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

Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados

Ralph Guenther Hammer

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Engenharia de Sistemas Agrícolas

Piracicaba
2016

Ralph Guenther Hammer
Diplomagraringenieur

Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **PAULO CESAR SENTELHAS**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Engenharia de Sistemas Agrícolas

Piracicaba
2016

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

Hammer, Ralph Guenther
Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados / Ralph Guenther Hammer. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2016.
98 p. : il.

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

1. Predição 2. Planejamento agrícola 3. Random forest 4. Boosting 5. Support vector machines I. Título

CDD 633.61
H224m

“Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

AGRADECIMENTOS

Aos meus pais Celina e Gerhard, que jamais mediram esforços em proporcionar aos filhos uma vida digna e uma educação de qualidade. Ao meu irmão Jens, companheiro nos tempos difíceis e nos momentos de alegrias.

Ao Fernando Aparecido Benvenuti, que me proporcionou esta oportunidade ímpar em minha vida, pelo incentivo e pelo apoio na realização desse trabalho.

Ao Prof. Paulo Cesar Sentelhas, cuja competente orientação e total apoio, principalmente em momentos difíceis, em que cogitei desistir, foram decisivos para a realização desse trabalho.

Aos Professores Luiz Roberto Angelocci, Rubens Duarte Coelho e Sonia Maria de Stefano Piedade, pelos *insights* na qualificação e pelos ensinamentos em suas disciplinas. Aos Professores Gabriel Adrián Sarriés, José Laércio Favarin e Ricardo Leite Camargo, pelos ensinamentos dentro e fora das salas de aulas.

Ao Professor Stanley Robson de Medeiros Oliveira, por acreditar nesse projeto e pelo apoio. Ao Flavio Margarito Martins de Barros, pelos ensinamentos na área de Mineração de Dados, bem como pelo apoio nas análises.

À Davilmar Collevatti e à Angela Márcia Derigi Silva, que tantas vezes me apoiaram em questões administrativas durante o mestrado.

Ao Diego Cardoso, por seu apoio na revisão desse trabalho.

A todos meus colegas e amigos, pela compreensão nos momentos de minha ausência.

E, finalmente, ao Jean Carlos Quaresma Mariano, por seu apoio incondicional e imprescindível em todas as etapas do trabalho. Sem sua ajuda, essa tese simplesmente não teria sido finalizada.

SUMÁRIO

RESUMO....7
ABSTRACT....9
1 INTRODUÇÃO....11
2 REVISÃO BIBLIOGRÁFICA....15
2.1 A Cana de açúcar no Brasil....15
2.2 Fatores que afetam a produtividade da cana de açúcar....16
2.2.1 Fatores agrometeorológicos....16
2.2.2 Fatores de manejo....18
2.3 Métodos de estimação da produtividade da cana de açúcar....20
2.3.1 Levantamentos de campo....20
2.3.2 Biometrias de campo....20
2.3.3 Imagens de satélite....20
2.3.4 Modelagem de produtividade agrícola....21
2.4 Mineração de Dados....24
2.4.1 Grupo das Árvores de Decisão....28
2.4.2 Support Vector Machine (SVM)....33
2.5 O uso de Mineração de Dados na agricultura....38
3 MATERIAL E MÉTODOS....41
3.1 Entendimento dos dados empregados....41
3.2 Preparação dos dados....42
3.2.1 Estimação da produtividade com base em modelo agrometeorológico para comparação da performance com os modelos de mineração de dados....46
3.3 Modelagem da produtividade da cana de açúcar por meio da mineração de dados....54
3.4 Avaliação dos resultados....54
3.5 Distribuição....55
4 RESULTADOS E DISCUSSÃO....57
4.1 Correlações entre as variáveis....57
4.2 Determinação dos principais fatores determinantes da produtividade da cana de açúcar utilizando-se diferentes técnicas de Mineração de Dados....61
4.2.1 Base de dados integral....61
4.2.2 Base de dados regionalizada....66

4.3 Avaliação dos modelos de Mineração de Dados para a estimação da produtividade da cana de açúcar 71
4.3.1 Base de dados integral 71
4.3.2 Base de dados regionalizada..... 75
5 CONCLUSÕES..... 81
REFERÊNCIAS 83
APÊNDICE 93

RESUMO

Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados

O entendimento da hierarquia de importância dos fatores que influenciam a produtividade da cana de açúcar pode auxiliar na sua modelagem, contribuindo assim para a otimização do planejamento agrícola das unidades produtoras do setor, bem como no aprimoramento das estimativas de safra. Os objetivos do presente estudo foram a ordenação das variáveis que condicionam a produtividade da cana de açúcar, de acordo com a sua importância, bem como o desenvolvimento de modelos matemáticos de produtividade da cana de açúcar. Para tanto, foram utilizadas três técnicas de mineração de dados nas análises de bancos de dados de usinas de cana de açúcar no estado de São Paulo. Variáveis meteorológicas e de manejo agrícola foram submetidas às análises por meio das técnicas *Random Forest*, *Boosting* e *Support Vector Machines*, e os modelos resultantes foram testados por meio da comparação com dados independentes, utilizando-se o coeficiente de correlação (r), índice de Willmott (d), índice de confiança de Camargo (C), erro absoluto médio (EAM) e raíz quadrada do erro médio (RMSE). Por fim, comparou-se o desempenho dos modelos gerados com as técnicas de mineração de dados com um modelo agrometeorológico, aplicado para os mesmos bancos de dados. Constatou-se que, das variáveis analisadas, o número de cortes foi o fator mais importante em todas as técnicas de mineração de dados. A comparação entre as produtividades estimadas pelos modelos de mineração de dados e as produtividades observadas resultaram em RMSE variando de 19,70 a 20,03 t ha^{-1} na abordagem mais geral, que engloba todas as regiões do banco de dados. Com isso, o desempenho preditivo foi superior ao modelo agrometeorológico, aplicado no mesmo banco de dados, que obteve RMSE $\approx$ 70% maior ($\approx$ 34 t ha^{-1}).

Palavras-chave: Predição; Planejamento agrícola; Random forest; Boosting; Support vector machines

ABSTRACT

Modeling sugarcane yield through Data Mining techniques

The understanding of the hierarchy of the importance of the factors which influence sugarcane yield can subsidize its modeling, thus contributing to the optimization of agricultural planning and crop yield estimates. The objectives of this study were to ordinate the variables which condition the sugarcane yield, according to their relative importance, as well as the development of mathematical models for predicting sugarcane yield. For this, three Data Mining techniques were applied in the analyses of data bases of several sugar mills in the State of São Paulo, Brazil. Meteorological and crop management variables were analyzed through the Data Mining techniques Random Forest, Boosting and Support Vector Machines, and the resulting models were tested through the comparison with an independent data set, using the coefficient of correlation (r), Willmott index (d), confidence index of Camargo (c), mean absolute error (MAE), and root mean square error (RMSE). Finally, the predictive performances of these models were compared with the performance of an agrometeorological model, applied in the same data set. The results allowed to conclude that, within all the variables, the number of cuts was the most important factor considered by all Data Mining models. The comparison between the observed yields and those estimated by the Data Mining techniques resulted in a RMSE ranging between 19,70 to 20,03 t ha^{-1}, in the general method, which considered all regions of the data base. Thus, the predictive performances of the Data Mining algorithms were superior to that of the agrometeorological model, which presented RMSE $\approx$ 70% higher ($\approx$ 34 t ha^{-1}).

Keywords: Prediction; Agricultural planning; Random forest; Boosting; Support vector machines

1 INTRODUÇÃO

Contando com uma área colhida de 9,7 milhões de hectares na safra 12/13, o que rendeu uma colheita de 588,5 milhões de toneladas (UNIAO DA INDÚSTRIA DE CANA DE AÇÚCAR – UNICA - UNICA, 2014), a cana de açúcar é uma das mais importantes culturas agrícolas do Brasil, atualmente o maior produtor mundial (FAO, 2015). De acordo com Neves et al. (2009), o setor sucro-energético brasileiro movimenta, por ano, um total de US$ 80 bilhões, considerando-se a somatória das vendas estimadas de toda a cadeia produtiva do setor. Com isso, representa um setor de fundamental importância para o país, com expressiva capacidade de geração de empregos e movimentação de recursos e impostos.

A produtividade da cana de açúcar, comumente expressa em toneladas de cana por ha (TCH), depende de uma combinação de fatores, dentre eles: clima, ambientes de produção (solo) e aspectos agronômicos associados ao manejo agrícola, como a escolha de variedades, a idade do canavial, o número médio de cortes, a adubação, o controle de pragas e doenças, o manejo de plantas invasoras, a aplicação de vinhaça, entre outros.

O entendimento da hierarquia das variáveis climáticas e de manejo, de acordo com seu grau de importância na variabilidade do TCH, pode auxiliar tanto na otimização do planejamento agrícola, bem como no aprimoramento das previsões de safra. Porém, há diferentes percepções com relação à ordem de importância desses fatores. Ramburan et al. (2013), ao estudarem os fatores que causam o declínio de produtividades das soqueiras ao longo do ciclo na África do Sul, constatam que o grau de influência dos mesmos ainda não é muito claro. Apesar da importância de variáveis como pragas e doenças ou dano às soqueiras, há uma percepção geral de que as diferenças entre as variedades contribuem de forma mais acentuada na variabilidade das produtividades. Nesse trabalho, os autores concluem que uma maior ênfase deveria ser dada ao manejo da cultura do que à escolha das variedades, de modo a garantir maior longevidade dos canaviais. Já Ferraro et al. (2009), em seu trabalho de análise de bancos de dados de produtividade da cana de açúcar utilizando técnica de mineração de dados, classificaram diversas variáveis, tanto de manejo quanto ambientais, de acordo com sua importância hierárquica na variabilidade das produtividades (TCH), e concluíram que as chuvas tiveram um peso relativamente pequeno para explicar a variabilidade das produtividades quando

comparada com outros fatores não associados ao clima, como número de cortes e variedade.

As estimações de produtividade são uma ferramenta importante na agricultura, principalmente no que tange o planejamento estratégico das empresas, permitindo um melhor ajuste das atividades de compra de insumos, vendas e exportações das produções. Nesse sentido, é imperativo para o setor sucro-energético o desenvolvimento e uso contínuo de técnicas que integrem as variáveis climáticas e de manejo, de forma que os modelos gerados a partir dessa integração possam ser usados como ferramentas de gestão para otimização de recursos e das tomadas de decisão.

Em função da importância das previsões de produtividade no contexto da indústria sucro-energética, bem como a incerteza sobre com que intensidade os diferentes fatores influenciam as produtividades, este estudo permitiu formular as seguintes hipóteses:

a) A variabilidade interanual da produtividade da cana de açúcar é afetada não somente pelas condições edafo-climáticas, mas também pela interação dessas condições com a cultura e com as ações de manejo agrícola;

b) O emprego das técnicas de Mineração de Dados em bancos de dados operacionais das usinas permite elencar os principais fatores determinantes da produtividade da cana de açúcar e gerar modelos matemáticos capazes de estimar sua produtividade.

Diante dessas hipóteses, os objetivos do presente estudo foram:

a) Identificar os fatores que afetam a produtividade da cana de açúcar no estado de São Paulo por meio da técnica de Mineração de Dados (MD), empregando-se bancos de dados operacionais das usinas;

b) Desenvolver modelos matemáticos de estimação da produtividade da cana de açúcar, ao nível de agrupamento de talhões, empregando-se técnicas de mineração de dados;

c) Testar os modelos propostos com dados independentes, avaliando-se sua viabilidade para a estimação da produtividade em condições operacionais;

d) Comparar o desempenho dos modelos gerados por meio de técnicas de mineração de dados com o de um modelo agrometeorológico, empregado para os mesmos grupos de talhões.

2 REVISÃO BIBLIOGRÁFICA

2.1 A cana de açúcar no Brasil

A produção de cana de açúcar no Brasil se concentra na Região Centro-Sul do país, principalmente nos estados de São Paulo, Minas Gerais, Goiás, Paraná, Mato Grosso do Sul e Mato Grosso, bem como na Região Nordeste, basicamente nos estados de Alagoas, Pernambuco e Paraíba. De acordo com dados da UNICA (2014), a produção da cana de açúcar na região Centro-Sul do Brasil foi da ordem de 533 milhões de toneladas na safra 2012/13, cerca de dez vezes maior do que o produzido no Nordeste Brasileiro. Nesse contexto, São Paulo é de longe o maior estado produtor, com 330 milhões de toneladas na mesma safra.

Os períodos de safra da cana de açúcar no Brasil também se distinguem entre as regiões. De acordo com Scarpari (2002), enquanto que no Nordeste esse período compreende os meses de abril a setembro, no Centro-Sul a safra ocorre entre maio e dezembro. O plantio da cana de açúcar ocorre em duas modalidades, conhecidas como “cana de ano” e “cana de ano e meio”. No Centro-Sul do Brasil, a cana de ano é plantada entre setembro e novembro, aproveitando-se o início das chuvas; já a cana de ano e meio é plantada entre janeiro e abril, sendo colhida apenas no ano consecutivo, entre maio e novembro. De acordo com Rolim (1993), a grande vantagem do plantio de ano é o retorno mais rápido do capital investido. Porém, essa modalidade exige solos férteis para o rápido crescimento da cultura bem como rapidez nas operações de preparo de solo, pois, logo após o último corte, há a necessidade de se destruir as soqueiras e preparar o solo para o plantio em outubro (COLETI, 1983). Com isso, essa operação acaba coincidindo com as atividades de colheita, exigindo maior concentração de mão-de-obra nesse período (ROLIM, 1993). Já a cana de ano e meio inicia seu desenvolvimento no início do ano, época que apresenta boas condições de umidade do solo e temperatura, o que garante uma rápida brotação. Após esse período, permanece em repouso até agosto, voltando a vegetar com alta intensidade durante a primavera e o verão, para então amadurecer nos meses do novo inverno, completando, assim, o ciclo de 18 meses (COLETI, 1983). A brotação rápida garante um bom desenvolvimento da parte aérea da planta, bem como do sistema radicular, de forma que ela possa suportar o período de inverno. Além disso, como no inverno a infestação por plantas

daninhas tende a diminuir, a cana de ano e meio exerce forte influência sobre elas no reinício das chuvas, dominando-as (ROLIM, 1993).

Há também no Centro-Sul uma terceira modalidade, conhecida como “plantio de inverno”, o qual se estende por uma parte da estação seca, mais precisamente até julho, para atender a demanda de matéria-prima das unidades produtoras (EMBRAPA, 2014).

2.2 Fatores que afetam a produtividade da cana de açúcar

Dentre os diversos fatores que afetam a produtividade da cana, têm-se aqueles atribuídos às condições climáticas, e aqueles associados ao manejo agrícola. Esses fatores são descritos a seguir:

2.2.1 Fatores agrometeorológicos

A produtividade da cana de açúcar é altamente dependente do total de energia solar e temperatura, bem como da disponibilidade de água no solo. Marin et al. (2009) destaca a importância de um período de alta radiação, além de altos níveis de temperatura do ar, bem como de alta disponibilidade de água no solo, para o bom crescimento vegetativo da cana. Por ser uma planta de metabolismo C4, a cana possui a vantagem de apresentar uma alta taxa fotossintética, se comparada às plantas de metabolismo do tipo C3. Porém, vale destacar que essa maior eficiência fotossintética só se pronunciará em sua plenitude sob altas taxas de radiação solar, somadas a uma longa duração do dia, temperaturas elevadas e com boa disponibilidade de água no solo. Nesses casos, o perfilhamento é estimulado; por outro lado, em condições de dias curtos e com baixa irradiância solar, o perfilhamento é menos intenso (MONTEIRO, 2011) e as canas se tornam mais longas e finas (RODRIGUES, 1995).

A temperatura do ar influencia nos processos de crescimento e desenvolvimento vegetal, tendo impacto direto na produtividade das culturas (PEREIRA et al., 2002). De acordo com Marin et al. (2009), a cana de açúcar tem taxas máximas de crescimento quando submetida a temperaturas do ar na faixa de 30°C a 34°C. Tanto para temperaturas abaixo de 16°C a 18°C, quanto para temperaturas acima de 38°C, o crescimento da cultura se estagna, devido ao

estresse térmico (MARIN et al., 2009). Para o crescimento ideal da planta, a temperatura média do ar no período vegetativo deve estar na faixa de 22°C a 30°C (DOORENBOS; KASSAM, 1979). Já na fase de maturação, as temperaturas mais baixas auxiliam no processo de acúmulo de sacarose no colmo. De acordo com Câmara (1993), de forma geral, regiões com temperatura média mensal do mês mais frio abaixo de 21°C são mais favoráveis para o processo de maturação. Já em regiões sem déficit hídrico, a temperatura média do ar deve ser abaixo de 21°C por três meses para que se inicie o processo de maturação (GLOVER, 1971 apud CARDOZO; SENTELHAS, 2013).

Por fim, a água desempenha papel vital nas produtividades agrícolas. A necessidade hídrica de uma cultura pode ser definida como o montante de água necessário para atender a demanda evapotranspirativa imposta pelas condições meteorológicas. Para se compensar as dificuldades de se obter medições de campo precisas, métodos de estimação da necessidade hídrica das culturas são usados, empregando-se o coeficiente de cultura (Kc) (ALLEN et al., 1998). De acordo com Marin et al. (2009), diversos autores vêm tentando determinar o consumo de água da cultura da cana de açúcar, porém os resultados diferem significativamente de acordo com o clima, a variedade e o ambiente de produção. De acordo com Blackburn (1984), o suprimento adequado de água para cana é de aproximadamente 1200 mm ano^{-1}, dependendo da região. Conforme Holden e McGuire (1998), em condições adequadas de temperatura e radiação solar, a cana de açúcar cresce em proporção direta ao montante de água consumida, ou seja, para cada 100 mm de água utilizada pela cultura, são produzidas aproximadamente 10 toneladas de colmos por hectare.

Baseando-se na premissa de que o crescimento da cultura e sua produtividade são proporcionais à água transpirada, o balanço hídrico é fundamental para definir o déficit hídrico, variável diretamente relacionada à quebra de produtividade da cultura da cana de açúcar (MONTEIRO; SENTELHAS, 2014). Monteiro (2011), empregando dados provenientes de um modelo de produtividade devidamente calibrado para o estado de São Paulo, determinou a relação entre a redução de produtividade e o déficit hídrico anual das regiões canavieiras do Estado de São Paulo e concluiu que para cada 100 mm de déficit hídrico há uma quebra de produtividade da cultura da cana de açúcar de 11,2 t ha^{-1}.

2.2.2 Fatores de manejo

Dentre os fatores de manejo que influenciam a produtividade da cana de açúcar, destaca-se o manejo varietal. Nas últimas décadas, grandes investimentos foram feitos no desenvolvimento de variedades mais produtivas, mais adaptadas aos diferentes ambientes de produção, assim como mais resistentes às doenças e pragas. A correta alocação das variedades existentes nos diferentes ambientes de produção constitui-se em um importante fator para se maximizar o potencial genético das mesmas e se obter ganhos de produtividade. Além disso, as doenças mais importantes da cultura da cana de açúcar são controladas pelo uso de variedades resistentes (ORLANDO FILHO et al., 1994).

Também a adubação desempenha um papel importante na produtividade da cana de açúcar. De acordo com Orlando Filho et al. (1994), apesar de os produtos finais, açúcar e etanol, conterem apenas carbono, hidrogênio e oxigênio, os nutrientes, dentre eles, os macronutrientes como nitrogênio (N), fósforo (P), potássio (K), cálcio (Ca), magnésio (Mg) e enxofre (S), bem como os micronutrientes boro (B), cloro (Cl), cobre (Cu), ferro (Fe), manganês (Mn), molibdênio (Mo) e zinco (Zn) são importantes para a produção da matéria-prima. Os referidos autores também ressaltam os ganhos de produtividade das canas planta e socas com o uso de gesso agrícola, em complemento à aplicação de calcário, este último com papel mais nutricional, em função do cálcio, do que o de correção da acidez do solo, visto que a cana de açúcar apresenta certa tolerância a tal condição. Conforme Dias e Rosseto (2006), a aplicação de fertilizantes na cultura da cana de açúcar é feita no pré-plantio, com a fosfatagem e aplicação de corretivos, no plantio, durante a operação de sulcação, bem como nos tratos culturais das soqueiras. A adubação pode ser mineral e/ou orgânica, sendo que neste último caso os resíduos agroindustriais, torta de filtro e vinhaça, são muito utilizados. De acordo com Orlando Filho et al. (1983), a torta de filtro é um resíduo da filtração de lodo durante o processo da fabricação de açúcar, sendo um material rico em matéria orgânica, cálcio, fósforo e nitrogênio, com produção média de 30 kg t^{-1} de cana moída. Esta pode ser aplicada antes do plantio, em área total, no sulco de plantio e nas entrelinhas das soqueiras, contribuindo de forma significativa para a elevação da produtividade (ORLANDO FILHO et al., 1994).

