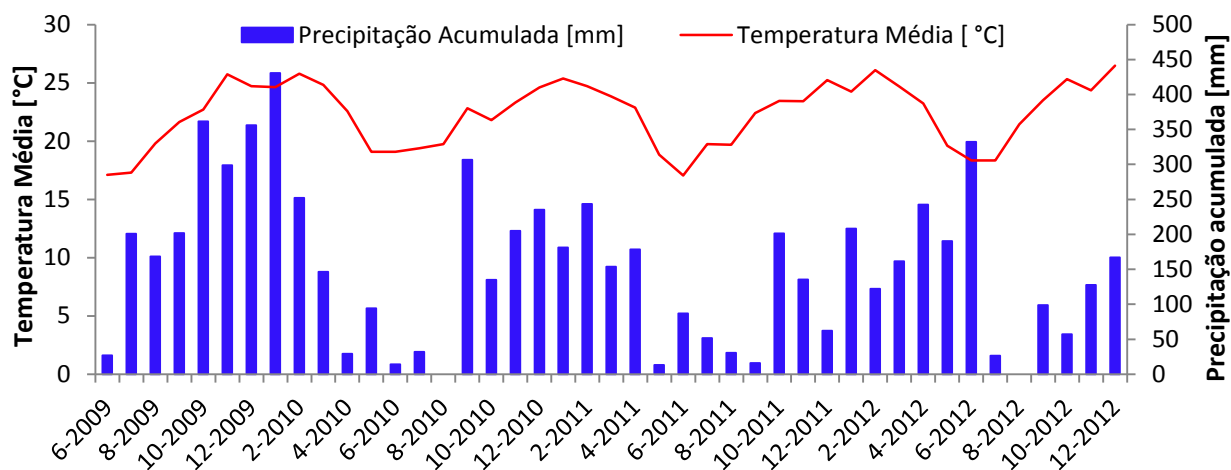


Figura 17. Clima ocorrido entre Junho de 2009 a Dezembro de 2012 (início do primeiro ciclo de desenvolvimento até a colheita do último).

### 4.3.3 Preparação de dados

#### 4.3.3.1 Estimativa de Fenologia e Atributos climáticos

O clima foi caracterizado em quatro períodos ao longo do ciclo de desenvolvimento, com objetivo de representar o efeito diferenciado nas diferentes fases de crescimento da cana-de-açúcar. Devido à inexistência de dados relativos à fenologia, foram utilizadas estimativas baseadas no comportamento local típico da cultura, conforme representado na Tabela 3.

**Tabela 3. Fenologia típica para região de estudo em função do tipo de ciclo ao longo do ano.**

Logotípica para regime de estudo em função do tipo de ciclo ao longo do ano.																																				
F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N															
1	2	3	4	5	6	7	8	9	10	11	12	13	14																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ciclo de Ano e Meio																					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14																					
				1	2	3	4	5	6	7	8	9	10	11	12	13																				
					1	2	3	4	5	6	7	8	9	10	11	12	13	Inverno																		
						1	2	3	4	5	6	7	8	9	10	11	12	13																		
							1	2	3	4	5	6	7	8	9	10	11	12	13																	
								1	2	3	4	5	6	7	8	9	10	11	12																	
									1	2	3	4	5	6	7	8	9	10	11	12																
										1	2	3	4	5	6	7	8	9	10	11	12															
											1	2	3	4	5	6	7	8	9	10	11	12														
												1	2	3	4	5	6	7	8	9	10	11	12													
													1	2	3	4	5	6	7	8	9	10	11	12												
														1	2	3	4	5	6	7	8	9	10	11	12											
															1	2	3	4	5	6	7	8	9	10	11	12										
																1	2	3	4	5	6	7	8	9	10	11	12									
																	1	2	3	4	5	6	7	8	9	10	11	12								
																		1	2	3	4	5	6	7	8	9	10	11	12							
																			1	2	3	4	5	6	7	8	9	10	11	12						
																				1	2	3	4	5	6	7	8	9	10	11	12					
																					1	2	3	4	5	6	7	8	9	10	11	12				
																						1	2	3	4	5	6	7	8	9	10	11	12			
																							1	2	3	4	5	6	7	8	9	10	11	12		
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11	12	
																								1	2	3	4	5	6	7	8	9	10	11		

Para as diferentes fases foram calculados a precipitação total e média, e criados atributos também relacionados ao número de dias consecutivos sem chuva (Estiagem) e a contagem de veranicos (Veranicos). Foram considerados veranicos os períodos de 10 dias consecutivos sem chuva entre os meses de outubro a fevereiro (PORTO DE CARVALHO et al., 2013). Devido à aplicação de vinhaça no início do desenvolvimento em parte da área da usina, essa lâmina foi considerada para os cálculos de veranico e estiagem, assim como foram feitas correções na soma e na média de precipitação do período de brotação. Para o período, foram calculados os valores médios da temperatura mínima, média e máxima diária. Utilizando os valores diários de temperatura foi calculado, também, o acúmulo de tempo térmico em Graus-Dias (LIU et al., 1998), considerando uma temperatura base de 18 °C.



#### 4.3.3.2 Atributos de Manejo

Entre as informações de manejo fornecidas, foram combinadas a fórmula de adubação aplicada e sua quantidade para calcular as diferentes quantidades de nutriente aplicadas, gerando as quantidades individuais de N, P e K, gerando os atributos *insumoN*, *insumoP* e *insumoK*. Uma vez que não é aplicado K em áreas de vinhaça, foi calculada a quantidade desse nutriente aportada pela vinhaça, tomando por base uma concentração média de 1,8 kg/m<sup>3</sup>, de acordo com dados fornecidos pela usina. Os dados relativos à aplicação de Molibdênio via foliar também foram convertidos para a quantidade de nutriente de acordo com a formulação e quantidade aplicada, gerando o atributo *insumoMo*.

Para os dados de química do solo, foram calculadas algumas relações entre atributos químicos, sendo:

**Tabela 4. Atributos derivados para descrição do solo.**

Atributo	Referência
Ca/Mg, Gradiente Textural (% Argila Camada B / % Argila Camada A =)	(BRÜGGEMANN et al., 2001; DIAS et al., 1999)
Mg/K, K/CTC, Ca/K	(ORLANDO FILHO et al., 1996)
K/(Ca + Mg) <sup>1/2</sup>	(ORLANDO FILHO et al., 1996; REIS JUNIOR, 2001)

#### 4.3.4 Seleção de Atributos

A seleção de atributos foi realizada utilizando como critério a importância dos atributos dentro da estrutura dos modelos, utilizando a função *varImp*, disponível no pacote *fscaret*, conforme descrito por Kuhn (2008). Essa forma de seleção de atributos foi utilizada para Regressão Múltipla, Árvore de Modelos, Árvore de Regressão, ANN e RF e tem como resultado uma lista de atributos com um valor associado a sua importância. Foi incluído no conjunto de dados um atributo correspondente a um número aleatório extraído de uma distribuição normal padrão  $\sim N(0,1)$ . Essa variável foi utilizada como critério de corte nas listas de importância, o que pode ser considerado uma heurística para seleção. Para modelos onde não é possível inferir a importância dos atributos a partir da estrutura do próprio modelo, a abordagem disponível no *fscaret* seria do tipo filtro, o que foi preterido em prol de uma abordagem do tipo Wrapper, onde é buscado um subconjunto de atributos capaz de minimizar o erro de predição. Dada as limitações computacionais de uma busca exaustiva, optou-se por uma seleção de atributos para SVM por meio de um algoritmo genético binário utilizando o pacote *genalg* (WILLIGHAGEN, 2012), onde a solução foi codificada em um cromossomo binário, representando a inclusão ou

não dos atributos. Foram realizadas 50 iterações para uma população de 50 indivíduos, com uma chance de mutação de 20 % e 10 % de elitismo. Na etapa posterior, de treinamento, cada modelo foi treinado de acordo com a lista indicada. Com exceção do Wrapper da SVM, os métodos utilizaram uma divisão do conjunto de treino em 2/3 e 1/3 para a indução de modelos e avaliação da importância dos atributos. Devido à tendência de *overfitting* na seleção de atributos com algoritmo genético em conjunto com SVM (FROHLICH et al., 2003), foi utilizado Cross-Validation. Como modelo de referência, foi realizada uma regressão múltipla do tipo *stepwise* bidirecional.

#### 4.3.5 Modelagem e Avaliação

A determinação dos parâmetros dos algoritmos, ou *tunning*, foi realizada através de uma busca em grid para minimizar o erro em Cross-Validation de 10 folds. Os limites inferior e superior utilizados para o ajuste dos algoritmos e o valor ótimo determinado podem ser consultados na Tabela 5, na qual constam também os pacotes de origem dos algoritmos utilizados.