A vinhaça, resíduo da fabricação do álcool, também é rica em matéria orgânica e tem alto valor fertilizante, destacando-se os elementos minerais potássio,

nitrogênio e fósforo, com produção média de 13 litros de vinhaça para cada litro de álcool (ORLANDO FILHO et al., 1983; BARROS et al., 2010; LAIME et al., 2010). Sua disposição no solo pode ser na forma de fertirrigação, seja através de caminhões tanque ou sistemas de irrigação. Os benefícios da aplicação da vinhaça podem ser resumidos em aumento da produtividade e da longevidade dos canaviais, substituição da adubação mineral das soqueiras, aumento da fertilidade dos solos e melhorias nas propriedades físicas e da retenção de água dos mesmos (LEME, 1993; BARROS et al., 2010).

As produtividades da cana de açúcar também estão relacionadas com o controle adequado de pragas e doenças. As principais pragas da cana de açúcar são a broca do colmo (*Diatraea* spp.), os cupins, o besouro (*Mygdolus fryanus*) e as cigarrinhas da folha (*Mahanarva posticata)* e da raíz (*Mahanarva fimbriolata)*. Os principais métodos de controle são o uso de inseticidas no sulco de plantio, bem como o uso de técnicas de controle biológico, como a aplicação do fungo *Metarhizium anisopliae* no controle das cigarrinhas e a liberação do parasitoide de larvas *Cotesia flavipes* (CENTRO DE TECNOLOGIA CANAVIEIRA – CTC, 2013; PROCÓPIO et al., 2003; ORLANDO FILHO et al., 1994).

Segundo Bailey (2004), as doenças da cana de açúcar podem ser divididas de acordo com o agente causal, em fungos, como por exemplo, o carvão e a ferrugem da cana, bactérias, como a escaldadura das folhas e o raquitismo das soqueiras, e vírus, sendo o mais importante o Mosaico. As principais medidas de controle são o manejo varietal, o uso de viveiros sadios, o tratamento térmico de mudas de cana, a desinfecção de ferramentas de corte e máquinas de plantio e colheita, e a prática do *roguing*, que consiste na retirada de plantas sintomáticas do canavial (CTC, 2013; MACCHERONI; MATSUOKA, 2006).

Além dos fatores listados acima, a produtividade da cana de açúcar também é influenciada pela longevidade dos canaviais, em termos de número de cortes. Os fatores que influenciam a longevidade são a cultivar escolhida, a fertilidade do solo, as condições climáticas, práticas culturais, controle de pragas e doenças e método de colheita (TOWNSEND, 2000; SEGATO et al., 2006 apud VIANNA, 2014). Por fim, Câmara (1993) cita o trânsito de máquinas e caminhões com seus rodados sobre as linhas das soqueiras, como fator prejudicial à longevidade, uma vez que prejudicam demasiadamente o sistema radicular das soqueiras, promovendo atraso na emissão de novas raízes, falhas de brotação e redução da população de colmos.

2.3 Métodos de estimação da produtividade da cana de açúcar

A estimação de produtividade da cana de açúcar tem grande importância no planejamento estratégico das empresas sucro-energéticas, o que, em função da enorme gama de variáveis que a influenciam, bem como da interdependência desses fatores, acaba sendo muito complexo e difícil (PICOLI, 2007). Entre as ferramentas empregadas na estimação da produtividade da cultura da cana de açúcar destacam-se:

2.3.1 Levantamentos de campo

Essa é a maneira mais tradicional de estimação de produtividade nas usinas de cana de açúcar. Os técnicos responsáveis pelos canaviais percorrem as lavouras e estimam suas produtividades tomando como base suas observações sobre o crescimento da cultura ao longo da safra, bem como as produtividades registradas nas safras anteriores. Em função da subjetividade, esse procedimento está sujeito a erros frequentes, o que acaba comprometendo a estimação de safra das usinas. Outra desvantagem desse método é que ele demanda muito tempo e é oneroso, dado o grande número de pessoas envolvidas (PICOLI, 2007).

2.3.2 Biometrias de campo

As biometrias de campo consistem em medições de variáveis que estão altamente relacionadas com a produtividade dos canaviais, como: altura de colmo; altura das canas; comprimento dos internódios; peso dos colmos e dos palmitos; número de plantas por m de plantio ou de soqueira; número de folhas verdes por cana, entre outras (SILVA et al., 2012). As medições são efetuadas em pontos amostrais, de forma que representam adequadamente as áreas de produção das unidades produtoras.

2.3.3 Imagens de satélite

As ferramentas de sensoriamento remoto vêm sendo aplicadas com resultados satisfatórios para o monitoramento do crescimento dos canaviais, a

identificação de variedades cultivadas e também na estimação de produtividade (ABDEL-RAHMAN; AHMED, 2008). A partir de imagens de satélite pode-se gerar conhecimento sobre o comportamento espectral das culturas, por meio do cálculo dos índices de vegetação, como o NDVI e o GVI. Para a cana de açúcar, os índices de vegetação têm boa correlação com as produtividades, como mostram Abdel, Rahman e Ahmed (2008). De acordo com Benvenuti (2005), os dados espectrais do dossel da cana de açúcar, obtidos no período de seu máximo crescimento, têm relação direta com a produtividade da cultura. Em seu estudo, o autor demonstrou que o índice GVI é o dado espectral que melhor explica a produtividade da cana, com R^2 de até 0,83, e conclui que as imagens de satélite podem ser utilizadas como ferramenta complementar na estimativa de safra da cana de açúcar. Almeida et al. (2006), em seu trabalho de estimação de produtividade da cana de açúcar no Brasil, usaram imagens dos sensores ASTER e Landsat ETM+ e conseguiram maior precisão em comparação aos métodos visuais usados na indústria sucro-alcooleira do Brasil, com erros variando entre 2,6% e 5,7%, inferiores aos estimados nos levantamentos de campo (9,1%). Os autores não têm dúvida de que o sensoriamento remoto auxilia em muitos aspectos da cultura da cana, com o desenvolvimento de produtos de maior resolução espacial, temporal e espectral a um custo menor. Porém, há alguns desafios a serem superados, como o aumento da acurácia desta ferramenta, bem como a redução de problemas causados por nuvens.

2.3.4 Modelagem de produtividade agrícola

De acordo com Hoogenboom (2000), os modelos agrícolas integram conhecimentos de várias disciplinas, incluindo agrometeorologia, física dos solos, química dos solos, fisiologia das culturas, melhoramento vegetal e agronomia em um conjunto de equações matemáticas com o objetivo de estimar o crescimento e o desenvolvimento vegetal bem como as produtividades agrícolas. De forma geral, os modelos matemáticos empregados na agricultura podem ser classificados de várias maneiras. Conforme Thornley e France (2007), os modelos podem ser categorizados como sendo determinísticos ou estocásticos, assim como dinâmicos ou estáticos, e, por fim, mecanísticos ou empíricos.

Com respeito aos modelos que incluem variáveis agrometeorológicas, Baier (1979) sugere a seguinte forma de classificação: modelos de simulação de culturas (*crop growth simulation models*); modelos agrometeorológicos propriamente ditos (*crop-weather analysis models*); e modelos empíricos-estatísticos (*empirical-statistical models*). As principais características desses modelos são:

a) Modelos de simulação de culturas: podem ser definidos como uma representação matemática simplificada dos complexos mecanismos físicos, químicos e fisiológicos que governam o crescimento das plantas (BAIER, 1979). Esse tipo de modelagem utiliza equações matemáticas que visam descrever os processos de crescimento e desenvolvimento vegetal (HOOGENBOOM, 2000), sendo que essas equações são baseadas em resultados de experimentos ou no conhecimento existente sobre um processo específico (BAIER, 1979). Conforme Vianna (2014), existem vários modelos de simulação dos componentes de produção da cana de açúcar, como o APSIM-Sugarcane (KEATING et al., 1999), criado na Austrália, e o CSM-DSSAT-CANEGRO (INMAM BAMBER, 1991; SINGELS et al., 2008), de origem sul africana, ambos com potencial para a simulação da cana de açúcar no Brasil, desde que devidamente ajustados às condições de solo, clima e variedades brasileiras. O CSM-DSSAT-CANEGRO pode ser considerado um dos principais modelos de crescimento de cana de açúcar, sendo extensivamente utilizado em pesquisas científicas e aplicadas (SINGELS et al., 2008). Em um estudo, com o objetivo de parametrizar e avaliar o modelo CANEGRO para cultivares brasileiras, Nassif (2010) concluiu que, apesar de apresentar superestimação no acúmulo de massa de matéria fresca durante o ciclo de produção, os valores finais dessa variável foram compatíveis com os dados experimentais. Marin et al. (2011) avaliaram a estimação de diferentes variáveis da cultura da cana de açúcar pelo modelo CANEGRO, como IAF, massa seca da parte aérea e massa seca de colmos, de duas cultivares de cana de açúcar, e obtiveram resultados satisfatórios quando comparados os dados estimados e observados em experimentos de campo, com erro (RMSE) entre 9,6 e 9,8 t ha^{-1} para a variável massa seca de colmos, e entre 8,5 e 9,9 t ha^{-1} para a variável massa seca da parte aérea. Também Vianna (2014) obteve desempenho satisfatório utilizando o CANEGRO, conseguindo estimar a produtividade comercial da cultura com um RMSE da ordem de 15 t ha^{-1}.

b) Modelos agrometeorológicos: fornecem informações que auxiliam na tomada de decisão na agricultura, como planejamento do uso do solo, monitoramento e controle de pragas e doenças, bem como na estimação de safras (LAZINSKI, 1993 apud TAVARES SOUZA; WILHELM, 2009). De acordo com Baier (1979), esse tipo de modelagem utiliza variáveis agrometeorológicas como umidade do solo e evapotranspiração (oriundas de observações ou medições climáticas diárias) como *inputs* primários e as correlacionam com crescimento vegetativo ou produtividade. Nesse sentido, algumas das funções matemáticas, como, por exemplo a resposta ao uso de fertilizantes ou ao déficit hídrico, são pré-programadas e técnicas estatísticas convencionais, como a regressão múltipla, são usadas para validar os pesos de cada coeficiente da equação final. Considerando-se que os modelos empíricos descrevem as relações entre determinadas variáveis, sem entrar no âmbito dos processos envolvidos nessas relações (DOURADO NETO, 1998) e que os modelos empíricos-estatísticos utilizam dados amostrais de produtividade de determinada área e as correlacionam com dados climáticos da mesma área (BAIER, 1979), fica claro que, a um determinado nível de organização, todos os modelos podem ser considerados empíricos (CARDOZO, 2012), o que torna difícil sua classificação mais clara e precisa. Na modelagem de produtividade da cana de açúcar, podem ser destacados os trabalhos de Ometto (1974), que relacionou a evapotranspiração estimada com o rendimento da cana de açúcar, e de Delgado-Rojas (1998), que obteve equações matemáticas de estimação de produtividade da cana de açúcar por meio de regressão linear múltipla usando variáveis agrometeorológicas como precipitação e evapotranspiração. Monteiro (2011) utilizou o modelo da Zona Agroecológica, apresentado por Doorenbos e Kassam (1994), associado à penalização da produtividade pelo déficit hídrico, para calcular a produtividade potencial e a produtividade atingível da cultura da cana de açúcar em 178 localidades do estado de São Paulo, e estimou a eficiência climática (η) por meio do quociente entre essas produtividades. Assim, foi possível a obtenção de classes de ambientes de produção para a cultura da cana de açúcar sob o enfoque climático, que foram espacializadas, gerando mapas úteis para o planejamento das usinas quanto aos manejos varietal e operacional dos canaviais. O mesmo modelo foi empregado por Gouvêa (2008) para estimar a produtividade da cana de açúcar na região de Piracicaba, SP, para diferentes cenários de mudanças climáticas, baseados no quarto relatório do Painel Intergovernamental sobre Mudanças

Climáticas (IPCC). Nesse estudo se avaliou os possíveis impactos das alterações das variáveis meteorológicas sobre a produtividade da cultura.

c) Modelagem empírica/estatística: A análise estatística de informações oriundas das lavouras para o entendimento dos padrões e dos mecanismos que influenciam a variabilidade da produtividade têm trazido resultados promissores, sendo que sua utilidade tem sido incrementada com o desenvolvimento de modernas ferramentas estatísticas bem como com a disponibilização de bancos de dados de produtividades com uma ampla cobertura de ambientes de produção (FERRARO et al., 2009). Porém, tal abordagem com o uso de ferramentas estatísticas “tradicionais” acaba se tornando muitas vezes inviável, em função de problemas na estrutura dos bancos, como interações de ordem mais elevada e não linearidade dos dados. Para contornar esses problemas, métodos não paramétricos de análise estatística vêm sendo utilizados para aumentar o entendimento da variabilidade das produtividades. Entre esses modelos, estão os baseados em mineração de dados e inteligência artificial, os quais viabilizam especialmente as análises de relações em grandes bancos de dados (*Big Data*), envolvendo uma grande gama de variáveis (FERRARO et al., 2009).

2.4 Mineração de Dados

A crescente geração de dados possibilitada pelos avanços tecnológicos na era da informação vem criando o desafio da obtenção de conhecimento a partir de grandes quantidades de dados. De acordo com Witten et al. (2011), o montante de dados armazenados nos bancos de dados ao redor do mundo se duplica a cada 20 meses, aumentando assim a lacuna entre a geração e o entendimento dessas informações. Nesse sentido, a Mineração de Dados surge como ferramenta capaz de elucidar os padrões implícitos nessas grandes massas de dados. Mineração de Dados, também conhecida como “Descoberta de Conhecimento a partir de Dados” (“*Knowledge Discovery from Data*, KDD) pode ser definida como a extração de padrões que representam o conhecimento armazenado em grandes bancos de dados (HAN et al., 2011). As informações e os conhecimentos adquiridos podem ser utilizados em diversas áreas, como, por exemplo, marketing, medicina, finanças e agropecuária (GALVÃO; MARIN, 2009). O processo da Mineração de Dados integra

técnicas oriundas de diferentes áreas do conhecimento, como tecnologia de bancos de dados, estatística, computação, processamento e visualização de dados, entre outros (HAN et al., 2011). De acordo com Camilo e Silva (2009), existem diversos processos que descrevem as fases e atividades da Mineração de Dados. Um dos mais utilizados é o modelo CRISP-DM (*Cross Industry Standard Process for Data Mining*), que consiste das seguintes fases (IBM SOFTWARE BUSINESS ANALYTICS, 2010; CAMILO; SILVA, 2009):

a) Fase de compreensão do contexto (*business understanding*): compreende o entendimento dos objetivos e necessidades do projeto que deverão ser atendidos por meio da mineração de dados, bem como um planejamento inicial do projeto;

b) Fase de entendimento dos dados (*data understanding*): uma vez que os dados podem vir de diversas fontes e estarem disponíveis em diferentes formatos, essa fase possibilita a familiarização com os dados e suas estruturas, bem como o entendimento de suas especificidades, com o objetivo de identificar possíveis problemas de qualidade dos dados;

c) Fase de preparação dos dados (*data preparing*): esta fase compreende as atividades relacionadas à construção do conjunto final de dados, que será utilizado para analise por meio de técnicas de mineração de dados, como por exemplo, alterações de estruturas para adequação aos softwares de mineração de dados, transformações de variáveis, inclusão de novas variáveis, mudanças de escala, eliminação de inconsistências e de dados faltantes;

d) Fase de modelagem (*modeling*): aqui são aplicados os diferentes algoritmos de mineração de dados. Tipicamente, para cada tipo de problema há diferentes soluções possíveis, sendo que algumas técnicas têm exigências específicas com relação ao formato dos dados;

e) Fase de avaliação (*evaluation*): uma vez desenvolvidos os modelos, é importante revisar o processo e verificar se os objetivos foram alcançados, de forma que, ao final desta etapa, se tome decisões sobre o uso dos resultados;

f) Fase de Implantação (*deployment*): um projeto de mineração de dados não termina com a geração de modelos, mas deve também incorporar a aplicação e distribuição do conhecimento gerado durante o processo. Portanto, nessa fase os resultados devem ser organizados para sua apresentação ao usuário final, de forma que este possa fazer uso prático do conhecimento adquirido.

Como se pode observar na Figura 1, as seis fases do processo de Mineração de Dados são organizadas de maneira cíclica, sendo que o fluxo do processo não é unidirecional, ou seja, pode-se ir e voltar entre as fases do processo.

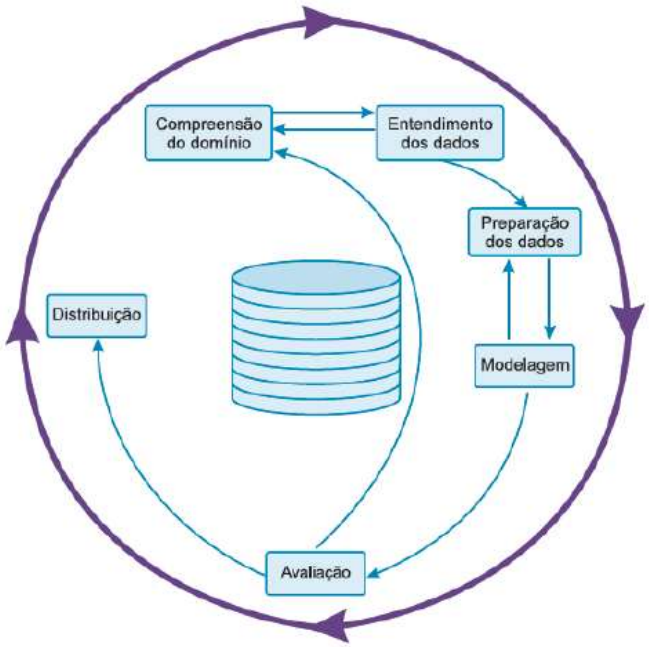

Figura 1 - Fases da metodologia da mineração de dados, denominada de CRISP DM (CHAPMAN et al., 2000)

De acordo com Han et al. (2011), as tarefas de Mineração de Dados podem ser divididas em preditivas, que consistem em se fazer predições a partir de inferências sobre os dados em questão, ou descritivas, com o foco na caracterização das propriedades de forma geral dos dados. As descrições dessas tarefas são apresentadas a seguir:

a) Tarefas preditivas

Classificação: pode ser definida como um processo para encontrar modelos que descrevem e distinguem classes. Esses modelos são gerados com base na

análise de uma base de *dados de treinamento* (em que os dados são conhecidos). Os modelos gerados são utilizados para predizer a categoria dos objetos, cujas classes não são conhecidas, podendo ser representadas de diferentes formas, como por regras, árvores de decisão, fórmulas matemáticas ou redes neurais.

Regressão: a análise de regressão é uma metodologia estatística usada principalmente para predições numéricas. Enquanto que a classificação estima fatores categóricos (discretos), os modelos de regressão estimam funções de valores contínuos.

b) Tarefas descritivas

Associação: Essa tarefa consiste em identificar o relacionamento entre atributos, sendo muito conhecido em função dos bons resultados que se obtém (CAMILO; SILVA, 2009). Dentre as diversas aplicações dessa técnica, se destacam as análises de informações médicas ou de perfil de compras de clientes (BARROS, 2013).

Clusterização ou Agrupamento: Ao contrário das técnicas de classificação e regressão, que analisam conjuntos de dados categorizados, as análises de cluster são destinadas para dados sem o conhecimento prévio de seus agrupamentos em classes. A clusterização pode ser utilizada para gerar um agrupamento de dados formando assim diferentes classes. Os objetos são agrupados baseados no princípio da maximização das similaridades dentro de uma classe (*intraclass similarity*) ou da minimização das similaridades entre as classes (*interclass similarity*). Ou seja, os clusters são gerados de forma que os objetos dentro de um cluster tenham alta similaridade em comparação aos demais, ao mesmo tempo em que são diferentes dos objetos de outros clusters. Sendo assim, cada cluster pode ser visto como uma classe de objetos, da qual podem ser derivadas regras (HAN et al., 2011).

Para a realização de cada tarefa, diferentes técnicas podem ser utilizadas, sendo que não há aquela ideal, com cada uma delas apresentando vantagens e desvantagens. A análise do problema em questão definirá a escolha de determinada técnica, sendo que se pode aplicar mais de um método para solucionar um mesmo problema (VIEIRA, 2014). Dentre as diversas técnicas disponíveis para realizar a tarefa de predição, pode-se destacar as denominadas de *Boosting*, *Random forest*,

ambos pertencentes ao grupo das Árvores de Decisões, e o *Support vector machine*. A seguir é apresentada uma breve descrição dessas técnicas.