**Tabela 5. Ajuste dos parâmetros dos atributos, valores determinados e implementação utilizada.**

Técnica/Pacote	Parâmetro Ajustado	Limite Inferior	Limite Superior	Ponto ótimo
<b>SVM/e1071</b> (MEYER, 2012)	Custo	2	1024	4,28
	Gama	0,001	0,25	0,06
	Epsilon	0,01	0,1	0,03
<b>RF/randomForest</b> (LIAW; WIENER, 2002)	Número de Árvores	8	200	100
	Número de Atributos	4	16	10
	Número máximo de Nós	200	1100	550
	Tamanho Mínimo das Folhas	1	100	3
<b>Árvore de Modelos/Cubist</b> (KUHN et al., 2013)	Número de folhas	1	100	7
<b>ANN/AMORE</b> (LIMAS et al., 2010)	Número de nós na primeira camada	3	77	20
	Número de nós na segunda camada	0	77	0
<b>RT/rpart</b> (THERNEAU et al., 2012)	Tamanho mínimo das folhas	2	512	16
	Profundidade da árvore	2	16	11

Após a determinação dos parâmetros ótimos, os modelos foram criados a partir dos dados de treino e utilizados para prever a produtividade no conjunto de teste. Foram utilizados para avaliação dos modelos o Erro Médio Absoluto (*Mean Absolute Error*, MAE) e o Viés Médio de Erro (*Mean Bias Error*, MBE) de acordo com Willmott e Matsuura (2005). Foi utilizada também a curva REC (BI; BENNETT, 2003). Mesmo diante das limitações em utilizar o coeficiente de

determinação entre valores reais e preditos (HARRISON, 1990; MITCHELL, 1997), o mesmo foi calculado a título de comparação, dado seu uso generalizado.

## 4.4 Resultados e Discussão

### 4.4.1 Seleção de Atributos

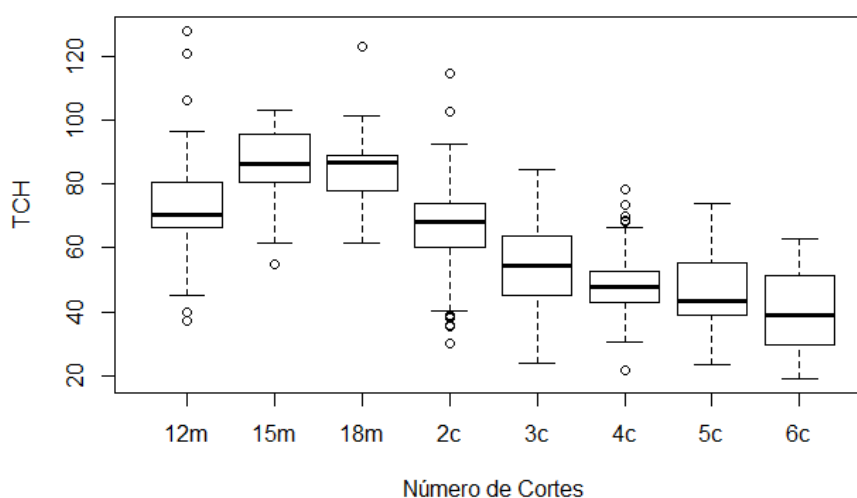
Ao obter a importância de cada atributo para as técnicas de modelagem, foram excluídos os atributos nos quais a importância era igual ou menor que o atributo de valor aleatório introduzido no conjunto de dados. Para cada modelo, o procedimento produziu diferentes subconjuntos de atributos (Tabela 6), um reflexo das diferentes estratégias embutidas nos modelos, fazendo com que consigam extrair informação de forma mais ou menos eficiente de um mesmo atributo. O fato de um atributo receber uma importância menor de que um atributo aleatório não implica que não exista relação entre aquele atributo e o atributo meta e sim que utilizar aquele atributo não irá ter efeito diferente de um atributo aleatório para a predição, na técnica em questão. A inclusão do atributo aleatório forneceu um critério objetivo quanto ao ponto de corte para técnicas de seleção de atributos onde é fornecida a importância do atributo, sendo responsabilidade do usuário estabelecer o ponto de corte. Foram selecionados 28 atributos para árvore de regressão, 37 para regressão múltipla, 40 para SVM, 42 para regressão *stepwise*, 44 para RNA, 55 para Árvore de Modelos e 75 para *Random Forest*.

**Tabela 6. Atributos selecionados por método. ( X ) – Selecionado e ( - ) – Não-selecionado. Para atributos de clima, algarismos romanos indicam o estágio para o qual o atributo foi calculado (I – Brotação, II – Perfilhamento, III – Crescimento Rápido e IV – Maturação), enquanto as letras S e M são referentes à Soma e Média, respectivamente.**

Atributo		RT	RM	SVM	Step	ANN	AM*	RF	Atributo		RT	RM	SVM	Step	ANN	AM*	RF
1	Corte	X	X	X	X	X	X	X	39	Precipitação.S.II	X	-	X	X	-	-	X
2	Graus Dias.S.II	X	X	X	X	X	X	X	40	Ca	X	-	X	-	-	X	X
3	TMax.M.III	X	X	X	X	X	X	X	41	Precipitação.M.IV	X	-	-	X	-	X	X
4	Matéria Orgânica	X	X	X	X	-	X	X	42	Saturação de Alumínio	-	X	X	-	-	X	X
5	insumoN	X	X	X	-	X	X	X	43	Al	-	X	-	X	-	X	X
6	Precipitação.S.I	X	X	X	-	X	X	X	44	Silte.3aCamada	-	X	-	X	-	X	X
7	Precipitação.S.IV	X	X	X	-	X	X	X	45	Ca/K	-	X	-	-	X	X	X
8	Graus Dias.S.I	X	X	-	X	X	X	X	46	Estiagem.IV	-	X	-	-	X	X	X
9	Mg	X	-	X	X	X	X	X	47	Silte.2aCamada	-	X	-	-	X	X	X
10	Tmin.M.III	X	-	X	X	X	X	X	48	Classe de Fertilidade	-	-	X	-	X	X	X
11	Ca/Mg	-	X	X	X	X	X	X	49	Mg/K	-	-	X	-	X	X	X
12	Densidade	-	X	X	X	X	X	X	50	Epoca do Corte	-	-	-	X	X	X	X
13	Estiagem.I	-	X	X	X	X	X	X	51	K/CTC	-	-	-	X	X	X	X
14	Veranico.II	-	X	X	X	X	X	X	52	Areia.1aCamada	X	-	X	-	-	-	X
15	insumoK	-	X	X	X	X	X	X	53	Areia. 2aCamada	X	-	X	-	-	-	X
16	Precipitação.M.II	-	X	X	X	X	X	X	54	Areia. 3aCamada	X	-	-	X	-	-	X
17	Precipitação.M.III	-	X	X	X	X	X	X	55	Soma Bases	X	-	-	X	-	-	X
18	Textura	-	X	X	X	X	X	X	56	SoloEsp	-	X	-	-	-	X	X
19	Ciclo de Crescimento [dias]	X	X	X	-	X	-	X	57	TMin.M.II	-	-	X	X	-	-	X
20	Veranico.I	X	X	X	-	-	X	X	58	Estiagem.III	-	-	-	-	X	X	X
21	Precipitação.M.I	X	X	-	-	X	X	X	59	KCaMg	-	-	-	-	X	X	X
22	CTC	X	-	X	X	-	X	X	60	Precipitação.S.III	-	-	-	-	X	X	X
23	pH	X	-	X	X	-	X	X	61	Silte.1aCamada	-	-	-	-	X	X	X
24	GrausDias.S.III	X	-	X	-	X	X	X	62	TMax.M.IV	-	-	-	-	X	X	X
25	TMax.M.II	-	X	X	X	-	X	X	63	Saturação de Bases	-	-	-	-	X	X	X
26	Ambiente de Manejo	-	X	-	X	X	X	X	64	Argila. 2aCamada	X	-	-	-	-	-	X
27	Estiagem.II	-	X	-	X	X	X	X	65	Argila. 3aCamada	X	-	-	-	-	-	X
28	Veranico.III	-	X	-	X	X	X	X	66	TMed.M.III	X	-	-	-	-	-	X
29	Veranico.IV	-	X	-	X	X	X	X	67	insumoMo	-	-	X	-	-	-	X
30	Gradiente Textural	-	X	-	X	X	X	X	68	K	-	-	X	-	-	-	X
31	P	-	X	-	X	X	X	X	69	TMin.M.I	-	-	X	-	-	-	X
32	SiBCS	-	X	-	X	X	X	X	70	Tmin.M.IV	-	-	X	-	-	-	X
33	TMax.M.I	-	X	-	X	X	X	X	71	Vinhaça	-	-	X	-	-	-	X
34	TMed.M.IV	-	X	-	X	X	X	X	72	TMed.M.II	-	-	-	X	-	-	X
35	Variedade	-	X	-	X	X	X	X	73	fonteK	-	-	-	-	-	X	X
36	HA1	-	-	X	X	X	X	X	74	Torta	-	-	-	X	-	-	-
37	Argila.1aCamada	X	-	X	X	-	-	X	75	TMed.M..I	-	-	-	-	-	-	X
38	insumoP	X	-	X	X	-	-	X									

AM\* : Árvore de Modelos

Entre os atributos, os únicos incluídos em todos os modelos foram o número de cortes, a soma de Graus-Dias no perfilhamento e a média da temperatura máxima diária na fase de crescimento rápido. A importância do número de cortes e do manejo do primeiro ciclo é de fácil constatação, dada a distribuição da produtividade em função desta variável (Figura 18). Os outros dois atributos guardam relação com a temperatura em duas fases essenciais para definição da produtividade. O impacto da temperatura na cana-de-açúcar está relacionado ao próprio desenvolvimento da biomassa (EBRAHIM et al., 1998) e afeta fatores importantes como a formação do dossel e perfilhamento (INMAN-BAMBER, 1994).



**Figura 18. Produtividade em função do primeiro ciclo (12, 15 e 18 m) e número de cortes.**

Dos 15 atributos selecionados por 6 dos 7 modelos (atributos 4 a 18 na Tabela 6), temos a presença de 8 atributos relacionados ao clima, sendo 6 deles relacionados à precipitação e dois relacionados à temperatura. Essa importância dos atributos relacionados ao clima é natural, dado que em linhas gerais, até 90 % da variabilidade do resultado agrícola pode ser atribuído ao clima (HOOGENBOOM, 2000), e pode ser vista também se analisarmos a importância dada aos atributos em cada método (Tabela 7). Entre os 15 atributos mais importantes, temos uma maioria de atributos climáticos com exceção da árvore de regressão, que destoa dos demais pela presença dos atributos das frações de areia e argila nas três camadas disponíveis no conjunto de dados.

De uma forma geral, os atributos selecionados mais vezes são atributos ligados ao clima, enquanto os atributos relacionados ao solo foram selecionados menos vezes. Os atributos relacionados ao acúmulo de Graus-Dias foram escolhidos em média 6 vezes, os atributos

relacionados à soma ou média de precipitação no período têm média 5 e os de veranico e estiagem têm média 4,87. Na sequência, temos os atributos de Manejo, escolhidos em média para 4,45 dos algoritmos e os atributos de solo, com média 4,05 e 3,69 para parte química e física respectivamente. Os atributos de temperatura tiveram média 3,58, porém ocorreram atributos escolhidos por todos ou para 6 dos 7 algoritmos (atributos 3 e 10 na Tabela 6), e por outro lado outros que foram escolhidos apenas para dois algoritmos, como a média da temperatura mínima nos períodos I e IV (atributos 69 e 70 da Tabela 6), ou mesmo apenas uma vez, como a média da temperatura média no período I. Cabe destacar que os atributos de temperatura mínima, média e máxima apresentam uma auto-correlação intrínseca, sendo natural que após o modelo utilizar um deles não haja ganho de informação ao utilizar outro. Isso reflete na dispersão de posições de alguns atributos de temperatura na lista.

O fato de alguns atributos de manejo e solo serem selecionados (Tabela 6) para vários algoritmos, até mais vezes que alguns atributos relacionados ao clima, e possuírem importância elevada na listagem de importância (Tabela 7), é interpretado como um indicativo da importância de incorporar os dados sobre solo e manejo para os modelos preditivos. Chama atenção o fato da matéria orgânica (MO) e a quantidade de magnésio do solo (Mg) estarem tão bem posicionados quanto as taxas de aplicação de N e K. Outro atributo bem posicionado foi a relação Ca/Mg, que também foi importante na modelagem realizada por Brüggemann et al. (2001) onde a relação entre Cálcio e Magnésio possuía a mesma significância estatística que a taxa de aplicação de nitrogênio. Os outros atributos de relações químicas no solo foram selecionados em 4 de 7 técnicas com exceção da relação entre Potássio e da raiz quadrada da soma do Cálcio e Magnésio (atributo 59 da Tabela 6), enquanto a relação da fração de argila nos horizontes A e B foi selecionada por 5 dos métodos. A relação entre os nutrientes disponíveis se mostrou mais importante que alguns dos nutrientes individualmente. A taxa de aplicação de Nitrogênio não foi selecionada pela regressão Stepwise, a de Potássio não foi selecionada pela árvore de regressão, enquanto a taxa de aplicação de Fósforo não foi selecionada pela regressão múltipla, rede neural e árvore de modelos. A ausência destes atributos nos modelos reforça a interpretação de que as técnicas empregadas não foram capazes de extrair informações a partir do atributo e não que o atributo não seja relevante para a produtividade.

**Tabela 7. Atributos mais importantes para métodos de acordo com a função *varImp*.**

Rank	Regressão Múltipla	Árvore de Regressão	Árvore de Modelos	ANN	<i>Random Forest</i>
1	Corte	Corte	Corte	Estiagem.I	Corte
2	Variedade	GrausDias.S.I	Veranico.I	K/CTC	Variedade
3	Precipitação.M.I	GrausDias.S.II	Estiagem.I	K/CaMg	GrausDias.S.I
4	GrausDias.S.I	GrausDias.S.III	GrausDias.S.I	TMin.M.III	TMed.M.IV
5	Precipitação.M.II	Precipitação.S.II	Veranico.III	Estiagem.IV	TMed.M.IV
6	Textura	pH	pH	Precipitação.M.II	Ciclo de Crescimento [Dias]
7	Saturação de Alumínio	Argila.2aCamada	Variedade	Veranico.III	Precipitação.M.IV
8	TMax.M.I	Argila.1aCamada	Estiagem.III	SiBCS	Argila.2aCamada
9	Precipitação.S.I	Areia.2aCamada	GrausDias.S.III	Gradiente Textural	TMax.M.I
10	Precipitação.S.IV	Argila.3aCamada	TMax.M.IV	GrausDias.S.III	Ambiente de Manejo
11	Matéria Orgânica	Areia.1aCamada	EpocaCorte	EpocaCorte	TMax.M.IV
12	Densidade	Areia.3aCamada	Precipitação.S.IV	Corte	Ca/Mg
13	Precipitação.M.III	Veranico.I	tmin.media.III	Precipitação.M.III	Precipitação.S.I
14	Veranico.I	Precipitação.M.I	Estiagem.II	Estiagem.III	K
15	Estiagem.II	Soma de Bases	Veranico.II	Veranico.IV	Precipitação.M.III

#### 4.4.2 Modelagem

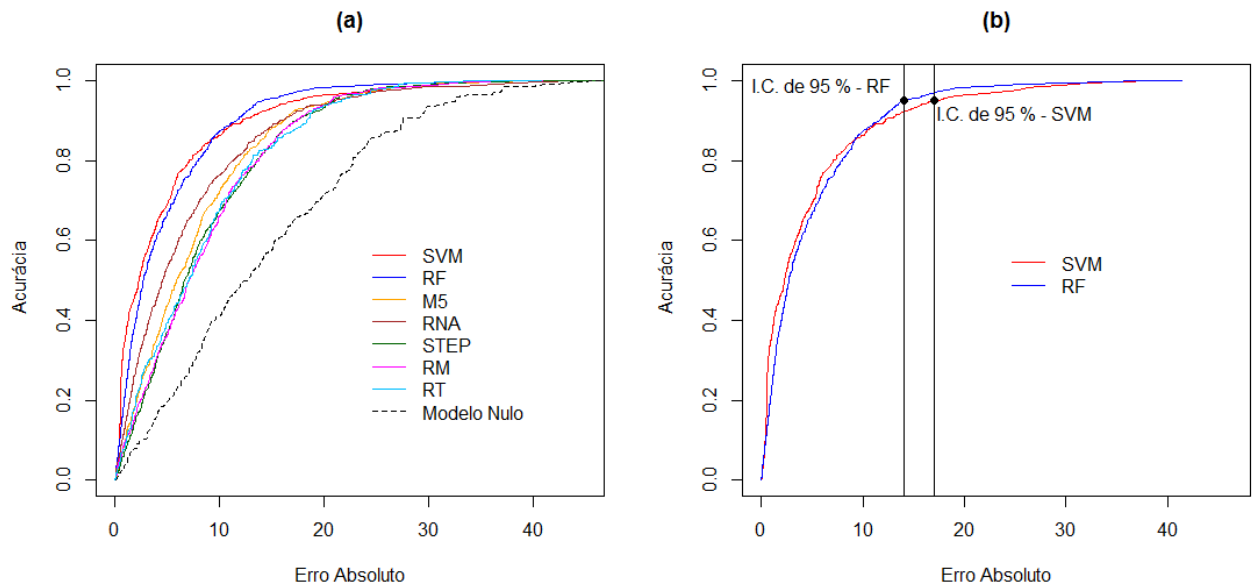
Reforçando os resultados encontrados em outras áreas de aplicação, a *Random Forest* e a SVM se mostraram técnicas de modelagem mais eficientes. O MAE de ambas pode ser considerado equivalente (Tabela 8), havendo pequeno viés de superestimava na SVM, que possui também um menor coeficiente de determinação (0,79), quando comparada à *Random Forest* (0,84). De forma geral, o desempenho dos demais modelos se mostra compatível com os resultados de modelagem obtidos em outras culturas (JI et al., 2007; KAUL et al., 2005; PARK et al., 2005), com exceção do desempenho inferior para a árvore de regressão. Nos trabalhos de Ji et al. (2007) e Kaul et al. (2005), onde houve comparação direta entre a regressão múltipla e a árvore de regressão e RNA, os autores não especificam se houve seleção de atributos para regressão múltipla. No presente trabalho, a realização da regressão *stepwise* e a heurística utilizada para seleção de atributos foram capazes de reduzir o erro da regressão, que sem nenhum dos procedimentos teria MAE igual a 8,67 e R<sup>2</sup> de 0,37.