2.4.1 Grupo das Árvores de decisão

Esse grupo engloba métodos usualmente simples e de fácil interpretação, podendo ser aplicados tanto para tarefas preditivas, quanto descritivas. Árvores de decisão são estruturas bastante utilizadas em mineração de dados, sendo chamadas de árvores de classificação quando a variável dependente é categórica, e árvores de regressão quando a mesma é contínua (SPINDEL, 2008). As árvores de decisão captam relações não lineares do sistema estudado, quando o uso de abordagens lineares como a regressão fica restrito (JAMES et al., 2013). Para tanto, é utilizada o método de *particionamento recursivo*, que consiste na divisão do conjunto de dados em sub-grupos ou regiões, que são tratados por modelos específicos (SPINDEL, 2008). A representação gráfica é efetuada por nós e ramos, como uma árvore, porém em sentido inverso (WITTEN et. al., 2011), conforme a Figura 2. Seu ciclo de construção começa com um único nó de decisão, situado mais acima na árvore, chamado de nó raiz. Nos nós internos, ou nós de intermediários, são realizados testes sobre os valores das variáveis independentes, gerando ramos, sendo que cada ramo representa uma saída do teste, particionando os sub-grupos (nós intermediários) em mais sub-grupos. Essa ramificação é executada enquanto a divisão de nós gere um ganho considerado suficiente. Assim, o dado entrante é submetido a sucessivas avaliações e ao fim, um ou mais modelos localizados nos nós terminais, ou folhas, são combinados para gerar um resultado final (SPINDEL, 2008; TACONELI, 2008).

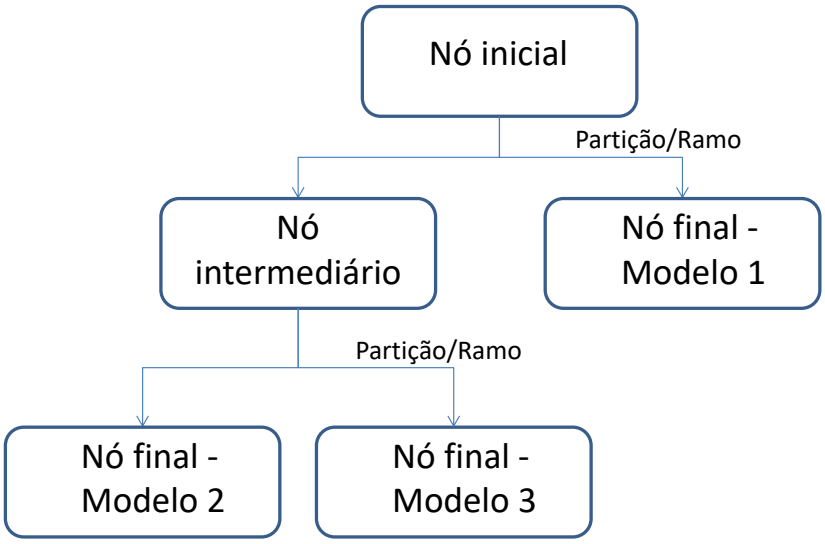

Figura 2 – Ilustração de uma árvore de decisão
(Adaptado de TACONELI, 2008)

Em árvores de regressão e classificação, deve-se buscar uma árvore de tamanho (número de nós finais) reduzido, com baixa heterogeneidade em seus nós finais, evitando-se assim uma alta complexidade da árvore, o que poderia comprometer sua capacidade preditiva (TACONELI, 2008). Um dos meios para se reduzir o número excessivo de nós é um procedimento denominado de poda, que deve ser aplicado após a construção da árvore (JAMES et al., 2013). Esse procedimento consiste em desfazer, uma a uma, aquelas partições que menos contribuem para a explicação da variável resposta (TACONELI, 2008). Apesar de apresentarem grandes vantagens, como a facilidade de explicação bem como sua representação gráfica de simples interpretação, mesmo por não especialistas, as árvores de regressão e classificação não são competitivas com outras técnicas, como, por exemplo, aprendizado de máquina. O poder preditivo das árvores de decisão pode ser melhorado consideravelmente com as técnicas do *Random Forest* e do *Boosting,* nas quais são geradas grandes quantidades de árvores múltiplas, que em seguida são recombinadas para se obter uma predição de consenso (JAMES et al., 2013). Ambas as técnicas são descritas a seguir:

a) *Random Forest*

Essa técnica, desenvolvida por Breiman (2001), utiliza uma variação do *Bagging*, ou *Bootstrap Aggregation*, que consiste em um método de combinação de

várias árvores de decisão criadas com base em amostras aleatórias do conjunto de dados de treinamento.

Considerando-se o conjunto de treino, $\boldsymbol{Z} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, a ideia do *Bagging* é criar um conjunto de B amostras *bootstrap* da amostra original (HAN et al., 2011). A amostragem *bootstrap* é do tipo com reposição, ou seja, cada novo conjunto poderá conter alguns registros incluídos mais de uma vez e outros não inclusos nenhuma vez. Para cada amostra *bootstrap* $\boldsymbol{Z^b}, \mathrm{b} = 1,2, \ldots, \mathrm{B}$, uma árvore de decisão é construída, sem que a mesma seja podada. Seja $\mathrm{g^b(x)}$ a função de predição obtida segundo a *b*-ésima árvore. A função de predição dada pelo *Bagging* é (HASTIE, 2009):

$$g_{bg}(\boldsymbol{x}) = \frac{1}{\mathrm{B}}\sum_{\mathrm{b}=1}^{\mathrm{B}} \mathrm{g^b(x)}$$

no caso de regressão, e, para classificação:

$$\mathrm{g_{bg}(x) = moda_b g^b(x)}.$$

Cada árvore de decisão apresenta alta variância e baixo viés. O propósito desta técnica é que, combinando várias funções de predição, obtenha-se um estimador com variância menor. O *Bagging* geralmente apresenta uma maior acurácia para predição quando comparado com os resultados obtidos por meio de uma única árvore de decisão. Entretanto, o modelo resultante pode ser de difícil interpretação e, desta forma, o *Bagging* melhora a acurácia da predição ao custo da interpretabilidade.

O *Random Forest* admite a mesma abordagem do *Bagging*. Para cada amostra *bootstrap* $\boldsymbol{Z^b}, \mathrm{b} = 1,2, \ldots, \mathrm{B}$, uma árvore de decisão é construída, no entanto, cada vez que uma partição da árvore é realizada, uma amostra aleatória de $\mathrm{m} < \mathrm{p}$ covariáveis é sorteada para selecionar a covariável, dentre os m possíveis candidatos, que será responsável por particionar os nós da árvore (HASTIE, 2009). O valor de m pode ser obtido por meio de validação cruzada. Normalmente, utiliza-se $m \approx \sqrt{p}$, isto é, o número de covariáveis consideradas em cada partição é aproximadamente igual à raiz quadrada do número total de covariáveis.

Seja $\mathrm{g^b(x)}$ a função de predição obtida segundo a *b*-ésima árvore utilizando-se uma amostra aleatória de $\mathrm{m} < \mathrm{p}$ covariáveis, a função de predição dada pelo *Random Forest* é (HASTIE, 2009):

$$g_{RF}(\boldsymbol{x}) = \frac{1}{\mathrm{B}}\sum_{\mathrm{b}=1}^{\mathrm{B}} \mathrm{g^b(x)}$$

A restrição de se utilizar no máximo $m < p$ covariáveis tem por objetivo diminuir a correlação entre os diferentes $g^{b'}s$, caso todas as variáveis fossem utilizadas (como é feito no *Bagging*). Caso contrário, os $g^{b'}s$ seriam muito próximos uns dos outros, de modo que apresentariam alta covariância e, portanto, a variância do estimador combinado não seria tão menor. Desta forma, aumenta-se o viés em troca de uma diminuição da variância. Assim como no *Bagging*, a técnica de *Random Forest* constrói uma árvore para cada amostra *bootstrap* gerada. Com isso, o *Random Forest* pode ser visto como um conjunto de várias árvores de decisão, formando assim uma "floresta". De acordo com Caruana et al. (2008), a técnica de *Random Forest* é mais precisa que outras abordagens, como redes neurais artificiais e máquinas de vetores. Além disso, esta técnica é capaz de evitar sobre ajuste (*overfitting*) e é pouco sensível a ruídos (BREIMAN, 2001).

b) Boosting

O *Boosting* é baseado no conceito de obtenção de algoritmos de classificação fortes (*strong learner*) por meio da transformação de algoritmos ruins (*weak learner*) que, de acordo com Shapire (1990), podem ser descritos como tendo desempenho um pouco acima que o acaso. Como o *Random Forest*, aqui também ocorre a atribuição de pesos às árvores individuais, ou classificadores, para se gerar um consenso a partir da recombinação das mesmas (JAMES et al., 2013). Porém, no *Boosting*, o peso de um classificador será maior, quanto pior for o desempenho de seu antecessor, previamente determinado ao longo do processo de treinamento. O método dessa abordagem funciona aplicando-se sequencialmente o(s) classificador(es) a versões reponderadas do conjunto de treinamento, dando maior peso aos registros classificados erroneamente no passo anterior, e menos peso aos classificados corretamente (HAN, 2011). Dessa forma, o *Boosting* gera uma sequência de árvores que usam a informação contida no treinamento de suas antecessoras, e com isso "aprendem" com o erro (ou pior desempenho) de suas antecessoras (JAMES et al., 2013).

Inicialmente, o método de *Boosting* foi desenvolvido para solucionar problemas envolvendo a classificação de dados. Desta forma, a definição do método, descrita nos próximos parágrafos, será fundamentada em termos da tarefa de classificação.

O algoritmo mais popular de *Boosting* foi desenvolvido por Freund e Schapire (1997) e denominado por “Ada.Boost.M1”. Considerando-se o problema em que a variável aleatória de resposta Y é dicotômica, assumindo apenas os valores -1 e 1, sendo $\boldsymbol{X}$ um vetor contendo as variáveis explicativas, o conjunto de treino, $\boldsymbol{Z} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, e $\mathrm{g(x)}$ uma função responsável por predizer um de dois valores possíveis $\{-1,1\}$ (HASTIE, 2009), o erro do conjunto de treino será dado por:

$$res = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq g(x_i)),$$

em que $I[]$ é a função preditora.

A ideia do algoritmo Ada.Boost.M1 é construir uma função de predição $\mathrm{g(x)}$ incrementalmente. Este algoritmo pode ser definido pelos seguintes passos (HASTIE, 2009):

1) Inicializar o peso das observações $w_i = {}^{1}/_{N}, i = 1, 2, \dots, N$;
2) Para $m = 1, \dots, M$:
   a. Ajustar uma árvore com d folhas para o conjunto de treinamento. Seja $\mathrm{g}_m(\mathrm{x})$ sua respectiva função de predição utilizando o peso w_i;
   b. Calcular:

$$res_m = \frac{\sum_{i=1}^{N} w_i I\left(y_i \neq \mathrm{g}_m(x_i)\right)}{\sum_{i=1}^{N} w_i};$$

   c. Calcular:

$$\alpha_m = \log\left(\frac{1 - res_m}{res_m}\right);$$

   d. Fixar $w_i \leftarrow wi.\exp\{\alpha_m.I\left(y_i \neq g_m(x_i)\right)\}, i = 1, \dots, N.$
3) Saída (onde $sing$ é o sinal):

$$g_{boost} = sing\left[\sum_{m=1}^{M} \alpha_m \mathrm{g}_m(x)\right].$$

Como se pode observar no algoritmo Ada.Boost.M1, inicialmente todos os pesos são fixados como $w_i = 1/N$. O primeiro passo consiste simplesmente em calcular a função de predição da maneira usual, no entanto, nas sucessivas iterações, $m = 2, \dots, M$, o peso das observações é individualmente modificado e a função de predição $\mathrm{g}(x)$ é reaplicada para estas novas observações. No passo m, as observações que foram classificadas erroneamente pela função $\mathrm{g}_{m-1}(x)$ induzem ao aumento dos pesos, enquanto que os pesos diminuem para aquelas observações que foram corretamente classificadas (HASTIE, 2009).

Muitas variações do AdaBoost podem ser encontradas na literatura, no entanto, a mais popular é a Ada.Boost.M1 desenvolvida por Freund e Schapire (1997). Apesar do algoritmo de *Boosting* ter sido desenvolvido inicialmente para a solução de tarefas de classificação, muitas variações de seu algoritmo foram desenvolvidas. Freund e Schapire (1997) criaram o AdaBoost.M2, para casos envolvendo mais de duas classes, e o AdaBoost.R, para problemas de regressão.

Assim como no *Random Forest*, essa técnica gera uma lista das variáveis mais importantes no desenvolvimento do conjunto de classificadores, que são obtidas pela importância acumulada da variável nas divisões de nós de cada árvore construída (JAMES et al., 2013).

2.4.2 Support Vector Machine (SVM)

A abordagem SVM representa um procedimento de aprendizado não-paramétrico para classificação e regressão de alto desempenho (LIMA, 2004). Esse grupo engloba funções ou classificadores que implementam modificações espaciais nos dados, levando-os a um plano onde a classificação se torna mais fácil (Di GIROLAMO NETO, 2013). Suponha que seja necessário escolher, em um plano bidimensional, um classificador que separe os círculos dos triângulos na ilustração abaixo. Na Figura 3a tem-se um classificador com um erro considerável, em função de sua baixa capacidade em distinguir pontos próximos pertencentes a classes diferentes. Sua taxa de acerto é bem inferior à função da Figura 3c, que classifica corretamente a maioria dos exemplos do conjunto de treinamento, mas que pode ser muito específica a um determinado conjunto de dados, podendo causar assim uma situação de *overfitting*, que é quando o modelo se ajusta muito bem ao conjunto em que foi treinado, porém, quando confrontado com novos dados, fica muito suscetível a erros. Por outro lado, a função do classificador em (a) é muito mais simples que em (c). Um classificador mais complexo que (a) e não tão sobre-ajustado quanto (c) seria o classificador (b). Ele tem uma complexidade intermediária e classifica bem grande parte dos exemplos.

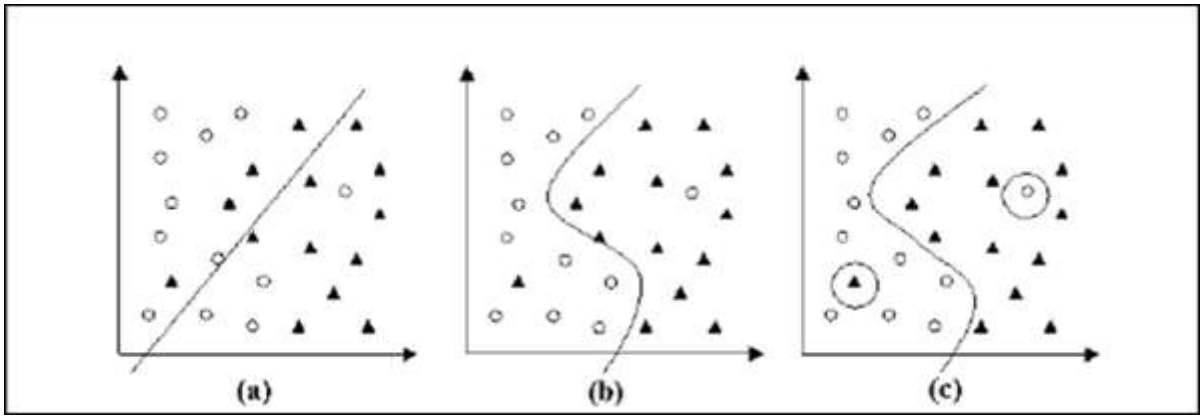

Figura 3 – Exemplo de conjunto de dados de treinamento pertencente a duas classes classificado segundo três diferentes hipóteses (LORENA; CARVALHO, 2003)

As SVM foram desenvolvidas com base na Teoria do Aprendizado Estatístico, que estabelece condições matemáticas que auxiliam na escolha de um classificador específico, a partir de um conjunto de dados de treinamento, que tenha bom desempenho tanto para os conjuntos de treinamento quanto os de teste. Originalmente, as SVM foram desenvolvidas para a solução de problemas que envolviam a classificação de dados em duas categorias, tanto para conjuntos linearmente separáveis, quanto para os casos não lineares, mas foram expandidas também para aplicações que envolviam mais que duas classes, por meio de técnicas específicas para gerar classificadores multiclasses a partir de preditores binários (LORENA; CARVALHO, 2003).

Para se encontrar um hiperplano separador que possa ser considerado ótimo, entre os diferentes hiperplanos capazes de classificar um conjunto de dados, é importante que o mesmo tenha não somente a habilidade de discriminação dos dados, como também tenha sua margem maximizada (MUCHERINO et al., 2009). A margem de um classificador pode ser definida como a menor distância entre as observações de um conjunto de treinamento e o hiperplano utilizado na categorização desses dados (LORENA; CARVALHO, 2003). Na Figura 4, a margem aparece como a distância entre as linhas contínuas e o hiperplano separador (linha pontilhada). Quanto maior a margem do classificador, maior será sua capacidade de generalização, ou seja, de classificar corretamente também as observações do conjunto de teste (MUCHERINO et al., 2009). Apesar de o conceito de hiperplanos de separação com margem máxima ser uma abordagem muitas vezes bem sucedida

no espaço bidimensional, ela pode levar ao *overfitting* se o número de dimensões for alto (JAMES et al., 2013).

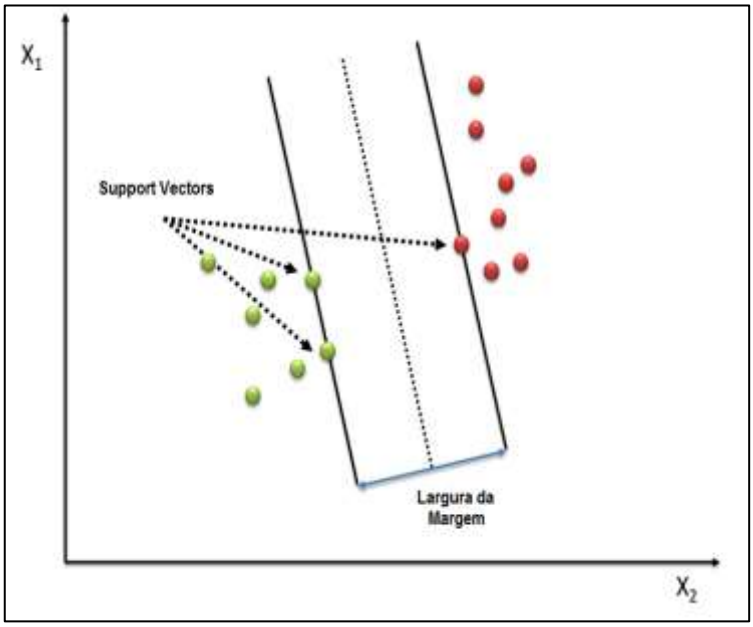

Figura 4 – Exemplo de observações separadas por hiperplano com respectivas margens (SAYAD, 2010)

Em casos de classificação linear, a separação dos dados ocorre por meio de hiperplanos definidos com base no conjunto de dados de treinamento (MUCHERINO et al., 2009). Considerando-se o problema com foco na utilização de uma SVM para a classificação de duas classes $y_i = +1, -1$, se o conjunto de treinamento é linearmente separável, então existe um par $(\vec{w}, b)$, tal que:

$$(\vec{w}.\vec{x}_i + b) \geq 1, \forall \vec{x}_i, \text{para } y_i = 1,$$

$$(\vec{w}.\vec{x}_i + b) \leq -1, \forall \vec{x}_i, \text{para } y_i = -1,$$

cujas inequações podem ser combinadas para se obter a seguinte função objetivo:

$$y_i(\vec{w}.\vec{x}_i + b) \geq 1, \forall \vec{x}_i.$$

A aprendizagem das SVM consiste em se encontrar w e b que satisfaçam o seguinte problema de optimização (HASTIE, 2009; SAYAD, 2010):

$$min \frac{1}{2} \left|\left|\vec{w}\right|\right|^2.$$

Como a solução que minimiza a função objetivo consiste em um problema convexo, então existe um valor mínimo global. A Figura 5 representa o problema de optimização de uma SVM para classificação linear.