**Tabela 8. Medidas de avaliação dos modelos criados.**

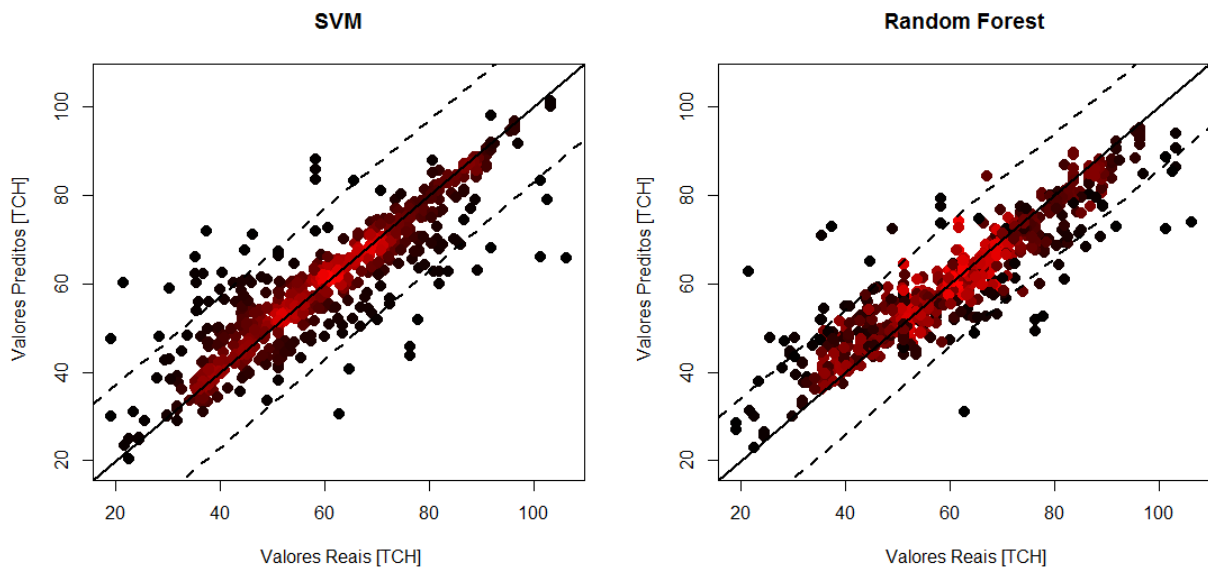
<b>Técnica</b>	<b>MAE</b>	<b>MBE</b>	<b>R<sup>2</sup></b>
<b>SVM</b>	4,61	0,0004	0,79
<b><i>Random Forest</i></b>	4,67	-0,14	0,84
<b>ANN</b>	6,76	-1,28	0,71
<b>Árvore de Modelos</b>	7,42	-0,25	0,67
<b>Regressão Múltipla</b>	8,06	-0,48	0,62
<b>Stepwise</b>	8,09	-0,55	0,63
<b>Árvore de Regressão</b>	8,37	-0,97	0,59

Na análise das curvas REC, modelos superiores se concentram na parte superior à esquerda do gráfico (BI; BENNETT, 2003), sendo evidente a dominância da RF e da SVM. Cabe destacar que os dois modelos se alternam, sendo a SVM dominante para erros de até 8,35, quando passa a ser dominada pela RF. Sendo a curva REC uma estimativa empírica da função de densidade de probabilidade acumulada, para acurácia de 95 % têm-se o Intervalo de Confiança de 95 % dos modelos, que corresponde a 14,08 para *Random Forest* e 17,06 para SVM. Intervalos de confiança menos conservadores, como o uso de um desvio padrão (68 % em uma distribuição normal), favoreceriam a SVM. Ao compararmos a curva REC dos modelos gerados com o modelo nulo, que corresponde a prever os valores pela média global, têm-se uma informação visual do ganho ao empregar as técnicas de modelagem. Do ponto de vista do modelo mais elementar para produtividade da cana-de-açúcar, seria mais coerente utilizar o número de cortes para prever a produtividade, situação na qual teríamos uma curva entre as dos modelos de pior desempenho (RM, RT e Stepwise) e a curva do modelo nulo, com um MAE de 9,58 e R<sup>2</sup> de 0,49.



**Figura 19.** Curvas REC dos modelos avaliados no conjunto de teste em comparação com o modelo nulo (a) e comparação da SVM e RF com destaque para o intervalo de confiança de 95 % (b).

A dominância da SVM para erros menores fica explícita no gráfico de valores preditos em função dos valores reais (Figura 20) no qual os pontos foram coloridos de vermelho de acordo com a densidade de pontos. É possível notar mais pontos vermelhos próximos a reta 1:1 para SVM em comparação com a *Random Forest*. Por outro lado, a SVM possui pontos mais distantes da reta 1:1 e para fora do intervalo de confiança de 95 %, o que é visto com a *Random Forest* dominando no gráfico de curva REC para erros maiores.



**Figura 20.** Valores reais e preditos para SVM e Random Forest, coloridos de acordo com a densidade de pontos, indicação da reta 1:1 (contínuo) e intervalos de confiança a 95 % (tracejado).

Em comparação com os resultados obtidos por Brüggemann et al. (2001), que obtiveram um intervalo de confiança de 36,8 ao modelar a TCH com  $R^2$  igual 0,55, foi obtido um ganho no intervalo de predição proporcionado pelo uso de técnicas mais avançadas. Comparando com os resultados da regressão múltipla, que apresentou um intervalo de confiança igual a 20,48, obteve-se um resultado 45 % menor. Esse ganho de acurácia na comparação de modelos gerados pela mesma técnica pode ser atribuído a um maior detalhamento dos atributos relacionados ao clima, disponibilidade de mais atributos ou pela menor janela temporal dos dados utilizados neste estudo. O uso de dois anos pode ser visto como uma deficiência na modelagem, porém deve ser considerada a exposição a diferentes climas ao longo de crescimento proporcionado pelo fato da safra durar nove meses e do plantio ser realizado durante o ano todo, fazendo com que talhões sejam expostos a diversas condições de clima, mesmo sendo colhidos no mesmo ano. Para ilustrar essa situação, basta fixar uma janela temporal de 10 a 18 meses ao longo da Figura 17. Por outro lado, alterações no manejo e a introdução de novas tecnologias fazem com que o sistema modelado – a produção agrícola – seja considerado não só um sistema estocástico, mas também dinâmico. Nesse contexto, com a criação de um modelo a partir de dados de uma janela temporal muito extensa, pode-se incorporar dados que não refletem mais a condição do sistema. No caso da incorporação desses dados, seria necessário sinalizar para o modelo os diferentes contextos de produção através do uso da safra (BRÜGGEMANN et al., 2001) ou de outro atributo análogo.

De forma geral, a utilização de técnicas mais avançadas levou a melhores resultados de modelagem, chamando atenção para a aplicação da SVM e *Random Forest* na modelagem de produtividade. Devido à natureza dos dados, e ao volume final de atributos, a abordagem do problema no framework de mineração de dados se mostrou necessária para criar modelos a partir dos dados disponíveis. Uma vez que os dados de clima utilizados correspondem ao clima ocorrido, a performance dos modelos para predição estará sujeita a incerteza da previsão climática utilizada e deverá ser analisada em trabalhos futuros, etapa considerada necessária para que os modelos possam ser utilizados como ferramenta de suporte no processo de planejamento. Com a disponibilização de dados de mais safras, essa avaliação deve ser viabilizada.

## 4.5 Conclusão

A avaliação da importância dos diferentes atributos mostrou que a interação entre eles e a técnica de modelagem utilizada faz com que devam ser recomendados diferentes subconjuntos para cada técnica.

A adoção de técnicas mais modernas para modelar a produtividade proporcionou ganhos na capacidade preditiva de modelos empíricos, que pode ser melhorada com a adoção da SVM ou da RF. Embora a *Random Forest* e SVM tenham apresentado desempenho similar, o número de atributos utilizados com *Random Forest* é quase o dobro, fazendo com que SVM possa ser considerado um modelo mais enxuto, capaz de realizar previsões a partir de menos atributos. O uso dessas técnicas se justifica, pois as de desempenho inferior não obtiveram performance superior à simples previsão pela média da produtividade em função do número de cortes.

A estratégia de modelagem utilizada pode ser empregada para gerar modelos *ad hoc* para previsão de produtividade a partir dos dados da usina, refletindo suas características de produção e adaptado à previsão no seu contexto, por força da estratégia utilizada. Por ser criada a partir de dados de talhões, a unidade fundamental de manejo em uma usina, o modelo pode ser utilizado para avaliar decisões desde a escala de talhões à escala geral da usina, através de uma estimativa *bottom-up*.

O framework de descoberta de conhecimento em conjuntos de dados se mostrou adequado para modelar a produtividade de cana-de-açúcar a partir de dados de produção e de clima. O modelo de produtividade deverá ser avaliado para utilização em conjunto com projeções de clima para previsão da produtividade.