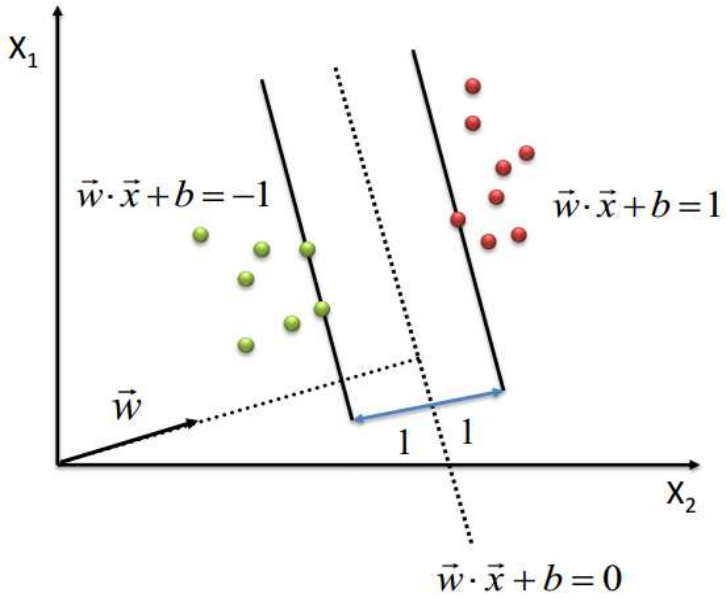

Figura 5 – Exemplo do problema de optimização de uma SVM para a classificação linear (SAYAD, 2010)

Assim, a análise de SVM ideal é aquela capaz de produzir um hiperplano que separe perfeitamente os vetores (observações) em duas classes não sobrepostas. No entanto, uma separação perfeita não é sempre possível ou pode-se obter modelos que não classifiquem corretamente um grande número de observações. Nestas situações, as SVM procuram encontrar um hiperplano que maximize as margens ao mesmo passo que minimizem os erros de classificação.

No caso em que o conjunto de dados não é linearmente separável, a utilização de uma SVM para a classificação de duas classes $y_i = +1, -1$ é possível por meio da adição de uma variável de penalização ξ_i. A Figura 6 representa um exemplo da aplicação de uma SVM no caso em que os dados não são linearmente separáveis.

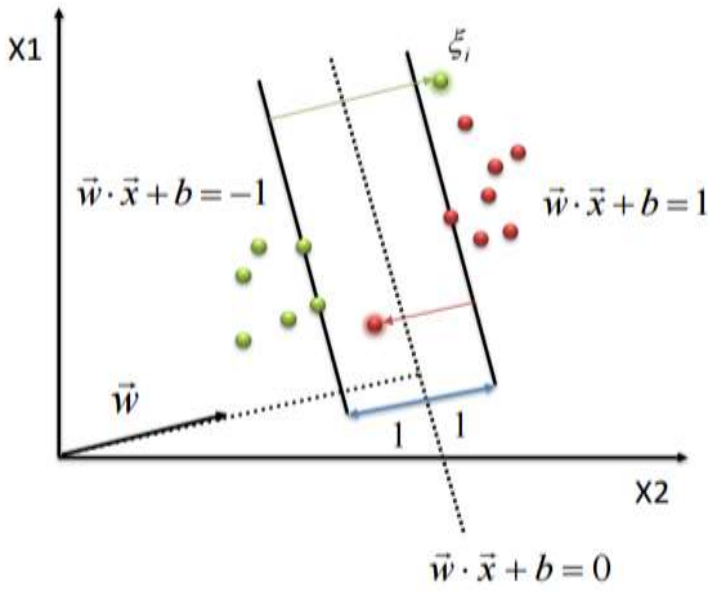

Figura 6 – Exemplo do problema de optimização de uma SVM para o caso em que o conjunto de dados não é linearmente separável (SAYAD, 2010)

No caso em que os dados não são linearmente separáveis, a variável ε_i será responsável por penalizar as observações que venham a ficar fora das margens. Neste caso, a função objetivo pode ser definida como:

$$y_i(\vec{w}.\vec{x}_i + b) \geq 1 - \xi_i, \forall \vec{x}_i \text{ e } \xi_i \geq 0,$$

sendo que neste caso a aprendizagem das SVM consiste em solucionar o seguinte problema de optimização (HASTIE, 2009; SAYAD, 2010):

$$min\frac{1}{2}\left|\left|\vec{w}\right|\right|^2 + c\sum_i \xi_i,$$

em que c é um *tuning parameter,* sendo que quanto maior este parâmetro, maiores as chances das observações caírem do lado “errado” das margens.

Para tarefas de classificação não linear, pode-se usar o artifício de elevação dos dados de treinamento para um espaço hiperdimensional, em que eles se tornam linearmente separáveis, por meio de um hiperplano de margem maximizada. Esse processo de generalização de SVMs para tarefas não lineares é possível graças à utilização de funções Kernel, que permitem o acesso a espaços complexos de forma simplificada. A elevação para um espaço hiperdimensional ocorre por meio de um produto interno chamado de *kernel*, que pode ser do tipo linear ou não linear (Polinomiais), os RBF (Radial-Basis Function) e os Sigmóides (LORENA; CARVALHO, 2003). Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as

redes neurais artificiais (HAYKIN, 2009; BRAGA et al., 2007). De forma geral, o desempenho da abordagem SVM depende da seleção dos parâmetros adequados dentro da função Kernel do classificador. A escolha de parâmetros inadequados pode resultar em decréscimo na acurácia do modelo. Atualmente, não há um método universal para a escolha de parâmetros das funções Kernel (BONESSO, 2013).

Apesar das SVM terem sido desenvolvidas, originalmente, para a solução de problemas envolvendo a classificação binária, outros tipos de SVM foram posteriormente formulados, como SVM para Regressão ou Clusterização. Apesar disso, tais modelos de SVM mantiveram todas as principais características que caracterizam o algoritmo, sendo a maximização das margens a principal delas.

2.5 O uso de Mineração de Dados na agricultura

De acordo com Chinchuluun et al. (2010), o uso de técnicas de mineração de dados é algo recente nas áreas de agrícola e ambiental. Porém, conforme os autores, ambas as áreas podem se beneficiar dessas técnicas.

Na agricultura, uma das aplicações encontra-se na área de solos. De acordo com Kumar e Kannathasan (2011), técnicas de mineração de dados podem ser desenvolvidas especificamente para resolver problemas complexos relacionados à classificação de solos, propiciando assim um melhor entendimento dos mesmos, com a possibilidade de redução da dependência de fertilizantes e o incremento de produtividade. Vibah et al. (2007) desenvolveram um modelo de classificação de solos por meio do uso da técnica de *Random Forest*, obtendo melhor acurácia de classificação se comparado ao algoritmo NBC (*Naive Bayes Classifier*), mundialmente utilizado. Outra área promissora é a de sistemas de alertas de doenças. Meira (2008) desenvolveu um sistema de alerta de ferrugem do cafeeiro, e concluiu que os modelos mais bem avaliados mostraram potencial para utilização na tomada de decisão referente à adoção de medidas de controle a referida doença.

Em cana de açúcar, destacam-se alguns trabalhos como o de Picoli (2007), que desenvolveu um modelo para estimar a produtividade de cana na região de Catanduva, SP, baseado em agregados de Redes Neurais Artificiais (RNA). As variáveis utilizadas no modelo foram: variedade plantada, tipo de solo, aplicação de vinhaça, ano safra, produtividade de cana de açúcar estimada pelos técnicos da Usina, produtividade real do ano safra anterior e o NDVI de uma imagem Landsat-

TM adquirida antes do início do período de colheita. Além disso, foram utilizadas produtividades estimadas a partir de um modelo agronômico-espectral, que por sua vez usa variáveis meteorológicas e agronômicas, como temperatura do ar, radiação solar e índice de área foliar (IAF). Nesse trabalho, a autora conclui que os modelos propostos apresentaram resultados satisfatórios, permitindo assim sua utilização no aprimoramento da estimação da produtividade agrícola da cana de açúcar.

Ferraro et al. (2009) avaliaram bancos de dados de produtividade na região norte da Argentina, do período de 1999 a 2005. Usando a técnica denominada de “árvore de classificação e regressão” (*CART*), os autores classificaram diversas variáveis, tanto do ambiente quanto de manejo, de acordo com sua importância hierárquica na variabilidade das produtividades (TCH), e concluíram que, entre os fatores analisados, a localidade, representada pelo agrupamento em uma das seis fazendas analisadas, foi o que teve maior influência nas produtividades, seguido do número de cortes e da variedade. Nesse caso, as condições climáticas não foram avaliadas de forma específica, estando representadas pelas localidades. Nonato (2010) desenvolveu modelos preditivos para identificar áreas cultivadas com cana de açúcar em imagens do satélite LANDSAT 5 no estado de SP, por meio de técnicas de árvores de decisão. Em seu trabalho, o autor conclui que as técnicas de mineração de dados voltadas para a classificação se mostraram como alternativas de grande potencial na resolução de problemas de identificação e classificação de regiões cultivadas com cana de açúcar em imagens de satélite, tanto pela melhoria na acurácia da classificação, como pela redução da necessidade de informações para a resolução deste tipo problema. Barros (2013) desenvolveu um sistema de recomendação web, baseado em regras de associação para recomendações de conteúdos para a cultura da cana de açúcar, com base em dados extraídos do banco de informações da Agência de Informação da Embrapa. Com esse trabalho, foi possível diminuir a taxa de rejeição dos usuários do site da Agência de Informação Embrapa – Árvore Cana de açúcar, os quais passaram a encontrar informações relevantes associadas as suas visitas ao site, aumentando assim seu tempo de permanência e intensificando o uso e visualização dos conteúdos no site.

3 MATERIAL E MÉTODOS

A descrição dos procedimentos utilizados no presente estudo foi realizada considerando-se as etapas que constituem o processo CRISP DM. Conforme já mencionado na revisão de literatura, essa metodologia encontra ampla aceitação em projetos de mineração de dados (CAMILO; SILVA, 2009).

3.1 Entendimento dos dados empregados

Foram usados dados de produtividade da cana de açúcar (TCH – toneladas de colmo por hectare) das safras 2012/2013, 2013/14 e 2014/2015, das seguintes regiões produtoras no estado de São Paulo: Campinas, Bauru, Araçatuba, Ourinhos e Ribeirão Preto (Tabela 1). Essas informações são oriundas dos bancos de usinas de açúcar e álcool situadas nas regiões mencionadas, disponíveis em nível de talhão.

Tabela 1 – Relação das regiões produtoras de cana de açúcar analisadas neste estudo, com suas respectivas coordenadas geográficas e altitudes médias e número de usinas das quais foram obtidos os dados para análise

Região	Latitude (graus e décimos)	Longitude (graus e décimos)	Altitude (m)	Número de Usinas por região
Araçatuba	-20,87	-50,48	439	5
Bauru	-22,31	-49,06	560	3
Campinas	-22,90	-47,07	678	4
Ourinhos	-22,98	-49,87	488	1
Ribeirão Preto	-21,17	-47,80	568	5

Da mesma forma empregada para os dados de produtividade, também foram levantados junto aos bancos de dados das usinas as seguintes informações referentes ao manejo da cultura da cana de açúcar ao nível de talhão:

a) Variedades utilizadas;
b) Ambientes de produção: relativo ao tipo de solos;
c) Aplicação de insumos: calcário, gesso, fosfato, vinhaça, inseticida;
d) Número de cortes do canavial;
e) Idade do canavial no corte: número de meses entre brotação e colheita;
f) Espaçamento de plantio – diferentes tipos de espaçamento da cultura.

Por fim, foram também utilizados dados meteorológicos, tanto na modelagem por meio de mineração de dados, como para o emprego do modelo agrometeorológico de estimação da produtividade da FAO (PEREIRA et al., 2002; MONTEIRO; SENTELHAS, 2014; MONTEIRO, 2015). Os dados meteorológicos utilizados foram: temperatura média do ar; precipitação; velocidade do vento; e radiação solar, obtidos a partir de estações meteorológicas automáticas situadas nas regiões produtoras em análise (Tabela 1), o que engloba tanto as estações próprias das usinas quanto aquelas da rede do Instituto Nacional de Meteorologia (INMET). Todas as estações seguem os padrões estabelecidos pela OMM (Organização Meteorológica Mundial). Já os dados pluviométricos são oriundos dos registros dos pluviômetros convencionais das unidades produtoras, totalizando 150 pluviômetros. Os dados de todas as variáveis meteorológicas foram obtidos a partir de 2011. Com isso, garantiu-se que fossem considerados os ciclos completos de todas as canas dos bancos de dados das usinas empregados na modelagem. Como exemplo, podem-se citar as canas de 18 meses colhidas na safra 2013/2014, cujos plantios datam do último trimestre de 2011.

3.2 Preparação dos dados

Todos os tipos de dados que fizeram parte da modelagem, ou seja, dados de TCH, de manejo agrícola e meteorológicos, são oriundos de diferentes tabelas, e passaram a constituir um único banco de dados, cujo sistema gerenciador utilizado para armazenamento e consulta foi o MS Access. Todos os processos de busca de dados entre as diferentes tabelas, os quais serão descritos posteriormente, foram programados por meio das linguagens SQL e Visual Basic.

Os dados de produtividade da cana de açúcar, em toneladas de cana por hectare (TCH), e manejo agrícola, conforme mencionado anteriormente, estavam

originalmente disponíveis em nível de talhão. Porém, para se contornar erros decorrentes de produtividades não representativas daqueles talhões, foi executado um agrupamento dessas produtividades em blocos de manejo. Esse problema ocorre porque, em muitos casos, há uma mistura da cana colhida nos talhões, uma vez que é logisticamente viável que a colhedora permaneça em sua linha de colheita, atravessando assim mais talhões, ao invés de colher um talhão após o outro, o que acarreta mais manobras. Essa mistura de canas de vários talhões na mesma carga da colhedora acaba sendo distribuída de forma aleatória aos talhões durante a inserção dos dados de produtividade no sistema, ocasionando uma distorção da produtividade real do talhão em análise. Esse problema é contornado no presente trabalho com o agrupamento dos talhões em blocos de manejo, que podem ser definidos como agrupamentos de talhões com as mesmas características, como variedade plantada e número de cortes do canavial. São os blocos de manejo que fornecerão as produtividades a serem analisadas pela mineração de dados, bem como comparadas com as produtividades geradas pelo modelo agrometeorológico FAO. Além disso, os dados de TCH, já agrupados em blocos de manejo, passaram por um processo de organização e consistência, de modo que aqueles valores discrepantes, muito acima ou abaixo das produtividades esperadas, foram descartados da análise, como tratamento de *outliers*. Foram utilizados os valores de 20 t ha^{-1} e de 160 t ha^{-1} como os limiares mínimo e máximo, respectivamente, para tal consistência dos bancos de dados. A Figura 6 apresenta a distribuição dos dados de produtividade ao nível de blocos de manejo. Quando se considera todas as variáveis que entraram na modelagem, bem como o número total de blocos de manejo analisados (13.274), tem-se um banco de grande dimensão, com 1,46 milhões de dados, justificando assim a utilização de técnicas de mineração de dados para a análise e geração de conhecimento a partir desses.

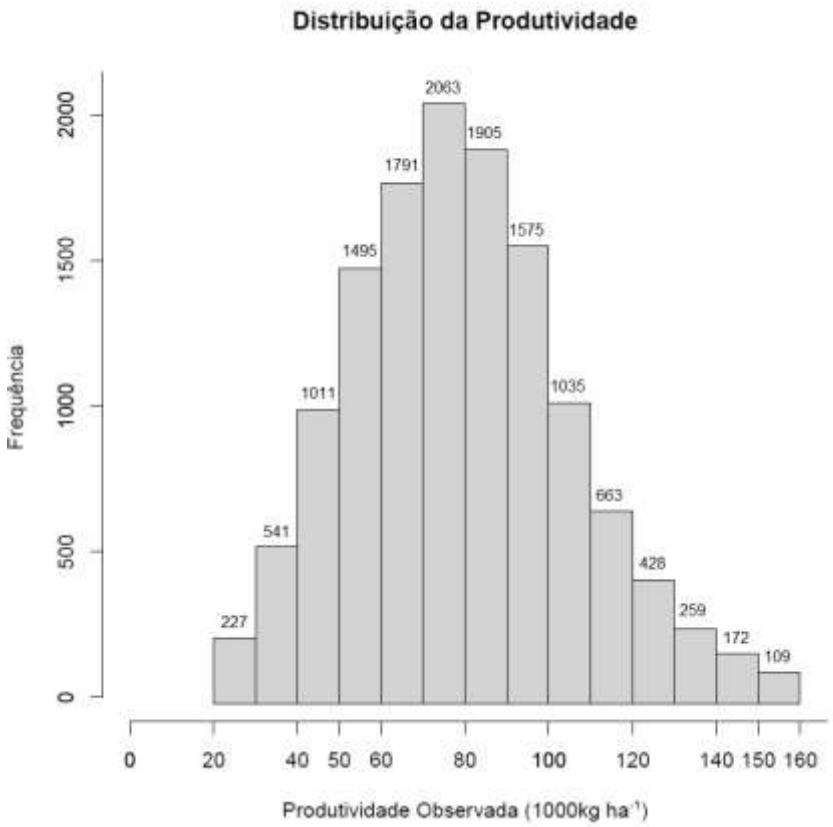

Figura 6 - Distribuição de frequência das produtividades de cana de açúcar observadas no banco de dados empregado no presente estudo

Na Figura 7 pode-se observar o diagrama que descreve os relacionamentos entre as tabelas com seus respectivos campos, que são necessários para o armazenamento das variáveis de entrada para a modelagem por meio de mineração de dados, bem como as variáveis de entrada e saída do modelo agrometeorológico da FAO. O bloco de manejo, sendo a menor unidade de área considerada na modelagem, contém várias informações cadastrais relativas à produtividade, ao manejo e à qualidade da cana. Assim como essas informações, todos os dados referentes às variáveis meteorológicas, que irão compor a base da modelagem, também são disponibilizados ao nível de bloco.

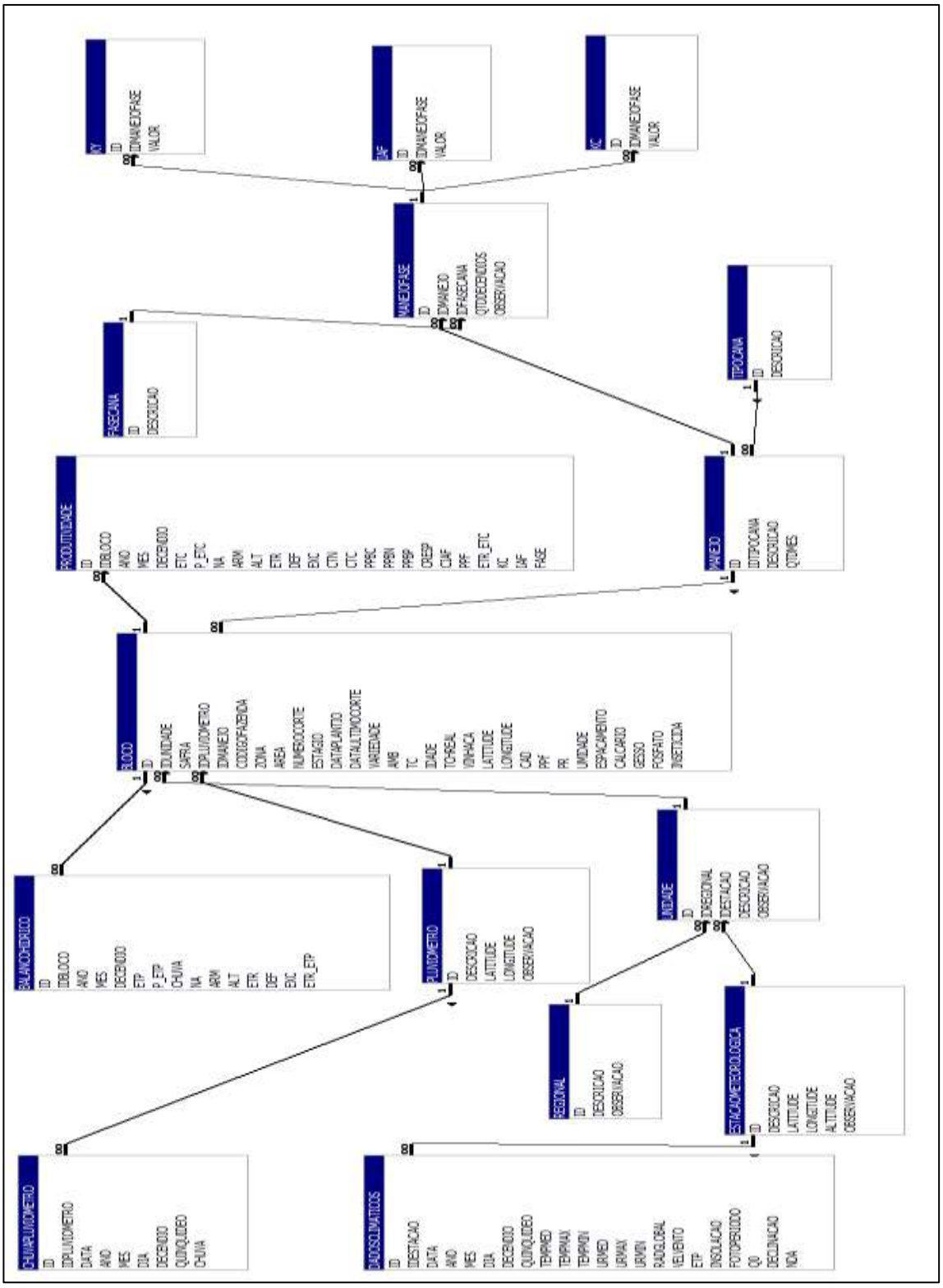

Figura 7 - Diagrama de entidade e relacionamento de todas as variáveis que compõem o banco de dados empregado na modelagem da produtividade da cana de açúcar por meio da técnica de mineração de dados e do modelo FAO

Com relação à determinação das chuvas de cada bloco de manejo, utilizou-se a técnica do Diagrama de Voronoi, a qual relaciona os dados de um pluviômetro com o bloco em análise mais próximo, conforme a Figura 8. O princípio do Diagrama de Voronoi admite que em um plano existem pontos que estão mais próximos de uma fonte geradora do que de outra, sendo possível se determinar as menores distâncias possíveis (MOURA, 2003). Por meio dessa técnica, foi, então, possível calcular a menor distância relativa entre o centro geográfico do bloco e os pluviômetros da região, relacionando assim cada bloco ao pluviômetro mais próximo. Para todos os demais dados meteorológicos, como temperatura e umidade relativa do ar, radiação solar global e velocidade do vento, o relacionamento da estação meteorológica com os blocos de manejo se deu por meio da definição das unidades produtoras nas quais as estações meteorológicas estão instaladas.

3.2.1 Estimação da produtividade com base em modelo agrometeorológico para comparação de seu desempenho em relação aos modelos de mineração de dados

Para a estimação da produtividade potencial da cultura da cana de açúcar, ao longo dos ciclos avaliados, foi utilizado o modelo da Zona Agroecológica da FAO (DOORENBOS; KASSAM, 1994), o qual considera as seguintes variáveis meteorológicas como dados de entrada: temperatura média (T); fotoperíodo (N); e insolação (n). O modelo foi programado em planilhas eletrônicas no ambiente EXCEL® na escala decendial, de acordo com a seguinte equação:

$$\mathrm{PP} = \sum_{i=1}^{m} (\mathrm{PPBp}_i \times \mathrm{c_{iaf}} \times \mathrm{c_{resp}} \times \mathrm{c_{colh}} \times \mathrm{c_{um}})$$

em que: PP é a produtividade potencial final da cultura, em kg ha^{-1}; *i* representa os decêndios do ciclo; PPBp é a produtividade potencial bruta padrão de matéria seca para uma cultura hipotética com IAF = 5, em kg MS ha^{-1} dia^{-1}; c_{iaf} é o coeficiente para correção do índice de área foliar; c_{resp} é o coeficiente de correção para respiração de manutenção da cultura, função da temperatura do ar; c_{colh} é o coeficiente relativo à parte colhida ou índice de colheita (no caso, colmos); c_{um} é o coeficiente que considera a umidade da parte colhida.

Com o intuito de se diferenciar os ciclos assumidos no estudo, principalmente em relação aos ciclos de cana planta e soca, foi feita uma distinção com relação ao índice de área foliar (IAF), de acordo com a duração das fases fenológicas da cultura, conforme apresentado em Monteiro e Sentelhas (2013). Os índices de área foliar (IAF) são apresentados na Tabela 2.

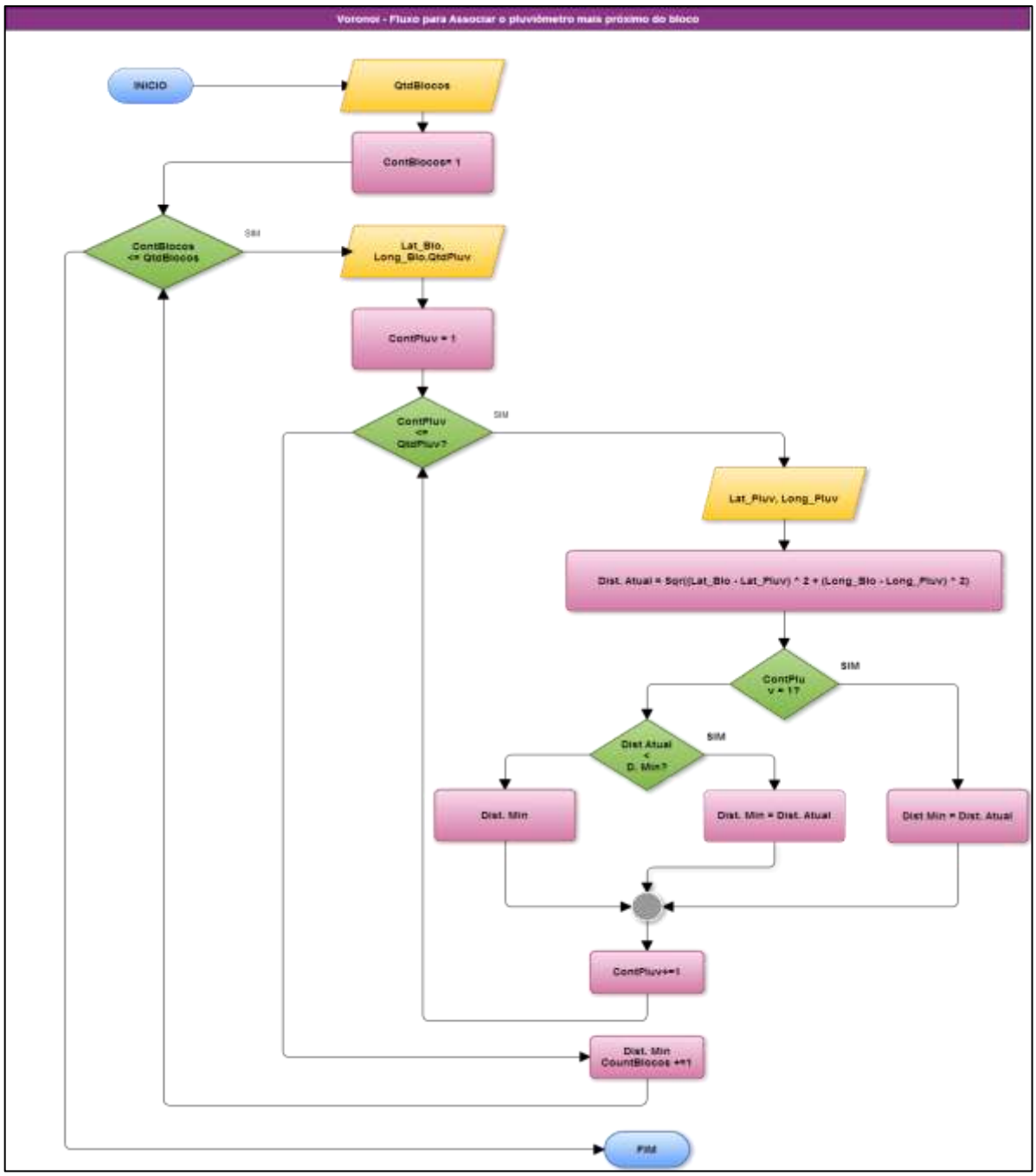

Figura 8 - Fluxo de associação dos pluviômetros mais próximos de cada um dos blocos considerados, de acordo com a técnica de Voronoi (MOURA, 2003)

Tabela 2 – Índice de área foliar (IAF) adotado para cada um dos decêndios do ciclo da cultura da cana de açúcar, considerando-se os diferentes tipos de cana (planta ou soca) e ciclos (12, 15 e 18 meses)

Decendio	Planta						SOCA	
	18		15		12		12	
	Fase	IAF	Fase	IAF	Fase	IAF	Fase	IAF
1	25 % de cobertura	2,50	25 % de cobertura	2,00	25 % de cobertura	2,00	25 % de cobertura	2,00
2	25 % de cobertura	2,50	25 % de cobertura	2,00	25 % de cobertura	2,00	25 % de cobertura	2,00
3	25 % de cobertura	2,50	25 % de cobertura	2,00	25 % de cobertura	2,00	25 % de cobertura	2,00
4	25 % de cobertura	2,50	25-50% de cobertura	2,50	25-50% de cobertura	2,50	25-50% de cobertura	2,50
5	25-50% de cobertura	3,00	25-50% de cobertura	2,50	25-50% de cobertura	2,50	25-50% de cobertura	2,50
6	25-50% de cobertura	3,00	25-50% de cobertura	2,50	25-50% de cobertura	2,50	25-50% de cobertura	2,50
7	25-50% de cobertura	3,00	50-75 % de cobertura	3,50	50-75 % de cobertura	3,00	50-75 % de cobertura	3,00
8	25-50% de cobertura	3,00	50-75 % de cobertura	3,50	50-75 % de cobertura	3,00	50-75 % de cobertura	3,00
9	50-75 % de cobertura	4,50	75-100 % de cobertura	4,00	75-100 % de cobertura	3,50	75-100 % de cobertura	3,50
10	50-75 % de cobertura	4,50	75-100 % de cobertura	4,00	75-100 % de cobertura	3,50	75-100 % de cobertura	3,50
11	50-75 % de cobertura	4,50	75-100 % de cobertura	4,00	75-100 % de cobertura	3,50	75-100 % de cobertura	3,50
12	75-100 % de cobertura	5,00	75-100 % de cobertura	4,00	75-100 % de cobertura	3,50	75-100 % de cobertura	3,50
13	75-100 % de cobertura	5,00	100 % de cobertura	5,00	75-100 % de cobertura	3,50	75-100 % de cobertura	3,50
14	75-100 % de cobertura	5,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
15	75-100 % de cobertura	5,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
16	75-100 % de cobertura	5,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
17	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
18	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
19	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
20	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
21	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
22	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
23	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
24	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
25	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
26	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
27	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
28	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
29	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
30	100 % de cobertura	6,00	100 % de cobertura	5,00	100 % de cobertura	4,50	100 % de cobertura	4,00
31	100 % de cobertura	6,00	100 % de cobertura	5,00	Senesescência	4,00	Senesescência	3,50
32	100 % de cobertura	6,00	100 % de cobertura	5,00	Senesescência	4,00	Senesescência	3,50
33	100 % de cobertura	6,00	100 % de cobertura	5,00	Senesescência	4,00	Senesescência	3,50
34	100 % de cobertura	6,00	100 % de cobertura	5,00	Maturação	3,50	Maturação	3,00
35	100 % de cobertura	6,00	100 % de cobertura	5,00	Maturação	3,50	Maturação	3,00
36	100 % de cobertura	6,00	100 % de cobertura	5,00	Maturação	3,50	Maturação	3,00
37	100 % de cobertura	6,00	100 % de cobertura	5,00				
38	100 % de cobertura	6,00	100 % de cobertura	5,00				
39	100 % de cobertura	6,00	Senesescência	4,00				
40	100 % de cobertura	6,00	Senesescência	4,00				
41	100 % de cobertura	6,00	Senesescência	4,00				
42	100 % de cobertura	6,00	Senesescência	4,00				
43	100 % de cobertura	6,00	Maturação	3,50				
44	100 % de cobertura	6,00	Maturação	3,50				
45	100 % de cobertura	6,00	Maturação	3,50				
46	100 % de cobertura	6,00						
47	Senesescência	5,00						
48	Senesescência	5,00						
49	Senesescência	5,00						
50	Senesescência	5,00						
51	Senesescência	5,00						
52	Maturação	4,50						
53	Maturação	4,50						
54	Maturação	4,50						

A PPBp foi calculada por meio do somatório da produtividade potencial bruta em condições de céu claro (PPBc) e de céu nublado (PPBn) a cada decêndio do ciclo da cultura. PPBc e PPBn (kg MS ha^{-1} dia^{-1}) foram estimados a partir das seguintes equações:

$$\mathrm{PPBc} = (107{,}2 + 8{,}604 \times \mathrm{Q_o}) \times {}^{n}/_{N} \times \mathrm{cTc}$$

$$\mathrm{PPBn} = (31{,}7 + 5{,}234 \times \mathrm{Q_o}) \times \left(1 - {}^{n}/_{N}\right) \times \mathrm{cTn}$$

em que: n é a insolação, em h dia^{-1}; N é o fotoperíodo diário, em h dia^{-1}; Qo é a irradiância solar extraterrestre, em MJ m^{-2} dia^{-1}; cTc e cTn representam os fatores de correção para a temperatura e metabolismo de espécie C_4 nos dias de céu limpo e céu nublado, respectivamente, como apresentado em Pereira et al. (2002):

Se $T \geq 16{,}5$ºC:

$$\mathrm{cTc} = -4{,}16 + 0{,}4325 \times \mathrm{T} - 0{,}00725 \times \mathrm{T}^2$$

$$\mathrm{cTn} = -1{,}064 + 0{,}173 \times \mathrm{T} - 0{,}0029 \times \mathrm{T}^2$$

Se $T < 16{,}5$ºC:

$$\mathrm{cTc} = -9{,}32 + 0{,}865 \times \mathrm{T} - 0{,}0145 \times \mathrm{T}^2$$

$$\mathrm{cTn} = -4{,}16 + 0{,}4325 \times \mathrm{T} - 0{,}00725 \times \mathrm{T}^2$$

em que: T é a temperatura média ao longo do decêndio, em ºC.

Como a área foliar da cultura varia ao longo de seu ciclo, o coeficiente de correção da PPBp relativo ao índice de área foliar (IAF) da cultura (c_{iaf}) foi calculado empregando-se a equação:

$$\mathrm{c_{iaf}} = 0{,}0093 + 0{,}185 \times \mathrm{IAF_{max}} - 0{,}0175 \times \mathrm{IAF_{max}^2} \qquad (\mathrm{IAF_{max}} \geq 5; \mathrm{c_{iaf}} = 0{,}5)$$

em que: IAF_{max} representa o valor máximo de IAF no decêndio considerado.

O coeficiente de correção para a respiração de manutenção de cultura (c_{resp}) é uma função da temperatura média do ar (Tmed), em que duas condições são consideradas:

Para Tmed $\geq$ 20ºC $\rightarrow$ c_{resp} = 0,5

Para Tmed $<$ 20ºC $\rightarrow$ c_{resp} = 0,6.

O coeficiente de colheita (C_{colh}) indica o quanto da planta representa a parte colhida, colmos no caso da cana de açúcar. Neste caso, esses representam, em média, 80% da massa total da planta de cana de açúcar (DOORENBOS; KASSAM, 1994).

Finalmente, considerou-se a correção para o teor de água na parte colhida. A equação relativa ao c_{um} empregada para a adição da umidade dos colmos foi dada por:

$$c_{um} = [1 - 0{,}01 \text{ x } U(\%)]^{-1}$$

em que: U(%) representa a umidade dos colmos, em porcentagem. No presente trabalho foi considerado 80%.

A produtividade atingível foi estimada a partir da relação entre a quebra relativa de produtividade (1 – PA/PP) e o déficit hídrico relativo (1 – ETr/ETc). A partir dessa relação, Doorenbos e Kassam (1994) propuseram um modelo que relaciona tais componentes para estimação da produtividade atingível (PA) de uma determinada cultura. Assim, é possível calcular a PA da cana de açúcar conhecendo-se a produtividade potencial (PP), o coeficiente de resposta ao déficit hídrico (ky), variável em função das fases fenológicas, e o déficit hídrico relativo, proveniente do balanço hídrico da cultura, neste trabalho calculado para cada bloco de manejo de acordo com o método proposto por Thornthwaite e Mather (1955). Os valores de capacidade de água disponível (CAD) utilizados nos balanços hídricos obedeceram, basicamente, ao tipo de solo (PRADO, 2008), disponibilizado nos bancos de dados das usinas para cada talhão. Os valores de ky, a serem empregados no modelo, são os apresentados em Monteiro e Sentelhas (2013) (Tabela 3).

Tabela 3 - Coeficiente de sensibilidade ao déficit hídrico (ky) e coeficiente de cultura (kc) considerados para cada um dos decêndios do ciclo da cultura da cana de açúcar, considerando-se os diferentes tipos de cana (planta ou soca) e ciclos (12, 15 e 18 meses)

Decendio	Planta									SOCA		
	18			15			12			12		
	Fase	kc	ky	Fase	kc	ky	Fase	kc	ky	Fase	kc	ky
1	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75
2	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75
3	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75	25 % de cobertura	0,50	0,75
4	25 % de cobertura	0,50	0,75	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75
5	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75
6	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75	25-50% de cobertura	0,80	0,75
7	25-50% de cobertura	0,80	0,75	50-75 % de cobertura	1,00	0,50	50-75 % de cobertura	1,00	0,50	50-75 % de cobertura	1,00	0,50
8	25-50% de cobertura	0,80	0,75	50-75 % de cobertura	1,00	0,50	50-75 % de cobertura	1,00	0,50	50-75 % de cobertura	1,00	0,50
9	50-75 % de cobertura	1,00	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50
10	50-75 % de cobertura	1,00	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50
11	50-75 % de cobertura	1,00	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50
12	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50
13	75-100 % de cobertura	1,10	0,50	100 % de cobertura	1,20	0,50	75-100 % de cobertura	1,10	0,50	75-100 % de cobertura	1,10	0,50
14	75-100 % de cobertura	1,10	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
15	75-100 % de cobertura	1,10	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
16	75-100 % de cobertura	1,10	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
17	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
18	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
19	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
20	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
21	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
22	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
23	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
24	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
25	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
26	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
27	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
28	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
29	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
30	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50
31	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	Senesescência	0,95	0,50	Senesescência	0,95	0,50
32	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	Senesescência	0,95	0,50	Senesescência	0,95	0,50
33	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	Senesescência	0,95	0,50	Senesescência	0,95	0,50
34	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	Maturação	0,65	0,10	Maturação	0,65	0,10
35	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	Maturação	0,65	0,10	Maturação	0,65	0,10
36	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50	Maturação	0,65	0,10	Maturação	0,65	0,10
37	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50						
38	100 % de cobertura	1,20	0,50	100 % de cobertura	1,20	0,50						
39	100 % de cobertura	1,20	0,50	Senesescência	0,95	0,50						
40	100 % de cobertura	1,20	0,50	Senesescência	0,95	0,50						
41	100 % de cobertura	1,20	0,50	Senesescência	0,95	0,50						
42	100 % de cobertura	1,20	0,50	Senesescência	0,95	0,50						
43	100 % de cobertura	1,20	0,50	Maturação	0,65	0,10						
44	100 % de cobertura	1,20	0,50	Maturação	0,65	0,10						
45	100 % de cobertura	1,20	0,50	Maturação	0,65	0,10						
46	100 % de cobertura	1,20	0,50									
47	Senesescência	0,95	0,50									
48	Senesescência	0,95	0,50									
49	Senesescência	0,95	0,50									
50	Senesescência	0,95	0,50									
51	Senesescência	0,95	0,50									
52	Maturação	0,65	0,10									
53	Maturação	0,65	0,10									
54	Maturação	0,65	0,10									

A equação final empregada para a estimação da produtividade atingível (PA) foi:

$$\mathrm{PA}_i = \prod_{i=1}^{7} \left\{ \mathrm{PA}_{i-1} \, \mathrm{x} \left[1 - \mathrm{ky}_i \, \mathrm{x} \, \left(1 - \frac{\mathrm{ETr}}{\mathrm{ETc}} \right)_i \right] \right\}$$

em que: PA_i é a produtividade atingível, em t ha^{-1}; *i* representa as fases fenológicas durante o ciclo da cultura; PA_{i-1} é a produtividade atingível do estádio fenológico anterior, em t ha^{-1}, sendo que no caso do estádio fenológico *i* = 1, PA_{i-1} = PP; ky é o coeficiente de resposta ao déficit hídrico da cultura em cada fase fenológica *i*; ETr é a evapotranspiração real da cultura, em mm; ETc é a evapotranspiração de cultivo, em mm, e é dada pelo produto entre ETP e kc. A ETP é a evapotranspiração potencial, estimada pelo método de Penman Monteith (ALLEN et al., 1998). Já o kc representa o coeficiente da cultura em cada fase fenológica *i,* conforme Tabela 3. O fluxograma contendo as equações desses modelos e o fluxo de cálculo desses é apresentado na Figura 9.