## Referências Bibliográficas

- AFFHOLDER, F.; TITTONELL, P.; CORBEELS, M.; et al. Ad Hoc Modeling in Agronomy: What Have We Learned in the Last 15 Years? **Agronomy Journal**, v. 104, n. 3, p. 735. doi: 10.2134/agronj2011.0376, 2012.
- ALVAREZ, J.; CRANE, D. R.; SPREEN, T. H.; KIDDER, G. A yield prediction model for Florida sugarcane. **Agricultural Systems**, v. 9, n. 3, p. 161–179. doi: 10.1016/0308-521X(82)90018-X, 1982.
- BEZUIDENHOUT, C.; BAIER, T. A global review and synthesis of literature pertaining to integrated sugarcane production systems. Proceedings of the 82nd Annual Congress of the South African Sugar Technologists' Association, Durban, South Africa, 26-28 August 2009. **Anais...** p.93–101, 2009.
- BEZUIDENHOUT, C. N.; SINGELS, A. Operational forecasting of South African sugarcane production: Part 1 – System description. **Agricultural Systems**, v. 92, n. 1-3, p. 23–38. doi: 10.1016/j.agsy.2006.02.001, 2007.
- BI, J.; BENNETT, K. P. Regression Error Characteristic Curves. Proceedings of the 20th International Conference on Machine Learning. **Anais...** p.43–50, 2003.
- BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32. doi: 10.1023/A:1010933404324, 2001.
- BRÜGGEMANN, E.; KLUG, J.; GREENFIELD, P.; DICKS, H. Empirical modelling and prediction of sugarcane yields from field records. Proc S Afr Sug Technol Ass. **Anais...** p.75, 2001.
- CHAPMAN, P.; CLINTON, R.; KERBER, R.; et al. CRISP-DM 1.0 - Step-by-step data mining guide. . SPSS Inc., 2000.
- DIAS, F. L. F.; MAZZA, J. A.; MATSUOKA, S.; PERECIN, D.; MAULE, R. F. Produtividade da cana-de-açúcar em relação a clima e solos da região noroeste do Estado de São Paulo. **Revista Brasileira de Ciência do Solo**, v. 23, p. 627–634, 1999.
- EBRAHIM, M. K.; ZINGSHEIM, O.; EL-SHOUBAGY, M. N.; MOORE, P. H.; KOMOR, E. Growth and sugar storage in sugarcane grown at temperatures below and above optimum. **Journal of Plant Physiology**, v. 153, n. 5-6, p. 593–602. doi: 10.1016/S0176-1617(98)80209-5, 1998.
- EVERINGHAM, Y. .; MUCHOW, R. .; STONE, R. .; et al. Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. **Agricultural Systems**, v. 74, n. 3, p. 459–477. doi: 10.1016/S0308-521X(02)00050-1, 2002.
- FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874. Retrieved December 3, 2012, , 2006.

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, n. 2–3, p. 149–157. doi: 10.1016/j.fcr.2009.02.014, 2009.

FRIED, R.; SCHETTLINGER, K.; BOROWSKI, M. **robfilter: Robust Time Series Filters**. < <http://CRAN.R-project.org/package=robfilter>>, 2012.

FROHLICH, H.; CHAPELLE, O.; SCHOLKOPF, B. Feature selection for support vector machines by means of genetic algorithm. Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on. **Anais...** p.142–148, 2003.

GILBERT, R. A.; SHINE, J. M.; MILLER, J. D.; RICE, R. W.; RAINBOLT, C. R. The effect of genotype, environment and time of harvest on sugarcane yields in Florida, USA. **Field Crops Research**, v. 95, n. 2-3, p. 156–170. doi: 10.1016/j.fcr.2005.02.006, 2006.

HARRISON, S. R. Regression of a model on real-system output: An invalid test of model validity. **Agricultural Systems**, v. 34, n. 3, p. 183–190, 1990.

HIGGINS, A.; THORBURN, P.; ARCHER, A.; JAKKU, E. Opportunities for value chain research in sugar industries. **Agricultural Systems**, v. 94, n. 3, p. 611–621. doi: 10.1016/j.agsy.2007.02.011, 2007.

HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. **Agricultural and Forest Meteorology**, v. 103, n. 1-2, p. 137–157. doi: 10.1016/S0168-1923(00)00108-8, 2000.

INMAN-BAMBER, N. G. Temperature and seasonal effects on canopy development and light interception of sugarcane. **Field Crops Research**, v. 36, n. 1, p. 41–51. doi: 10.1016/0378-4290(94)90051-5, 1994.

JL, B.; SUN, Y.; YANG, S.; WAN, J. Artificial neural networks for rice yield prediction in mountainous regions. **The Journal of Agricultural Science**, v. 145, n. 03, p. 249–261. doi: 10.1017/S0021859606006691, 2007.

JIMÉNEZ, D.; PÉREZ-URIBE, A.; SATIZÁBAL, H.; et al. A Survey of Artificial Neural Network-Based Modeling in Agroecology. In: B. Prasad (Ed.); **Soft Computing Applications in Industry**. v. 226, p.247–269. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

KAUL, M.; HILL, R. L.; WALTHALL, C. Artificial neural networks for corn and soybean yield prediction. **Agricultural Systems**, v. 85, n. 1, p. 1–18. doi: 10.1016/j.agsy.2004.07.009, 2005.

KEATING, B. .; CARBERRY, P. .; HAMMER, G. .; et al. An overview of APSIM, a model designed for farming systems simulation. **European Journal of Agronomy**, v. 18, n. 3-4, p. 267–288. doi: 10.1016/S1161-0301(02)00108-9, 2003.

KUHN, M. Building Predictive Models in R Using the caret Package. **Journal of Statistical Software**, v. 28, n. 5, p. 1–26, 2008.

KUHN, M.; WESTON, S.; KEEFER, C.; QUINLAN, N. C. C. CODE FOR C. BY R. **Cubist: Rule and Instance-Based Regression Modeling**. <<http://CRAN.R-project.org/package=Cubist>>, 2013.

LAWES, R. A.; LAWN, R. J. Applications of industry information in sugarcane production systems. **Field Crops Research**, v. 92, n. 2–3, p. 353–363. doi: 10.1016/j.fcr.2005.01.033, 2005.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002.

LIMAS, M. C.; MERÉ, J. B. O.; MARCOS, A. G.; et al. **AMORE: A MORE flexible neural network package**. <<http://CRAN.R-project.org/package=AMORE>>, 2010.

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, n. 2–3, p. 321–335. doi: 10.1016/j.fcr.2005.01.010, 2005.

LIU, D. L.; KINGSTON, G.; BULL, T. A. A new technique for determining the thermal parameters of phenological development in sugarcane, including suboptimum and supra-optimum temperature regimes. **Agricultural and Forest Meteorology**, v. 90, n. 1-2, p. 119–139. doi: 10.1016/S0168-1923(97)00087-7, 1998.

MEINKE, H.; STONE, R. C. Seasonal and Inter-Annual Climate Forecasting: The New Tool for Increasing Preparedness to Climate Variability and Change In Agricultural Planning And Operations. **Climatic Change**, v. 70, n. 1-2, p. 221–253. doi: 10.1007/s10584-005-5948-6, 2005.

MEYER, D. **Support Vector Machines. The Interface to libsvm in package e1071. Online-Documentation of the package e1071 for R**. Wien: Technische Universität. < <http://cran.r-project.org/web/packages/e1071/index.html> >, 2012.

MITCHELL, P. L. Misuse of regression for empirical validation of models. **Agricultural Systems**, v. 54, n. 3, p. 313 – 326, 1997.

ORLANDO FILHO, J.; BITTENCOURT, V. C.; CARMELLO, Q. A. . .; BEAUCLAIR, E. G. F. Relações K, Ca e Mg de solo areia quartzosa e produtividade da cana-de-açúcar. **Stab Açúcar, Álcool e Subprodutos**, v. 14, n. 5, p. 13–17, 1996.

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, n. 1, p. 59–81. doi: 10.1016/j.agsy.2004.06.021, 2005.

PORTO DE CARVALHO, J. R.; DELGADO ASSAD, E.; MEDEIROS EVANGELISTA, S. R.; DA SILVEIRA PINTO, H. Estimation of dry spells in three Brazilian regions — Analysis of extremes. **Atmospheric Research**, v. 132-133, p. 12–21. doi: 10.1016/j.atmosres.2013.04.003, 2013.

PYLE, D. **Data preparation for data mining**. San Francisco, Calif.: Morgan Kaufmann Publishers, 1999.

QUINLAN, J. R. Combining Instance-Based and Model-Based Learning. ICML. **Anais...** p.236–243, 1993.

RAMBURAN, S.; PARASKEVOPOULOS, A.; SAVILLE, G.; JONES, M. A decision support system for sugarcane variety selection in South Africa based on genotype-by-environment analyses. **Experimental Agriculture**, v. 46, n. 02, p. 243. doi: 10.1017/S001447970999086X, 2009.

RAMBURAN, S.; ZHOU, M.; LABUSCHAGNE, M. Interpretation of genotype×environment interactions of sugarcane: Identifying significant environmental factors. **Field Crops Research**, v. 124, n. 3, p. 392–399. doi: 10.1016/j.fcr.2011.07.008, 2011.

REIS JUNIOR, R. DOS A. Probabilidade de resposta da cana-de-açúcar à adubação potássica em razão da relação K (Ca+Mg) do solo. **Pesquisa Agropecuária Brasileira**, v. 36, n. 9, p. 1175–1183, 2001.

RUß, G. Data mining of agricultural yield data: A comparison of regression models. **Advances in Data Mining. Applications and Theoretical Aspects**, p. 24–37. , 2009.

SANTOS, H. DOS; JACOMINE, P.; ANJOS, L. DOS; et al. Sistema brasileiro de classificação de solos. ,2006.