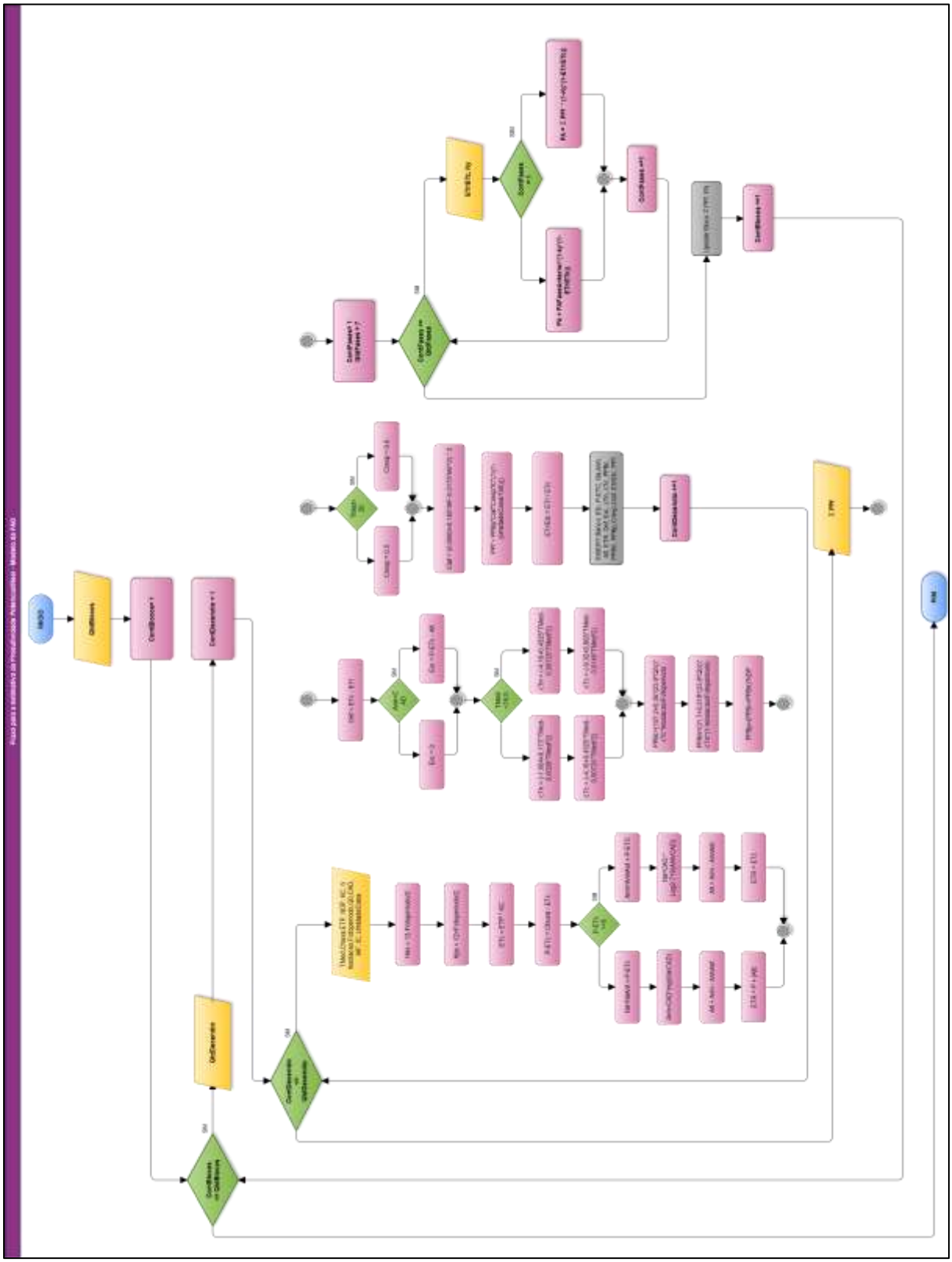

Figura 9 - Fluxo da estimação das produtividades potencial e atingível da cana de açúcar por meio do modelo da FAO (DOORENBOS; KASSAM, 1994)

3.3 Modelagem da produtividade da cana de açúcar por meio da mineração de dados

Todos os modelos foram gerados a partir da execução dos algoritmos no software R (http://www.r-project.org/). Três diferentes técnicas de mineração de dados foram empregadas neste estudo: o *Random Forest* (*RF*); o *Boosting* (*GBM*); e o *Support Vector Machine* (*SVM*). Como interface para a aplicação dessas técnicas, foi utilizado o pacote *Caret* (*Classification And Regression Training*). O *Caret* é um conjunto de funções desenvolvidas para simplificar o processo de criação de modelos preditivos, e contém ferramentas para diversas funcionalidades, como a seleção automática de parâmetros para as diferentes técnicas de mineração de dados, como por exemplo, o número de árvores ideal no *Random Forest,* bem como a caracterização das variáveis preditoras mais importantes dos modelos. Além disso, o pacote proporciona uma visualização mais amigável dos resultados (KUHN, 2008).

Em uma primeira etapa, foi feita uma abordagem mais geral, em que os algoritmos acima foram aplicados na base integral de dados, que contém os dados de todas as regiões e safras analisadas. Em uma segunda etapa, as mesmas técnicas foram utilizadas para a análise dos dados de uma das regiões produtoras integrantes do banco de dados. A região escolhida para essa análise foi a de Bauru, a qual, juntamente com os municípios em seu entorno, é um importante polo produtor de cana de açúcar no estado de São Paulo. Essa abordagem mais específica foi feita com o intuito de se tentar capturar particularidades da região escolhida e que poderiam estar ocultadas pela análise geral dos dados.

Com base nisso, os dados de produtividade da cana de açúcar, bem como os dados meteorológicos, das produtividades estimadas pelo método da FAO e de manejo da cultura das safras de 2011/12 a 2013/14 foram organizados para que fossem submetidos ao processo de treinamento dos algoritmos de mineração de dados (*Support Vector Machine* - SVM, *Boosting* - GBM e *Random Forest* - RF).

3.4 Avaliação dos resultados

Para se elaborar a modelagem por meio de Mineração de Dados, é necessário primeiramente se separar a base de dados em duas, sendo uma para o treinamento dos algoritmos, em que são determinadas automaticamente as

características necessárias para se representar a informação desejada, e outra para o teste dos modelos gerados, que neste estudo foi feito por meio de validação cruzada (*cross validation*) (REFAEILZADEH et al., 2009).

Para a avaliação de desempenho dos modelos desenvolvidos por meio de mineração de dados, 30% dos dados foram selecionados aleatoriamente da base original, criando-se assim uma base específica de teste, contendo 3.980 dados de produtividade de cana de açúcar. A análise do desempenho dos modelos foi efetuada pela comparação entre as produtividades estimadas pelos modelos e as produtividades observadas por meio da análise de regressão. O desempenho dos modelos foi avaliado considerando-se os seguintes índices estatísticos: coeficiente de correlação (r); coeficiente “d” de Willmott (1985), que representa a distância relativa dos valores estimados da reta 1:1; raiz quadrada do erro médio (*Root Mean Square Error, RMSE*), que indica a acurácia do modelo; erro absoluto médio (EAM), o qual indica a magnitude dos erros associados à estimação da produtividade; coeficiente de confiança “C” de Camargo e Sentelhas (1997), que representa o grau de concordância do modelo, englobando em um único índice a precisão e a acurácia.

3.5 Distribuição

A distribuição do conhecimento se dará por meio da divulgação dos resultados junto às usinas que forneceram os dados para este estudo, de modo a posicioná-los a respeito dos fatores mais importantes que afetam a produtividade da cultura da cana de açúcar, para que os mesmos possam considerados em ações que proporcionem uma maior eficiência na produção de cana de açúcar em cada uma das unidades consideradas.

4 RESULTADOS E DISCUSSÃO

4.1 Correlações entre as variáveis

Para se obter uma visão geral prévia da importância de cada atributo (variável) utilizado nos algoritmos de Mineração de Dados, os mesmos foram correlacionados entre si e com a produtividade observada (TCH observado) gerando um diagrama de correlações, apresentado na Figura 10. Esse diagrama foi elaborado empregando-se a *Correlação de Pearson* (r), sendo que quanto maior r, maior será a intensidade das cores azul (correlação positiva entre as variáveis) e vermelha (correlação negativa). Como se pode notar, as maiores correlações encontradas com a produtividade (consolidado$TCHREAL) foram para as variáveis independentes: número de corte (NCORTE), com correlação negativa; e idade de corte (IDADE), com correlação positiva (Figura 10). Em seguida, com correlações menores, mas relativamente maiores que as demais variáveis, estão a evapotranspiração potencial (ETP), evapotranspiração da cultura (ETC) e precipitação pluviométrica (CHUVA), tanto para o período total de crescimento da cultura, quanto para a fase de 100% de cobertura do solo (F5). A fase F5 é a mais longa do ciclo da cultura da cana de açúcar, com períodos que variam entre 17 decêndios (≈ 170 dias) para as canas de 12 meses (planta e soca) e 32 decêndios (≈ 320 dias) para a cana planta de 18 meses.

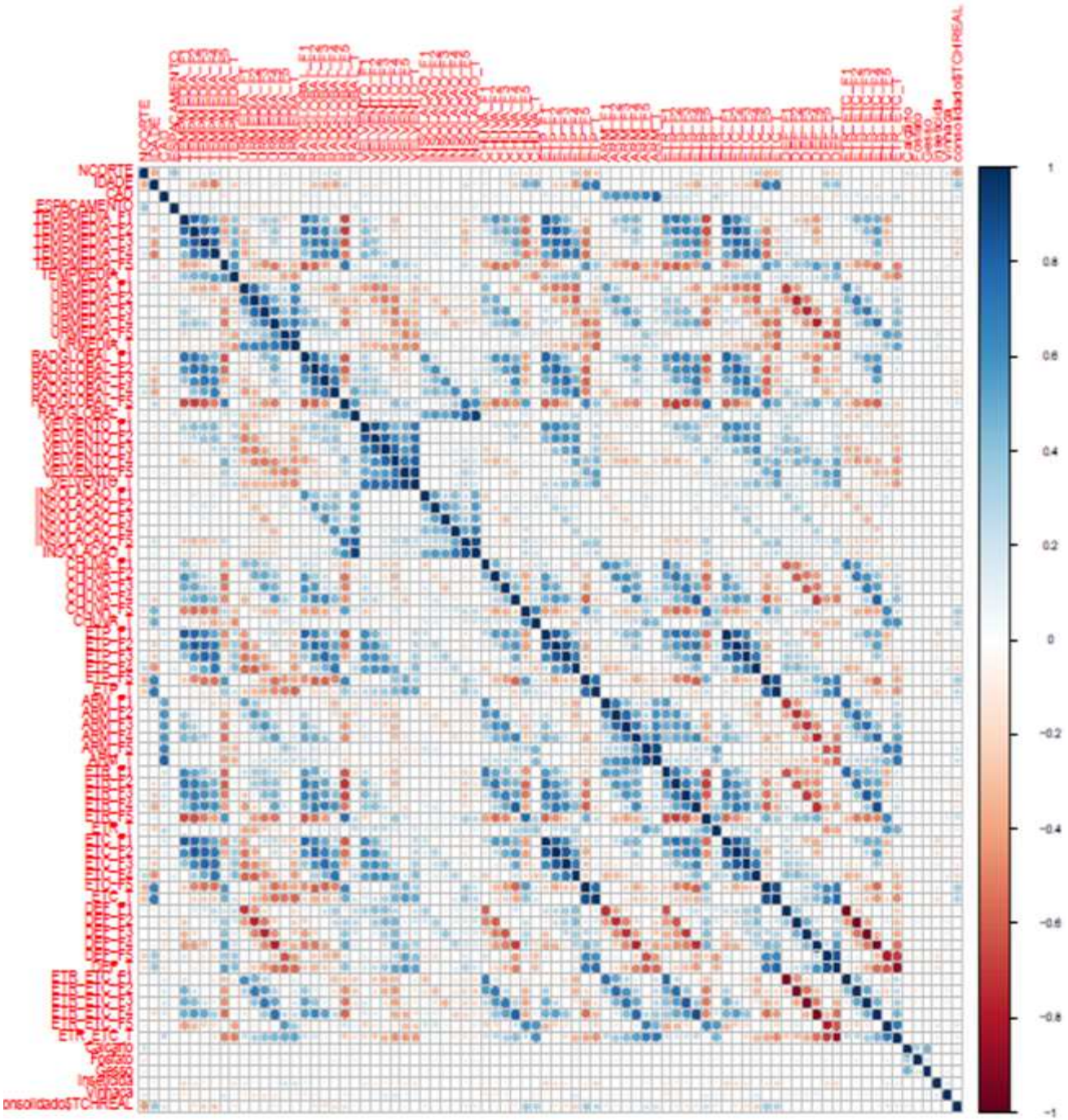

Figura 10 - Diagrama de correlações entre as variáveis independentes e entre essas e a produtividade observada da cana de açúcar (consolidado$TCHREAL). A relação do significado de cada uma dessas variáveis é apresentada no Apêndice A.

Na Tabela 4 são apresentados os valores numéricos das correlações das variáveis independentes com a produtividade observada (consolidado$TCHREAL). A aplicação do teste de significância das correlações, a 0,05 de significância e graus

de liberdade > 100, resultou em um valor crítico de 0,195 (IRELAND, 2010), indicando que coeficientes de correlação abaixo desse valor não são significativos.

Tabela 4 - Coeficientes de correlação de Pearson (r) da relação entre a produtividade observada de cana de açúcar (consolidado$TCHREAL) e as variáveis independentes

r	Variável	r	Variável	r	Variável	r	Variável
-0,42	NCORTE	0,17	ETR_ETC_F5	0,09	VELVENTO_F5	-0,05	Gesso
0,37	IDADE	-0,16	ETR_F4	-0,08	VELVENTO_F1	-0,04	RADGLOBAL_F1
0,35	CHUVA_F5	-0,15	ETP_F3	0,08	ARM_F3	-0,04	VELVENTO_F3
0,34	ETP_F5	-0,15	ETC_F3	-0,08	CHUVA_F2	0,04	INSOLACAO_F3
0,34	ETC_F5	-0,15	ETR_F3	-0,08	TEMPMEDIA_F1	-0,04	DEF_F1
0,30	CHUVA_T	-0,14	ETR_F2	0,07	ARM_F4	-0,04	URMEDIA_T
0,30	ETC_T	-0,14	TEMPMEDIA_T	-0,07	ETR_ETC_F4	-0,04	URMEDIA_F2
-0,29	TEMPMEDIA_F4	-0,14	RADGLOBAL_F2	-0,07	CHUVA_F3	-0,04	ETR_ETC_F2
0,28	ETP_T	0,13	ARM_F1	0,07	DEF_F5	-0,04	ETR_ETC_F3
0,23	ETR_F5	0,13	ETR_T	-0,06	VELVENTO_F2	-0,03	ESPACAMENTO
-0,23	TEMPMEDIA_F3	0,13	ETR_ETC_T	-0,06	ETP_F1	0,02	ETR_ETC_F1
-0,23	ETP_F4	-0,13	ETP_F2	-0,06	ETC_F1	0,02	Fosfato
-0,23	ETC_F4	-0,13	ETC_F2	0,06	INSOLACAO_F2	0,02	URMEDIA_F1
0,21	ARM_T	-0,13	Calcario	-0,05	URMEDIA_F3	0,02	VELVENTO_T
0,21	ARM_F5	0,12	INSOLACAO_F5	0,05	DEF_T	0,01	DEF_F2
0,20	CAD	-0,12	CHUVA_F4	0,05	TEMPMEDIA_F5	0,01	Vinhaca
-0,20	RADGLOBAL_F4	0,11	INSOLACAO_T	0,05	RADGLOBAL_T	0,01	DEF_F3
0,19	RADGLOBAL_F5	-0,11	URMEDIA_F4	0,05	INSOLACAO_F4	0,01	CHUVA_F1
-0,18	RADGLOBAL_F3	0,09	INSOLACAO_F1	-0,05	URMEDIA_F5	0,00	DEF_F4
-0,18	TEMPMEDIA_F2	0,09	ARM_F2	-0,05	ETR_F1	0,00	Inseticida
						0,00	VELVENTO_F4

Conforme já observado na Figura 10, a correlação mais significativa entre a produtividade e os fatores que a condicionam se deu com o número de cortes do canavial (NCORTE), com r de -0,42. Isso significa que à medida que se aumenta o número de cortes há um decaimento da produtividade da cana de açúcar. Apesar de a Correlação de Pearson não ser um indicativo da relação causa-efeito entre duas variáveis, e sim, da intensidade da relação entre as mesmas, pode-se inferir que a correlação negativa entre o NCORTE e a produtividade da cana de açúcar encontrada nesse trabalho significa que à medida que aumenta a idade do canavial em termos de números de cortes, sua produtividade decai. De acordo com Garside et al. (2005), o decaimento das produtividades ao longo de sucessivos cortes da

cana é um tema complexo, causado por uma combinação de fatores em um sistema de produção de cana de açúcar. Bernardes et al. (2008), ao estudarem a queda de produtividade ao longo do ciclo da cana-de-açucar, geraram uma função exponencial negativa para estimar a produtividade nos sucessivos cortes dos canaviais e concluíram que o decaimento da produtividade da cana de açúcar ao longo dos sucessivos cortes apresenta maior dependência do conjunto de práticas de manejo, como intensidade do controle de pragas e plantas daninhas, correção e adubação dos solos, bem como controle de danos às soqueiras durante a colheita mecanizada, do que do potencial produtivo do ambiente de produção. Também para a variável IDADE, que representa a idade do canavial em termos de meses entre o plantio e a colheita para a cana planta e entre sucessivos cortes na cana soca, foi encontrada uma correlação significativa com a produtividade observada, sendo esta positiva. A influência da idade do canavial no momento do corte tem sido objeto de muita discussão (LAWES et al., 2002a). Analisando bancos de dados de uma usina situada no estado de Queensland, Austrália, Lawes et al. (2002b) encontraram influência significativa da idade nas produtividades dos blocos analisados, sendo que canas com 8 meses no momento da colheita obtiveram em média 7 t ha^{-1} a menos que canas com idade de 14 meses. Apesar de óbvia, a influência da idade se confunde com a influência de outros fatores, como a época de corte do canavial (McDONALD, 2006) e padrões meteorológicos durante a safra (INMAN-BAMBER, 1994).

Com relação às variáveis ETP, ETC e CHUVA, todas apresentaram $r \geq 0,30$ quando correlacionadas com a produtividade da cana, tanto na fase F5 como no ciclo como um todo. É evidente o efeitos dessas variáveis sobre o processo produtivo da cana (MARIN et al., 2009; MONTEIRO, 2011), já que maiores valores de ETP indicam indiretamente maior disponibilidade de energia no ambiente para o crescimento das plantas, integrando as variáveis radiação solar, temperatura do ar e fotoperíodo, enquanto que maiores valores de CHUVA e a ETC demonstram que houve maior disponibilidade de água para o crescimento das plantas, reduzindo os impactos negativos da deficiência hídrica.

4.2 Determinação dos principais fatores determinantes da produtividade da cana de açúcar utilizando-se diferentes técnicas de Mineração de Dados

Como já mencionado, o pacote *caret* possibilita a identificação das variáveis preditoras segundo sua importâncias nos modelos de mineração de dados, ordenando-as da variável mais relevante a menos relevante.

4.2.1 Base de dados integral

a) *Random Forest*

A Figura 11 apresenta a importância relativa das variáveis determinantes da produtividade da cana de açúcar gerada pelo modelo *Random Forest*. Com a aplicação do algoritmo *Random Forest*, obteve-se uma listagem dos atributos mais importantes para a determinação da produtividade da cana-de-açucar. O modelo aplicado teve seus parâmetros fornecidos pelo pacote Caret, com 1.500 árvores. Como todos os atributos são ordenados de acordo com sua importância, foram utilizadas para as comparações com os outros modelos (*Boosting* e *Support Vector Machine*) as variáveis classificadas entre as dez melhores, uma vez que as demais variáveis contribuíram muito pouco para explicar a variabilidade da produtividade pelo modelo em questão. Nota-se na Figura 11 que a variável “NCORTE”, ou seja, uma variável de manejo que representa a idade dos canaviais em termos de número de cortes já efetuados teve o maior peso na explicação da variação da produtividade, o que está de acordo com o resultado obtido pela análise de correlação (Tabela 4). Na sequencia, as variáveis que representam o armazenamento de água no solo (“CAD” e “ARM_F5”) e as variáveis meteorológicas umidade relativa, velocidade do vento e chuva, todas para o ciclo como um todo (“URMEDIA_T”, “VELVENTO_T”, e “CHUVA_T”) também ajudam a explicar a variabilidade espaço-temporal da produtividade da cana de açúcar nas 18 usinas avaliadas.

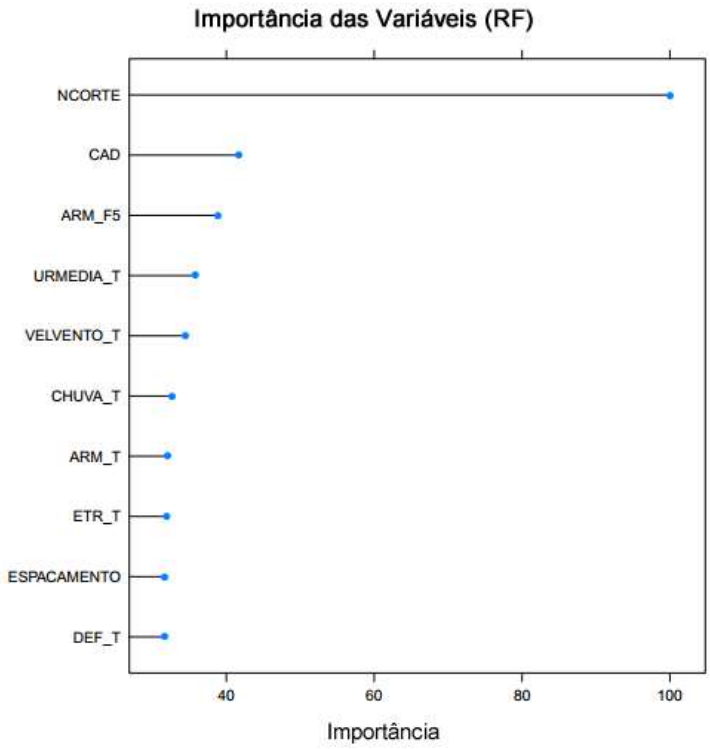

Figura 11 – Classificação das dez principais variáveis determinantes da produtividade da cana de açúcar pelo método do *Random Forest* para a base completa de dados

b) *Boosting*

Assim como o *Random Forest*, o algoritmo *Boosting* foi utilizado para determinar a classificação das variáveis mais importantes para a definição da produtividade da cana de açúcar (TCH). De forma semelhante, este algoritmo também foi aplicado com 1500 árvores, o que resultou na classificação da importância relativa das variáveis determinantes da produtividade (Figura 12). Nota-se que, do conjunto dos dez atributos mais importantes selecionados pelo *Boosting*, quatro também foram selecionados pelo *Random Forest*, sendo eles: o número de cortes do canavial (NCORTE); a capacidade de água disponível do solo (CAD); o armazenamento de água na fase fenológica de número 5 (ARM_F5); e a chuva acumulada de todo o período de desenvolvimento da cultura (CHUVA_T).

Também nesta abordagem, o NCORTE foi a variável que mais impactou a variabilidade da produtividade da cana de açúcar, corroborando os resultados obtidos pela técnica do Random Forest (Figura 11) e também a análise de correlação (Tabela 4). No entanto, por meio desta técnica foi possível se confirmar que a IDADE do canavial em meses é outra variável de grande importância, como já

evidenciado por meio da análise de correlação (Tabela 4) e discutido por Lawes et al. (2002 a, 2002b).

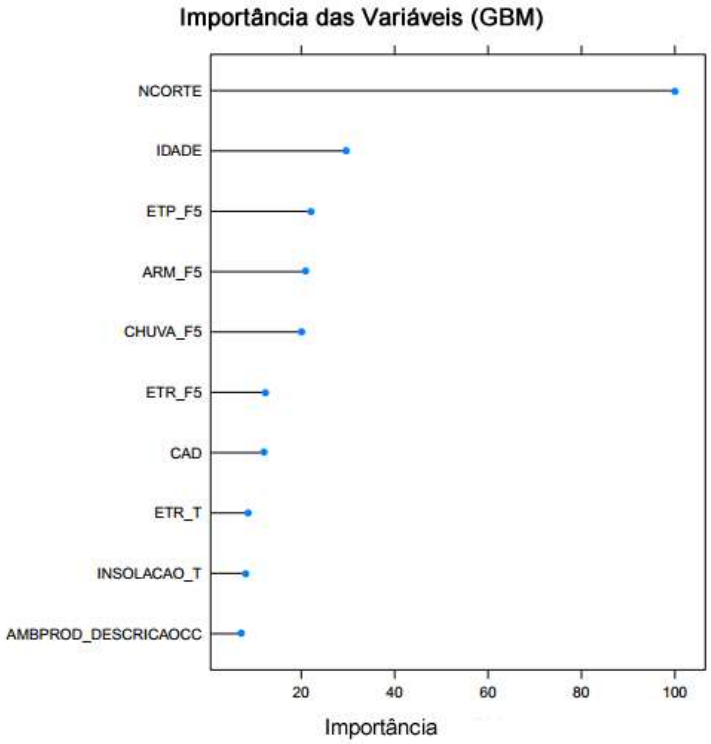

Figura 12 – Classificação das dez principais variáveis determinantes da produtividade da cana de açúcar pelo método do *Boosting* para a base total de dados

c) *Support Vector Machine*

A Figura 13 apresenta a importância relativa das variáveis determinantes da produtividade da cana de açúcar (TCH), gerada pela técnica *Support Vector Machine*. Das dez variáveis mais importantes selecionadas pelo *Support Vector Machine* para explicar a variação da produtividade da cana de açúcar, cinco também foram selecionadas pelo algoritmo *Boosting*, sendo elas: número de cortes do canavial (NCORTE); idade do canavial em meses no momento do corte (IDADE); as evapotranspirações potencial e real durante a fase fenológica 5 (ETP_F5 e ETR_F5); e a chuva acumulada da fase 5 (CHUVA_F5). Dessas variáveis em comum, duas delas aparecem inclusive na mesma ordem de importância (NCORTE e IDADE), o que também foi observado ao se avaliar os coeficientes de correlação de Person (Tabela 4). Já em comparação ao *Random Forest*, apenas o número de

corte (NCORTE) e a chuva acumulada no período de desenvolvimento total (CHUVA_T) aparecem como comuns para ambos os métodos.

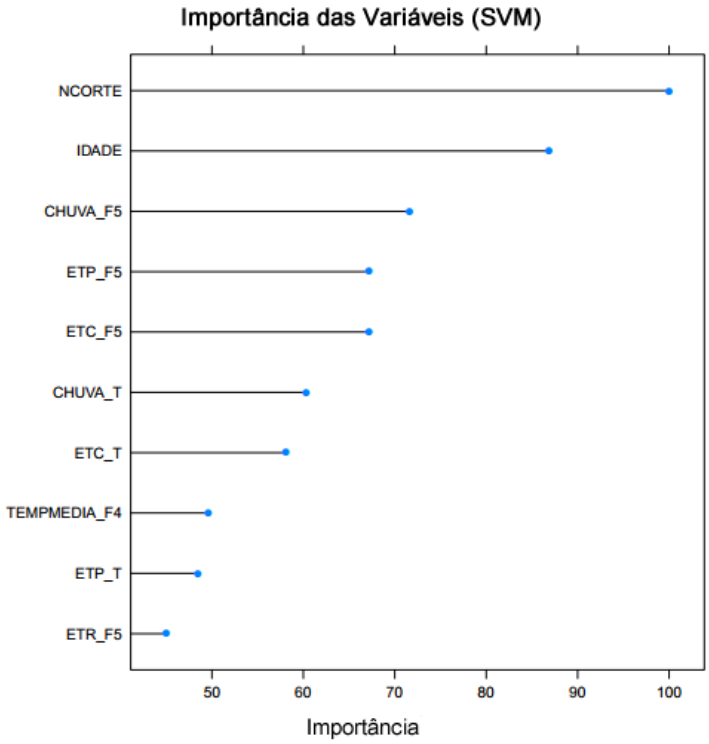

Figura 13 – Classificação das dez principais variáveis determinantes da produtividade da cana de açúcar pelo método do *Support Vector Machine* para a base total de dados

A Figura 14 apresenta o Diagrama de Venn, o qual apresenta as dez variáveis selecionadas pelos modelos empregados e a concordância entre eles para selecionar as variáveis que mais afetaram a produtividade da cana de açúcar.

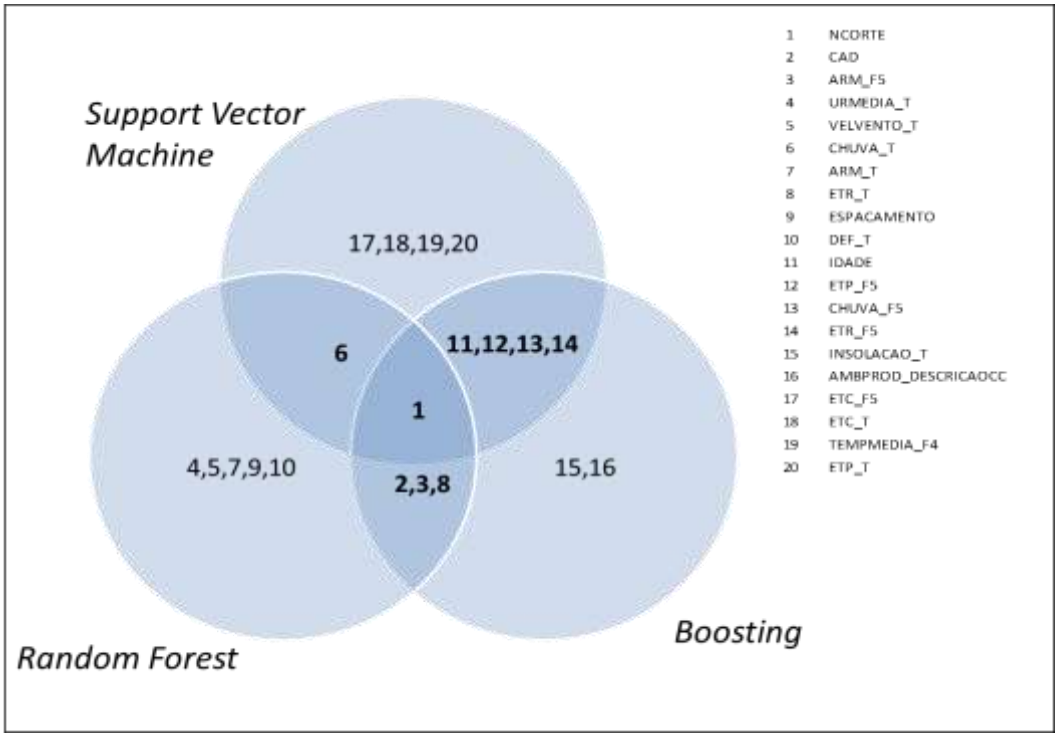

Figura 14 - Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis determinantes da produtividade da cana de açúcar, considerando-se a base total de dados

De forma resumida, é possível constatar que, para todos os modelos empregados, o número de cortes (NCORTE), ou seja, uma variável de manejo despontou como o fator com maior peso na determinação da produtividade da cana de açúcar. Constata-se também que essa variável tem importância relativa bem maior do que as demais variáveis. Em seguida, em dois dos três modelos, aparecem variáveis relacionadas ao regime hídrico da cultura, como a precipitação (CHUVA) e a evapotranspiração (ETP_T, ETP_F5), principalmente a da fase de desenvolvimento 5 (fase em que há 100% de cobertura do solo, F5), bem como atributos edafoclimáticos (CAD e ARM). Em trabalho similar, empregando o modelo CART, considerado como uma técnica de mineração de dados, Ferraro et al. (2009), ao analisarem dados de produtividade de seis fazendas de uma usina situada no norte da Argentina, observaram que a produtividade foi afetada principalmente pelas seguintes variáveis: variedade plantada; fazenda de origem de cada bloco colhido; número de cortes das canas socas; mês de corte de cada bloco; duração do período de desenvolvimento para cada corte (idade em dias); área dos blocos colhidos; e

precipitação de cada bloco (total e para o período de verão). No ranking de importância, a variável referente à fazenda de origem, que, de acordo com os autores, captura a influência do manejo local, da variabilidade edafoclimática, bem como a influência de fatores de estresse, como pragas e doenças, foi a que teve a maior colocação, seguida do número de cortes. Em outro trabalho de análise de dados ao nível de talhão, Bocca (2014) modelou as produtividades da cana de açúcar de uma usina situada no estado de SP, por meio de sete técnicas diversas, entre elas o *Random Forest* e o *Support Vector Machine*, e concluiu que o número de cortes foi uma das três variáveis consideradas por todas as técnicas como sendo importante para explicar a variação das produtividades, corroborando, assim, com o resultado deste estudo.

4.2.2 Base de dados regionalizada

Os resultados da aplicação dos algoritmos *Random Forest, Boosting* e *Support Vector Machine* na base de dados das três usinas situadas na região de Bauru são apresentados a seguir.

a) *Random Forest*

A Figura 15 apresenta a importância relativa das variáveis determinantes da produtividade da cana de açúcar gerada pelo modelo *Random Forest*, para a região de Bauru. Observa-se que a variável referente ao número de cortes do canavial (NCORTE) teve o maior peso na explicação da variação da produtividade da cana de açúcar pelo modelo, sendo esta expressivamente maior que a segunda variável mais importante, ou seja, a capacidade de água disponível dos solos (CAD). Essas duas variáveis também foram as principais quando da análise com os dados completos de todas as regiões, o que mostra a importância delas independentemente da escala espacial. Para as demais variáveis selecionadas não houve concordância, porém ficou claro que tanto as variáveis meteorológicas como as relativas ao ambiente de produção (CAD e tipo de ambiente) foram importantes na definição dos níveis de produtividade da cana de açúcar.

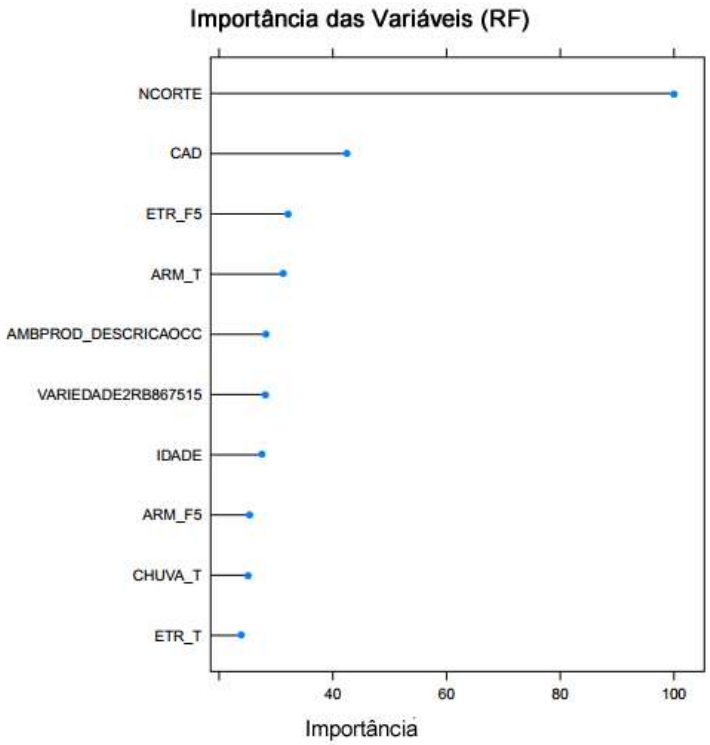

Figura 15 - Classificação das dez principais variáveis determinantes da produtividade da cana de açúcar pelo método do *Random Forest* para a base de dados da região de Bauru, SP

b) *Boosting*

A Figura 16 apresenta a importância relativa dos dez atributos mais importantes selecionados pela técnica do *Boosting* para a base de dados da região de Bauru. Destes, sete foram também selecionados pelo *Random Forest*, sendo que aqui também o número de cortes do canavial (NCORTE) foi a variável mais importante. Em relação à análise com o banco total de dados, houve concordância para as duas primeiras variáveis, relacionadas ao manejo da cultura, ou seja, número de cortes (NCORTE) e idade da cana na colheita em meses (IDADE). As demais variáveis selecionadas indicaram importância tanto de variáveis meteorológicas como do ambiente de produção (CAD e tipo de ambiente), assim como observado pela técnica do *Random Forest*.

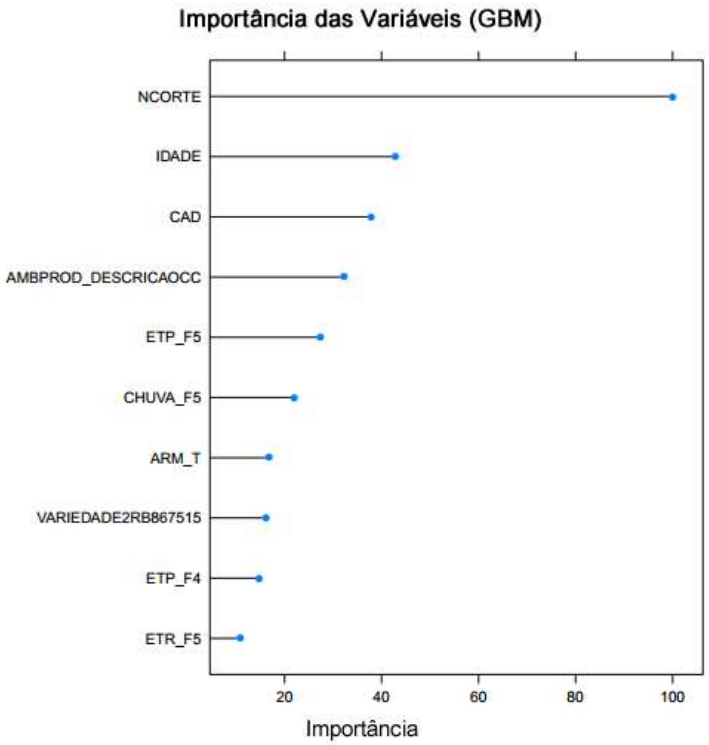

Figura 16 - Classificação das dez principais variáveis determinantes da produtividade da cana de açúcar pelo método do *Boosting* para a base de dados da região de Bauru

c) *Support Vector Machine*

A Figura 17 apresenta a importância relativa das variáveis determinantes da produtividade da cana de açúcar obtida pelo modelo *Support Vector Machine.* Neste caso, houve uma grande similaridade dos resultados com as variáveis selecionadas pelo modelo do *Boosting* e, de certa forma, com o *Random Forest*. Assim, o NCORTE e a IDADE foram as duas principais variáveis a definir a produtividade da cana de açúcar na região de Bauru, mostrando que, assim como para as demais análises, o ambiente de produção e as condições meteorológicas tiveram uma importância secundária.

De um modo geral, os resultados obtidos pela análise regionalizada não diferiram substancialmente dos resultados apresentados na análise com o banco geral de dados. No entanto, pode-se observar que, no caso da análise regionalizada, houve uma maior concordância entre os três modelos de mineração de dados para a classificação das variáveis de importância para a produtividade da cana, como se pode observar na Figura 18. Na análise regionalizada, houve a concordância de cinco variáveis (NCORTE, IDADE, CAD, AMBPROD e ARM_T), o que mostra que

ao nível regional a interação entre manejo, ambiente de produção e balanço de água no solo é o que define as produtividades na região de Bauru. Obviamente, isso pode se alterar para as diferentes regionais, já que os ambientes de produção e o clima diferem de uma regional para outra.