SCHÖLKOPF, B.; SMOLA, A. J.; WILLIAMSON, R. C.; BARTLETT, P. L. New Support Vector Algorithms. **Neural Comput.**, v. 12, n. 5, p. 1207–1245. doi: 10.1162/089976600300015565, 2000.

TEAM, R. CORE. **R: A Language and Environment for Statistical Computing**. Vienna, Austria. < <http://www.R-project.org/> >, 2013.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. **rpart: Recursive Partitioning**, 2012.

VAPNIK, V.; GOLOWICH, S. E.; SMOLA, A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. *Advances in Neural Information Processing Systems 9*. **Anais...** p.281–287. MIT Press, 1996.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. **Pattern Recognition**, v. 44, n. 2, p. 330–349. doi: 10.1016/j.patcog.2010.08.011, 2011.

WILLIGHAGEN, E. **genalg: R Based Genetic Algorithm**. < <http://CRAN.R-project.org/package=genalg> >, 2012.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate Research**, v. 30, n. 1, p. 79–82, 2005.

ZHENG, H.; CHEN, L.; HAN, X.; ZHAO, X.; MA, Y. Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. **Agriculture, Ecosystems & Environment**, v. 132, n. 1-2, p. 98–105. doi: 10.1016/j.agee.2009.03.004, 2009.



## 5 CONSIDERAÇÕES FINAIS

O modelo de produtividade construído na Seção 4 pode ser interpretado como o modelo inicial que comporia o framework proposto na Seção 3. Por ser um modelo que utilizou somente informações disponíveis antes da instalação da cultura e dados de clima, é possível que seja utilizado em conjunto com projeções de clima para realização de projeções de produtividade. A avaliação do modelo permitiu o estudo do erro de modelagem, porém é necessário uma avaliação do erro de projeção que trará mais um componente de erro, a incerteza da projeção climática. Atualmente, não há sentido em comparar o erro do modelo com o erro dos especialistas, uma vez que o erro de modelagem foi calculado a partir do clima ocorrido, que corresponde a utilizar uma projeção perfeita. O acréscimo de informações de sensoriamento remoto no modelo traz potencial para diminuição de seu erro e é uma estratégia de modelagem que pode ser explorada. É esperado que ao longo do tempo o erro diminua, devido à substituição de parte da série de clima projetada pelo clima ocorrido (HOOGENBOOM, 2000).

Embora uma melhor performance do modelo possa ser uma consequência esperada da substituição da série de clima e dos dados de sensoriamento remoto, cabe destacar os resultados de Danese e Kalchschmidt (2011), onde um menor erro de projeções de demanda não levou necessariamente a uma melhor performance operacional para cadeias de valor de manufatura. Esse resultado indica que uma busca apenas por menores erros de projeção pode não trazer benefícios para empresa, assim como a criação de um modelo com objetivo apenas de produzir um erro menor que o obtido atualmente por especialistas pode ser em vão. Assim, os benefícios do uso de um modelo podem estar mais ligados à geração automatizada de projeções, com potencial para atualizações mais frequentes e inclusão das tendências de clima na projeção. Outra vantagem para uma empresa que possui diversas usinas é a padronização do método de estimativas.

A variabilidade do clima traz impactos (positivos ou negativos) e usufruir dessa informação pode ser um fator de sucesso na agricultura. Se de um lado da discussão sobre o clima, temos as políticas públicas pautadas pela mudança climática, temos de outro lado as iniciativas em relação à variabilidade climática que começam a ser tomadas pela iniciativa privada (SURMINSKI, 2013). Uma das formas de reduzir o risco climático é incorporar projeções climáticas ao

planejamento, o que apresenta alguns desafios (COELHO; COSTA, 2010), boa parte ligados ao distanciamento entre as instituições que produzem as projeções e os possíveis usuários finais, uma relação que deveria ser intermediada por especialistas. O uso da projeção climática no modelo de produtividade pode ser vista também como uma forma de intermediar o resultado, mostrando direto ao tomador de decisão o impacto da projeção de clima na sua atividade fim. Em outras escalas, temos também o uso dessa informação em modelos de planejamento, mostrando as consequências além da própria produtividade. Os modelos, se implementados em sistemas automatizados que disponibilizem a informação para os gestores, passam a ser os intermediários entre a informação de clima e o tomador de decisão.

## REFERÊNCIAS

- ABDEL-RAHMAN, E. M.; AHMED, F. B. The application of remote sensing techniques to sugarcane ( *Saccharum* spp. hybrid) production: a review of the literature. **International Journal of Remote Sensing**, v. 29, n. 13, p. 3753–3767. doi: 10.1080/01431160701874603, 2008.
- AFFHOLDER, F.; TITTONELL, P.; CORBEELS, M.; et al. Ad Hoc Modeling in Agronomy: What Have We Learned in the Last 15 Years? **Agronomy Journal**, v. 104, n. 3, p. 735. doi: 10.2134/agronj2011.0376, 2012.
- AHUMADA, O.; VILLALOBOS, J. R. Application of planning models in the agri-food supply chain: A review. **European Journal of Operational Research**, v. 195, n. 1, p. 1–20. doi: 10.1016/j.ejor.2008.02.014, 2009.
- ALVAREZ, J.; CRANE, D. R.; SPREEN, T. H.; KIDDER, G. A yield prediction model for Florida sugarcane. **Agricultural Systems**, v. 9, n. 3, p. 161–179. doi: 10.1016/0308-521X(82)90018-X, 1982.
- ANDERSON, D. L.; PORTIER, K. M.; OBREZA, T. A.; COLLINS, M. E.; PITTS, D. J. Tree Regression Analysis To Determine Effects Of Soil Variability On Sugarcane Yields. **Soil Sci. Soc. Am. J.**, v. 63, n. 3, p. 592–600, 1999.
- ARGENTON, P. E.; BEAUCLAIR, E. G. F.; SCARPARI, M. S. Modelagem de variáveis climáticas, edáficas e de manejo para a predição de produtividade de cana-de-açúcar. In: C. A. C. Crusciol; M. de A. Silva; R. Rosseto; R. P. Soratto (Eds.); **Tópicos em ecofisiologia da cana-de-açúcar**. p.22–26. Botucatu: FEPAF, 2010.
- BEZUIDENHOUT, C.; BAIER, T. A global review and synthesis of literature pertaining to integrated sugarcane production systems. Proceedings of the 82nd Annual Congress of the South African Sugar Technologists' Association, Durban, South Africa, 26-28 August 2009. **Anais...** p.93–101, 2009.
- BEZUIDENHOUT, C. N.; SINGELS, A. Operational forecasting of South African sugarcane production: Part 1 – System description. **Agricultural Systems**, v. 92, n. 1-3, p. 23–38. doi: 10.1016/j.agsy.2006.02.001, 2007.
- BI, J.; BENNETT, K. P. Regression Error Characteristic Curves. Proceedings of the 20th International Conference on Machine Learning. **Anais...** p.43–50, 2003.
- BINBOL, N.; ADEBAYO, A.; KWON-NDUNG, E. Influence of climatic factors on the growth and yield of sugar cane at Numan, Nigeria. **Climate Research**, v. 32, p. 247–252. doi: 10.3354/cr032247, 2006.
- BONNETT, G. D. Rate of leaf appearance in sugarcane, including a comparison of a range of varieties. **Australian Journal of Plant Physiology**, v. 25, n. 7, p. 829. doi: 10.1071/PP98041, 1998.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32. doi: 10.1023/A:1010933404324, 2001.

BRÜGGEMANN, E.; KLUG, J.; GREENFIELD, P.; DICKS, H. Empirical modelling and prediction of sugarcane yields from field records. Proc S Afr Sug Technol Ass. **Anais...** p.75, 2001.

BRUGNARO, C.; SBRAGIA, R. **Gerência Agrícola em Destilarias de Álcool**. Piracicaba, 1982.

CHAPMAN, P.; CLINTON, R.; KERBER, R.; et al. CRISP-DM 1.0 - Step-by-step data mining guide. . SPSS Inc., 2000.

COELHO, C. A.; COSTA, S. M. Challenges for integrating seasonal climate forecasts in user applications. **Current Opinion in Environmental Sustainability**, v. 2, n. 5-6, p. 317–325. doi: 10.1016/j.cosust.2010.09.002, 2010.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da Safra Brasileira: Cana-de-Açúcar**, 2º Levantamento. Brasília, 2013.

DANESE, P.; KALCHSCHMIDT, M. The role of the forecasting process in improving forecast accuracy and operational performance. **International Journal of Production Economics**, v. 131, n. 1, p. 204–214. doi: 10.1016/j.ijpe.2010.09.006, 2011.

DIAS, F. L. F.; MAZZA, J. A.; MATSUOKA, S.; PERECIN, D.; MAULE, R. F. Produtividade da cana-de-açúcar em relação a clima e solos da região noroeste do Estado de São Paulo. **Revista Brasileira de Ciência do Solo**, v. 23, p. 627–634, 1999.

EBRAHIM, M. K.; ZINGSHEIM, O.; EL-SHOUBAGY, M. N.; MOORE, P. H.; KOMOR, E. Growth and sugar storage in sugarcane grown at temperatures below and above optimum. **Journal of Plant Physiology**, v. 153, n. 5-6, p. 593–602. doi: 10.1016/S0176-1617(98)80209-5, 1998.

EVERINGHAM, Y. .; MUCHOW, R. .; STONE, R. .; et al. Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. **Agricultural Systems**, v. 74, n. 3, p. 459–477. doi: 10.1016/S0308-521X(02)00050-1, 2002.

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, n. 3-4, p. 689–696. doi: 10.1016/j.agrformet.2008.10.018, 2009.

FAO - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. FAOSTAT database. . <[http://faostat3.fao.org/faostat-gateway/go/to/browse/rankings/countries\\_by\\_commodity/E](http://faostat3.fao.org/faostat-gateway/go/to/browse/rankings/countries_by_commodity/E)>, 30/11/ 2013.

FRANCO, H.C.J., OTTO, R., FARONI, C.E., VITTI, A.C., ALMEIDA DE OLIVEIRA, E.C., TRIVELIN, P.C.O., 2011. Nitrogen in sugarcane derived from fertilizer under Brazilian field conditions. **Field Crops Research** 121, 29–41. doi:10.1016/j.fcr.2010.11.011

- FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874. Retrieved December 3, 2012, , 2006.
- 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, n. 2–3, p. 149–157. doi: 10.1016/j.fcr.2009.02.014, 2009.
- FRIED, R.; SCHETTLINGER, K.; BOROWSKI, M. **robfilter: Robust Time Series Filters**. <<http://CRAN.R-project.org/package=robfilter>>, 2012.
- FROHLICH, H.; CHAPELLE, O.; SCHOLKOPF, B. Feature selection for support vector machines by means of genetic algorithm. Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on. **Anais...** p.142–148, 2003.
- GILBERT, R. A.; SHINE, J. M.; MILLER, J. D.; RICE, R. W.; RAINBOLT, C. R. The effect of genotype, environment and time of harvest on sugarcane yields in Florida, USA. **Field Crops Research**, v. 95, n. 2-3, p. 156–170. doi: 10.1016/j.fcr.2005.02.006, 2006.
- GLASZIOU, K.; BULL, T.; HATCH, M.; WHITEMAN, P. Physiology of Sugar-Cane VII. Effects of Temperature, Photoperiod Duration, and Diurnal and Seasonal Temperature Changes on Growth and Ripening. **Australian Journal of Biological Sciences**, v. 18, n. 1, p. 53–66, 1965.
- GREENLAND, D. Climate Variability and Sugarcane Yield in Louisiana. **Journal of Applied Meteorology**, v. 44, n. 11, p. 1655–1666. doi: 10.1175/JAM2299.1, 2005.
- GRUNOW, M.; GÜNTHER, H.-O.; WESTINNER, R. Supply optimization for the production of raw sugar. **International Journal of Production Economics**, v. 110, n. 1-2, p. 224–239. doi: 10.1016/j.ijpe.2007.02.019, 2007.
- HAN, J.; KAMBER, M.; PEI, J. **Data mining : concepts and techniques**. Amsterdam; Boston: Elsevier/Morgan Kaufmann, 2012.
- HARRISON, S. R. Regression of a model on real-system output: An invalid test of model validity. **Agricultural Systems**, v. 34, n. 3, p. 183–190, 1990.
- HIGGINS, A. J. Australian Sugar Mills Optimize Harvester Rosters to Improve Production. **Interfaces**, v. 32, n. 3, p. 15–25. doi: 10.1287/inte.32.3.15.41, 2002.
- HIGGINS, A.; THORBURN, P.; ARCHER, A.; JAKKU, E. Opportunities for value chain research in sugar industries. **Agricultural Systems**, v. 94, n. 3, p. 611–621. doi: 10.1016/j.agry.2007.02.011, 2007.
- HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. **Agricultural and Forest Meteorology**, v. 103, n. 1-2, p. 137–157. doi: 10.1016/S0168-1923(00)00108-8, 2000.
- IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Levantamento Sistemático da Produção Agrícola. . <[http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/lspa/lspa\\_201310\\_1.shtm](http://www.ibge.gov.br/home/estatistica/indicadores/agropecuaria/lspa/lspa_201310_1.shtm)>, 25/11/2013.

INMAN-BAMBER, N. G. Temperature and seasonal effects on canopy development and light interception of sugarcane. **Field Crops Research**, v. 36, n. 1, p. 41–51. doi: 10.1016/0378-4290(94)90051-5, 1994.

INMAN-BAMBER, N. G.; SMITH, D. M. Water relations in sugarcane and response to water deficits. **Field Crops Research**, v. 92, n. 2-3, p. 185–202. doi: 10.1016/j.fcr.2005.01.023, 2005.

VAN ITTERSUM, M. .; LEFFELAAR, P. .; VAN KEULEN, H.; et al. On approaches and applications of the Wageningen crop models. **European Journal of Agronomy**, v. 18, n. 3-4, p. 201–234. doi: 10.1016/S1161-0301(02)00106-5, 2003.

JENA, S. D.; POGGI, M. Harvest planning in the Brazilian sugar cane industry via mixed integer programming. **European Journal of Operational Research**, v. 230, n. 2, p. 374–384. doi: 10.1016/j.ejor.2013.04.011, 2013.

JIAO, Z.; HIGGINS, A. J.; PRESTWIDGE, D. B. An integrated statistical and optimisation approach to increasing sugar production within a mill region. **Computers and Electronics in Agriculture**, v. 48, n. 2, p. 170–181. doi: 10.1016/j.compag.2005.03.004, 2005.

Ji, B.; SUN, Y.; YANG, S.; WAN, J. Artificial neural networks for rice yield prediction in mountainous regions. **The Journal of Agricultural Science**, v. 145, n. 03, p. 249–261. doi: 10.1017/S0021859606006691, 2007.

JIMÉNEZ, D.; PÉREZ-URIBE, A.; SATIZÁBAL, H.; et al. A Survey of Artificial Neural Network-Based Modeling in Agroecology. In: B. Prasad (Ed.); **Soft Computing Applications in Industry**. v. 226, p.247–269. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

KAUL, M.; HILL, R. L.; WALTHALL, C. Artificial neural networks for corn and soybean yield prediction. **Agricultural Systems**, v. 85, n. 1, p. 1–18. doi: 10.1016/j.agsy.2004.07.009, 2005.

KEATING, B. .; CARBERRY, P. .; HAMMER, G. .; et al. An overview of APSIM, a model designed for farming systems simulation. **European Journal of Agronomy**, v. 18, n. 3-4, p. 267–288. doi: 10.1016/S1161-0301(02)00108-9, 2003.

KUHN, M. Building Predictive Models in R Using the caret Package. **Journal of Statistical Software**, v. 28, n. 5, p. 1–26, 2008.

KUHN, M.; WESTON, S.; KEEFER, C.; QUINLAN, N. C. C. CODE FOR C. BY R. **Cubist: Rule and Instance-Based Regression Modeling**. <<http://CRAN.R-project.org/package=Cubist>>, 2013.

LAWES, R. A.; LAWN, R. J. Applications of industry information in sugarcane production systems. **Field Crops Research**, v. 92, n. 2–3, p. 353–363. doi: 10.1016/j.fcr.2005.01.033, 2005.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002.

LIMAS, M. C.; MERÉ, J. B. O.; MARCOS, A. G.; et al. **AMORE: A MORE flexible neural network package**. <<http://CRAN.R-project.org/package=AMORE>>, 2010.

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, n. 2–3, p. 321–335. doi: 10.1016/j.fcr.2005.01.010, 2005.

LIU, D. L.; KINGSTON, G.; BULL, T. A. A new technique for determining the thermal parameters of phenological development in sugarcane, including suboptimum and supra-optimum temperature regimes. **Agricultural and Forest Meteorology**, v. 90, n. 1-2, p. 119–139. doi: 10.1016/S0168-1923(97)00087-7, 1998.

LOBELL, D. B.; BURKE, M. B. On the use of statistical models to predict crop yield responses to climate change. **Agricultural and Forest Meteorology**, v. 150, n. 11, p. 1443–1452. doi: 10.1016/j.agrformet.2010.07.008, 2010.

LUSTIG, I. J.; PUGET, J.-F. Program Does Not Equal Program: Constraint Programming and Its Relationship to Mathematical Programming. **Interfaces**, v. 31, n. 6, p. 29–53. doi: 10.1287/inte.31.6.29.9647, 2001.

MAPA - MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. Valor Bruto da Produção. <<http://www.agricultura.gov.br/comunicacao/noticias/2013/10/governo-estima-crescimento-de-9porcento-no-vbp-em-2013>>, 25/11/2013.

MARGARIDO, A. E. Planejamento agrícola em cana-de-açúcar. In: S. V. Segato; A. de S. Pinto; E. Jendiroba; J. C. M. de Nóbrega (Eds.); **Atualização em Produção de Cana-de-Açúcar**. p.69–78. Piracicaba, 2006.

MARIN, F. R.; JONES, J. W.; ROYCE, F.; 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.; JONES, J. W.; SINGELS, A.; et al. Climate change impacts on sugarcane attainable yield in southern Brazil. **Climatic Change**, v. 117, n. 1-2, p. 227–239. doi: 10.1007/s10584-012-0561-y, 2012.

MEINKE, H.; STONE, R. C. Seasonal and Inter-Annual Climate Forecasting: The New Tool for Increasing Preparedness to Climate Variability and Change In Agricultural Planning And Operations. **Climatic Change**, v. 70, n. 1-2, p. 221–253. doi: 10.1007/s10584-005-5948-6, 2005.

MEYER, D. **Support Vector Machines. The Interface to libsvm in package e1071. Online-Documentation of the package e1071 for R**. Wien: Technische Universität. < <http://cran.r-project.org/web/packages/e1071/index.html> >, 2012.

MITCHELL, P. L. Misuse of regression for empirical validation of models. **Agricultural Systems**, v. 54, n. 3, p. 313 – 326, 1997.

MOZAMBANI, A. E.; PINTO, A. DE S.; SEGATO, S. V.; MATTIUZ, C. F. M. História e morfologia da cana-de-açúcar. In: S. V. Segato; A. de S. Pinto; E. Jendiroba; J. C. M. de Nóbrega (Eds.); **Atualização em Produção de Cana-de-Açúcar**. p.11–18. Piracicaba, 2006.

MUCHERINO, A.; PAPAJORGJI, P.; PARDALOS, P. A survey of data mining techniques applied to agriculture. **Operational Research**, v. 9, n. 2, p. 121–140. doi: 10.1007/s12351-009-0054-6, 2009.

MUCHERINO, A.; RUß, G. Recent Developments in Data Mining and Agriculture. IbaI Conference Proceedings. **Anais...** p.90–98. New York, USA, 2011.

NASSIF, D. S. P.; MARIN, F. R.; PALLONE FILHO, W. J.; RESENDE, R. S.; PELLEGRINO, G. Q. Parametrização e avaliação do modelo DSSAT/Canegro para variedades brasileiras de cana-de-açúcar. **Pesq Agrop Brasileira**, v. 47, n. 3, p. 311–318, 2012.

ORLANDO FILHO, J.; BITTENCOURT, V. C.; CARMELLO, Q. A. . .; BEAUCLAIR, E. G. F. Relações K, Ca e Mg de solo areia quartzosa e produtividade da cana-de-açúcar. **Stab Açúcar, Álcool e Subprodutos**, v. 14, n. 5, p. 13–17, 1996.

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, n. 1, p. 59–81. doi: 10.1016/j.agsy.2004.06.021, 2005.

PASSIOURA, J. B. Simulation Models: Science, Snake Oil, Education, or Engineering? **Agronomy Journal**, v. 88, n. 5, p. 690. doi: 10.2134/agronj1996.00021962008800050002x, 1996.

PICOLI, M. C. A. **ESTIMATIVA DA PRODUTIVIDADE AGRÍCOLA DA CANA-DE-AÇÚCAR UTILIZANDO AGREGADOS DE REDES NEURAIS ARTIFICIAIS: ESTUDO DE CASO NA USINA CATANDUVA**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais - INPE, 2006.

PIEWTHONGNGAM, K.; PATHUMNAKUL, S.; SETTHANAN, K. Application of crop growth simulation and mathematical modeling to supply chain management in the Thai sugar industry. **Agricultural Systems**, v. 102, n. 1-3, p. 58–66. doi: 10.1016/j.agsy.2009.07.002, 2009.

POWER, D. J.; SHARDA, R. Model-driven decision support systems: Concepts and research directions. **Decision Support Systems**, v. 43, n. 3, p. 1044–1061. doi: 10.1016/j.dss.2005.05.030, 2007.

PORTO DE CARVALHO, J. R.; DELGADO ASSAD, E.; MEDEIROS EVANGELISTA, S. R.; DA SILVEIRA PINTO, H. Estimation of dry spells in three Brazilian regions — Analysis of extremes. **Atmospheric Research**, v. 132-133, p. 12–21. doi: 10.1016/j.atmosres.2013.04.003, 2013.

PYLE, D. **Data preparation for data mining**. San Francisco, Calif.: Morgan Kaufmann Publishers, 1999.

QUINLAN, J. R. Combining Instance-Based and Model-Based Learning. ICML. **Anais...** p.236–243, 1993.

RAMBURAN, S.; PARASKEVOPOULOS, A.; SAVILLE, G.; JONES, M. A decision support system for sugarcane variety selection in South Africa based on genotype-by-environment



analyses. **Experimental Agriculture**, v. 46, n. 02, p. 243. doi: 10.1017/S001447970999086X, 2009.

RAMBURAN, S.; ZHOU, M.; LABUSCHAGNE, M. Interpretation of genotype×environment interactions of sugarcane: Identifying significant environmental factors. **Field Crops Research**, v. 124, n. 3, p. 392–399. doi: 10.1016/j.fcr.2011.07.008, 2011.

REIS JUNIOR, R. DOS A. Probabilidade de resposta da cana-de-açúcar à adubação potássica em razão da relação K (Ca+Mg) do solo. **Pesquisa Agropecuária Brasileira**, v. 36, n. 9, p. 1175–1183, 2001.

ROBERTSON, M. J.; BONNETT, G. D.; HUGHES, R. M.; MUCHOW, R. C.; CAMPBELL, J. A. Temperature and leaf area expansion of sugarcane: integration of controlled-environment, field and model studies. **Australian Journal of Plant Physiology**, v. 25, n. 7, p. 819. doi: 10.1071/PP98042, 1998.

RUß, G. Data mining of agricultural yield data: A comparison of regression models. **Advances in Data Mining. Applications and Theoretical Aspects**, p. 24–37. , 2009.

SANTOS, H. DOS; JACOMINE, P.; ANJOS, L. DOS; et al. Sistema brasileiro de classificação de solos. ,2006.

SCHÖLKOPF, B.; SMOLA, A. J.; WILLIAMSON, R. C.; BARTLETT, P. L. New Support Vector Algorithms. **Neural Comput.**, v. 12, n. 5, p. 1207–1245. doi: 10.1162/089976600300015565, 2000.

SCHÖLKOPF, B.; SUNG, K.-K.; BURGESS, C. J. C.; et al. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. **Signal Processing, IEEE Transactions on**, v. 45, n. 11, p. 2758 –2765. doi: 10.1109/78.650102, 1997.

SEGATO, S. V.; MATTIUZ, C. F. M.; MOZAMBANI, A. E. Aspectos fenológicos da cana-de-açúcar. In: S. V. Segato; A. de S. Pinto; E. Jendiroba; J. C. M. de Nóbrega (Eds.); **Atualização em Produção de Cana-de-Açúcar**. p.19–38. Piracicaba, 2006.

SHARIAT, M.; NWAKANMA, H. Enterprise Resource Planning And Its Future Relationship To Decision Support System. **Journal of Business & Economics Research (JBER)**, v. 4, n. 12, 2006.

SILVA, M. DE A.; DOS SANTOS, C. M.; ARANTES, M. T.; PINCELLI, R. P. Fenologia da Cana-de-açúcar. In: C. A. C. Crusciol; M. de A. Silva; R. Rosseto; R. P. Soratto (Eds.); **Tópicos em ecofisiologia da cana-de-açúcar**. p.8–21. Botucatu: FEPAF, 2010.

SPIERTZ, H. Challenges for Crop Production Research in Improving Land Use, Productivity and Sustainability. **Sustainability**, v. 5, n. 4, p. 1632–1644. doi: 10.3390/su5041632, 2013.

STANEK, S.; SROKA, H.; TWARDOWSKI, Z. Directions for an ERP-based DSS. Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8. 3 International Conference. **Anais...** , 2004.

STONE, R. C.; MEINKE, H. Operational seasonal forecasting of crop performance. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 360, n. 1463, p. 2109–2124. doi: 10.1098/rstb.2005.1753, 2005.

SURMINSKI, S. Private-sector adaptation to climate risk. **Nature Clim. Change**, v. 3, n. 11, p. 943–945, 2013.

TEAM, R. CORE. **R: A Language and Environment for Statistical Computing**. Vienna, Austria. < <http://www.R-project.org/> >, 2013.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. **rpart: Recursive Partitioning**, 2012.

VAPNIK, V.; GOLOWICH, S. E.; SMOLA, A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. *Advances in Neural Information Processing Systems 9. Anais...* p.281–287. MIT Press, 1996.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. **Pattern Recognition**, v. 44, n. 2, p. 330–349. doi: 10.1016/j.patcog.2010.08.011, 2011.

WILLIGHAGEN, E. **genalg: R Based Genetic Algorithm**. < <http://CRAN.R-project.org/package=genalg> >, 2012.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate Research**, v. 30, n. 1, p. 79–82, 2005.

WITTEN, I. H.; FRANK, E. **Data mining : practical machine learning tools and techniques**. 2nd ed. Amsterdam; Boston, MA: Morgan Kaufman, 2005.

WEINTRAUB, A.; ROMERO, C. Operations Research Models and the Management of Agricultural and Forestry Resources: A Review and Comparison. **Interfaces**, v. 36, n. 5, p. 446–457. doi: 10.1287/inte.1060.0222, 2006.

WIEDENFELD, R. P. Effects of irrigation and N fertilizer application on sugarcane yield and quality. **Field Crops Research**, v. 43, n. 2-3, p. 101–108. doi: 10.1016/0378-4290(95)00043-P, 1995.

ZHENG, H.; CHEN, L.; HAN, X.; ZHAO, X.; MA, Y. Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. **Agriculture, Ecosystems & Environment**, v. 132, n. 1-2, p. 98–105. doi: 10.1016/j.agee.2009.03.004, 2009.