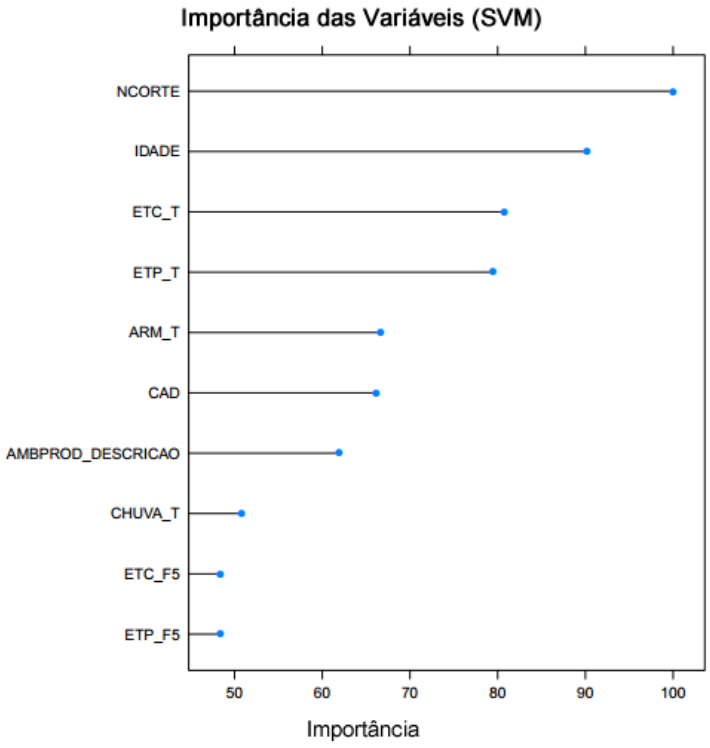

Figura 17 - Classificação das dez principais variáveis determinantes da produtividade da cana de açúcar pelo método do *Support Vector Machine* para a base de dados da região de Bauru, SP

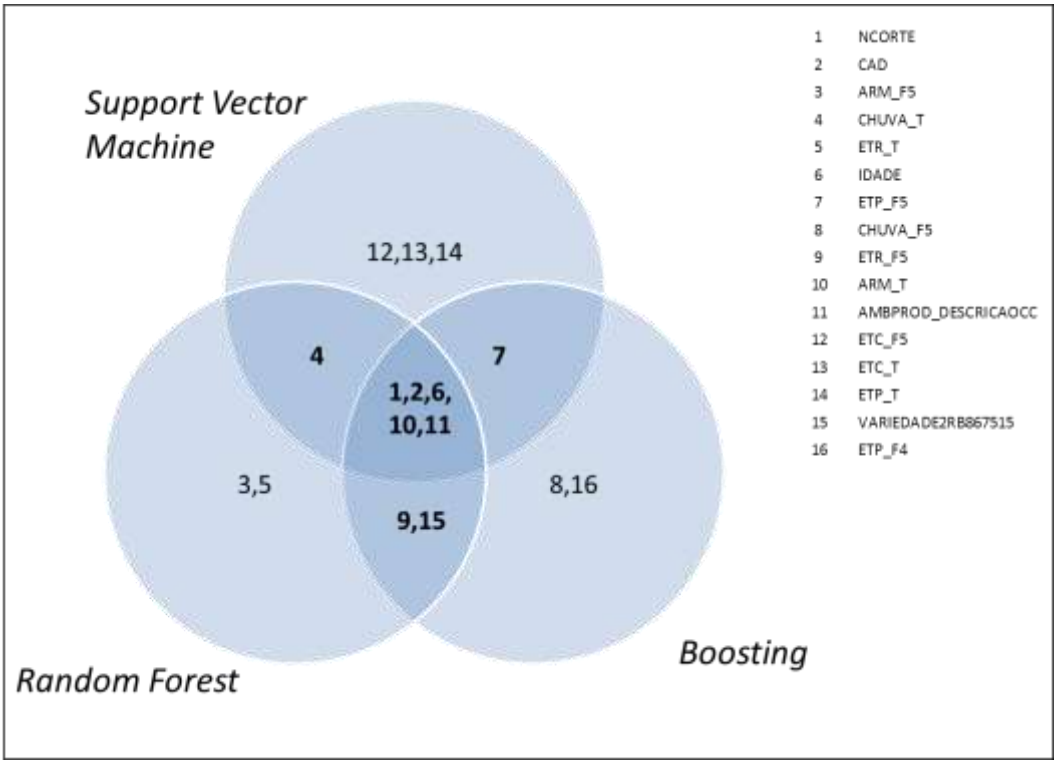

Figura 18 - Diagrama de Venn para a indicação da concordância entre os modelos de mineração de dados empregados na classificação das variáveis determinantes da produtividade da cana de açúcar, considerando-se a base de dados da região de Bauru, SP

Outro aspecto em comum em ambas as abordagens (geral e regionalizada), é o fato das variáveis relacionadas à aplicação de insumos não terem apresentado correlações significativas com a produtividade da cana de açúcar. A importância relativamente menor dessas variáveis nos modelos não implica que elas não sejam relevantes para a explicação da variável resposta, e sim, que, para os modelos em questão, a utilização desses atributos não gerou contribuição maior que a utilização dos demais, na explicação das produtividades. Uma possível razão para isso é que os dados relacionados às variáveis de manejo, devido à forma em que foram utilizadas (aplicação do insumo em termos de percentual de área dos blocos), geraram pouca variabilidade, de forma que, em função da natureza dos dados, os modelos não conseguiram extrair informações a partir dessas variáveis, de forma que pudessem contribuir de forma significativa para explicar pela modelagem a variabilidade das produtividades.

4.3 Avaliação dos modelos de Mineração de Dados para a estimação da produtividade da cana de açúcar

Assim como para a ordenação dos fatores determinantes da produtividade da cana de açúcar, descrita anteriormente, as técnicas de mineração de dados *Random Forest* (*RF*), *Boosting* (*GBM*) e *Support Vector Machine* (*SVM*) foram empregadas para a obtenção de modelos de estimação da produtividade dessa cultura. Em uma primeira etapa, os modelos gerados foram testados com dados independentes na estimação da produtividade da cana-de-açúcar, considerando-se a base integral de dados, contendo todas as regiões analisadas. Numa segunda etapa, o teste dos modelos se seu de forma regionalizada, para a região de Bauru, de modo a se identificar as diferenças decorrentes da utilização do banco de dados geral e de um banco específico para uma região.

4.3.1 Base de dados integral

Os algoritmos obtidos para a estimação da produtividade da cana de açúcar pelas técnicas de Mineração de Dados *Random Forest, Boosting* e *Support Vector Machine* foram aplicados ao conjunto de dados de validação, que contém 3980 dados de produtividades (30% da base completa). Os resultados do processo de teste desses modelos são apresentados nas Figuras 19 a 21, onde se observa um desempenho muito similar entre os três modelos baseados na mineração de dados, com r entre as produtividades estimadas e observadas variando de 0,65 a 0,66, e a RMSE de 19,70 a 20,03 t ha^{-1}.

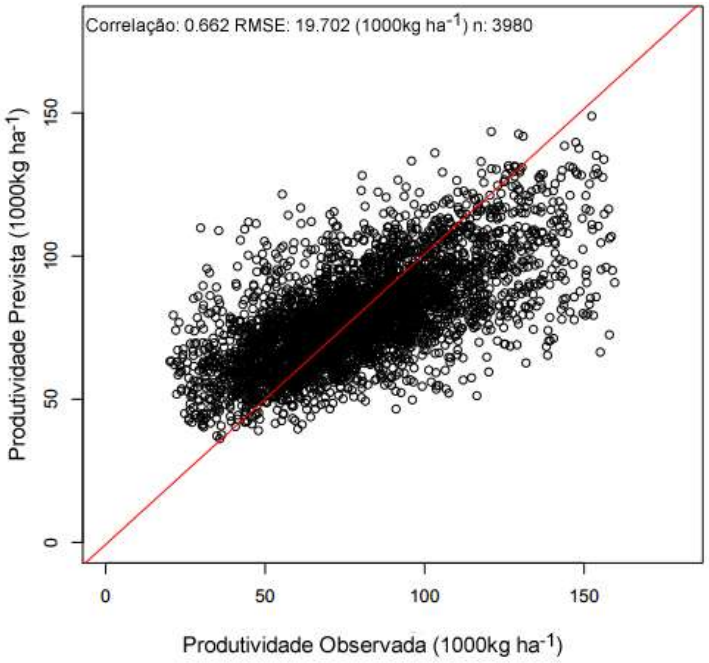

Figura 19 – Relação entre a produtividade da cana de açúcar observada e a estimada pelo algoritmo gerado pela técnica do *Random Forest*, considerando-se o conjunto de dados independentes para a base integral de dados

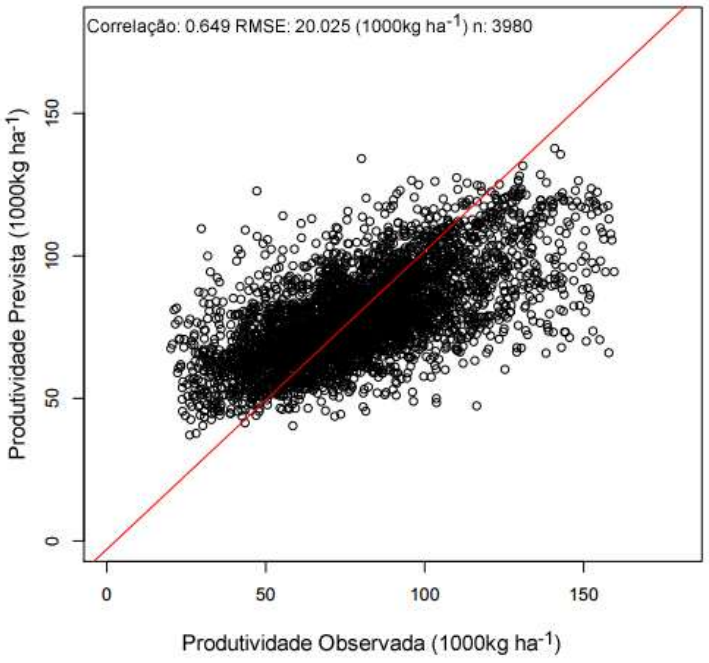

Figura 20 - Relação entre a produtividade da cana de açúcar observada e a estimada pelo algoritmo gerado pela técnica do *Boosting*, considerando-se o conjunto de dados independentes para a base integral de dados

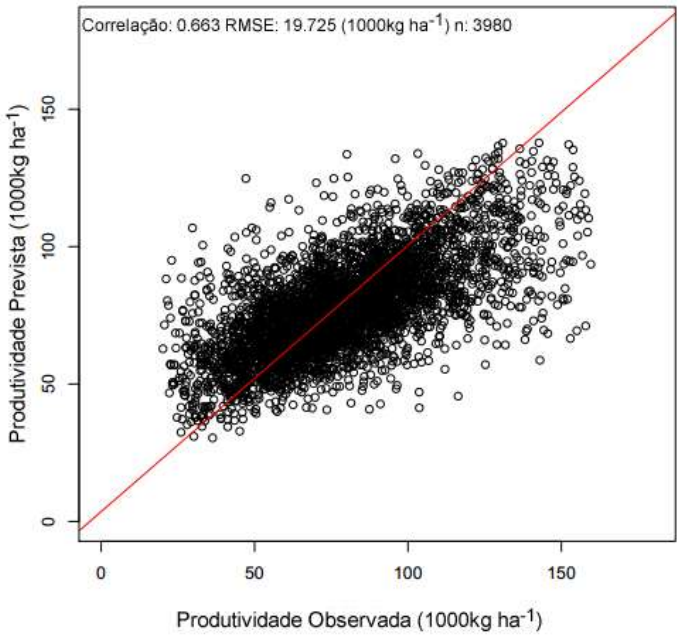

Figura 21 - Relação entre a produtividade da cana de açúcar observada e estimada pelo algoritmo gerado pela técnica do *Support Vector Machine*, considerando-se o conjunto de dados independentes para a base integral de dados

Além disso, a avaliação dos modelos utilizados foi também feita por meio de outros indicadores estatísticos como o índice d de Willmott, o índice C e o EAM. A Tabela 3 apresenta os resultados desses indicadores, que mostram uma similaridade muito grande entre os modelos de estimação, porém com um desempenho um pouco melhor para o modelo *Support Vector Machine*, que apresentou os maiores valores de r, d e C, o menor EAM e um tempo de processamento de 2,46 h, maior do que o *Boosting*, porém bem menor do que o modelo do *Random Forest*, o qual necessita de aproximadamente 17 h para o processamento dos dados.

Tabela 5 – Indicadores estatísticos para avaliação dos modelos empregados na estimativa da produtividade da cana de açúcar, empregando-se o conjunto de dados independentes: coeficiente de correlação (r); índice de Willmott (d); índice de confiança (C); erro absoluto médio (EAM); raíz quadrada do erro médio (RMSE); e tempo de processamento dos modelos

Modelo	**r**	**D**	**C**	**EAM (t ha^{-1})**	**RMSE (t ha^{-1})**	**Tempo de Processamento (h)**
Random Forest	0,662	0,772	0,511	14,928	19,702	16,962
Gradiente Boosting	0,649	0,754	0,490	15,347	20,025	0,034
Support Vector Machine	0,663	0,779	0,517	14,831	19,725	2,460

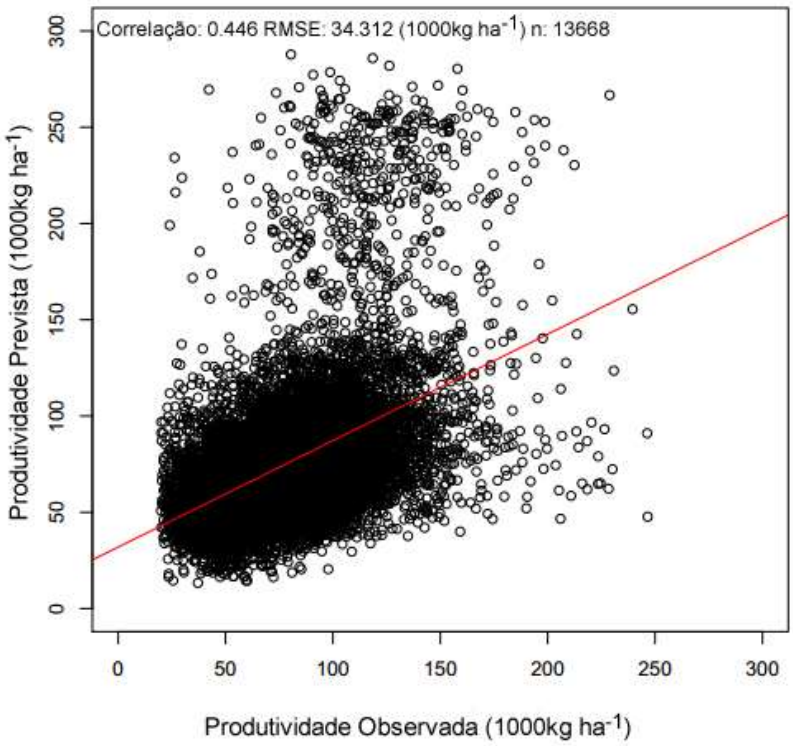

Figura 22 - Relação entre a produtividade da cana de açúcar observada e a estimada pelo modelo agrometeorológico da FAO, considerando-se o conjunto de dados independentes para a base integral de dados

Os resultados obtidos mostram que a estimação da produtividade comercial da cana de açúcar com modelos baseados em técnicas de mineração de dados apresentaram desempenho similar aos modelos empírico, matemático-fisiológicos e

mecanísticos de simulação da produtividade. Resultados de O´Leary (2000), Nassif et al. (2012), Marin et al. (2011), Monteiro e Sentelhas (2014), Vianna e Sentelhas (2015) e Monteiro (2015), empregando diferentes tipos de modelos de estimação da produtividade da cana de açúcar, apresentaram RMSE variando de 5,6 a 30 t ha^{-1}, porém nenhum desses estudos empregou uma base de dados tão grande quanto à utilizada aqui, ou seja, com 3980 dados de produtividade ao nível de talhão, oriundos de diferentes regiões climáticas, ambientes de produção (tipo de solo), tipos de cana (planta e soca), ciclos e manejos da cultura.

Essa enorme granularidade de dados nos dá a certeza de que mesmo com um RMSE da ordem de 19,7 t ha^{-1}, as estimativas obtidas por meio dos modelos gerados a partir das técnicas de mineração de dados são mais acuradas do que qualquer outro tipo de modelo, já que esses não levam em consideração aspectos do manejo da cultura. Isso pode ser provado ao se comparar os desempenhos obtidos pelo modelo da FAO, o mesmo empregado por Monteiro e Sentelhas (2014) e Monteiro (2015), e pelos modelos de mineração de dados. Enquanto os modelos de MD resultaram RMSE da ordem de 20 t ha^{-1}, o RMSE do modelo da FAO foi aproximadamente 70% maior (≈ 34 t ha^{-1}). O potencial preditivo das técnicas de mineração de dados pode ser constatado também no trabalho de Everingham et al. (2016), que, empregando o algoritmo *Random Forest*, utilizando dados de produtividades regionais de uma usina localizada no nordeste da Austrália, obtiveram resultados com RMSE variando de 6,3 a 8 t ha^{-1} na estimação das produtividades da região.

4.3.2 Base de dados regionalizada

Da mesma forma que na estimação da produtividade utilizando-se a base de dados integral, os algoritmos *Random Forest, Boosting* e *Support Vector Machine* foram aplicados ao conjunto de dados de validação da base da região de Bauru, que contém 1116 dados de produtividades (30% da base regionalizada). Os resultados do processo de teste desses modelos são apresentados nas Figuras 23 a 25, onde também se observa um desempenho muito similar entre os três modelos baseados na mineração de dados, em termos de correlação (r) entre as produtividades estimadas e observadas, bem como RMSE. A síntese dos testes RMSE, d, EAM e C, bem como o tempo de processamento que foi necessário para executar cada

modelo, são apresentados na Tabela 6. Pode-se observar que, assim como na base de dados integral, os resultados dos testes de todos os modelos foram similares entre si. Porém, constata-se que o algoritmo *Support Vector Machine* superou levemente os demais modelos em todos os testes.

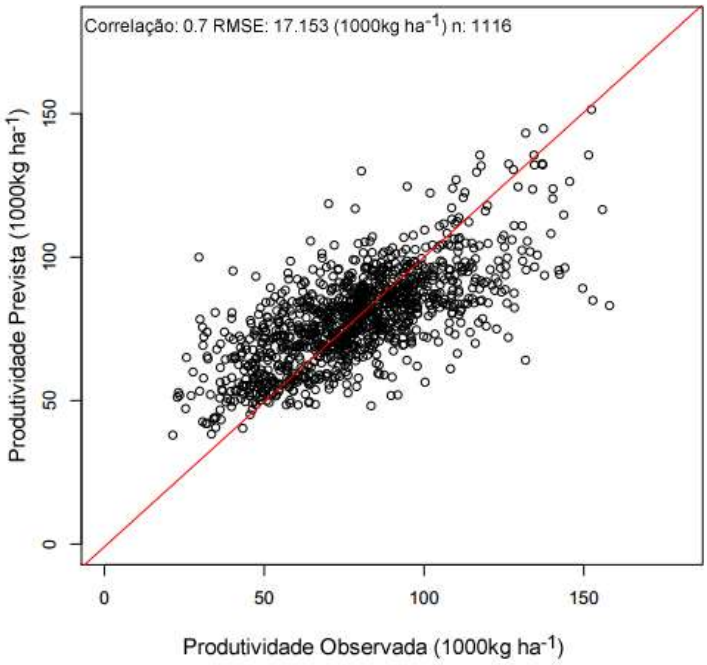

Figura 23 - Relação entre a produtividade da cana de açúcar observada e a estimada pelo algoritmo gerado pela técnica do *Random Forest*, considerando-se o conjunto de dados independentes para a base de dados da região de Bauru, SP

Apesar do desempenho pouco satisfatório dos modelos apresentados nas Figuras 19 a 21, esses foram melhores do que as estimativas de produtividade obtidas com o modelo agrometeorológico da FAO (Figura 22), empregado para os mesmos blocos que foram utilizados os modelos de Mineração de Dados. O modelo agrometeorológico, por considerar apenas a interação da cultura com as condições meteorológicas, não leva em conta as variáveis associadas ao manejo da cultura, resultando em um desempenho pior, não conseguindo representar com precisão e acurácia a enorme variabilidade da produtividade observada a campo de forma satisfatória, com r = 0,45 e RSME = 34,31 t ha^{-1}.

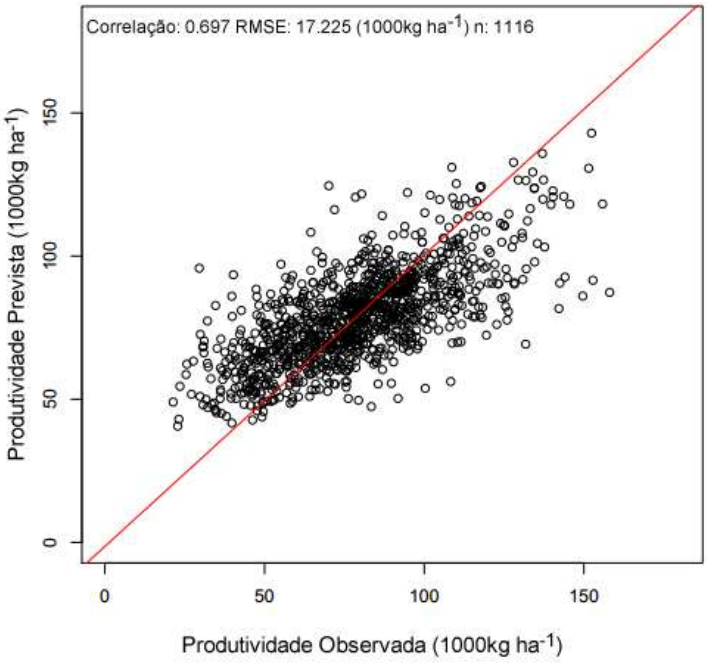

Figura 24 - Relação entre a produtividade da cana de açúcar observada e a estimada pelo algoritmo gerado pela técnica do *Boosting*, considerando-se o conjunto de dados independentes para a base de dados da região de Bauru, SP

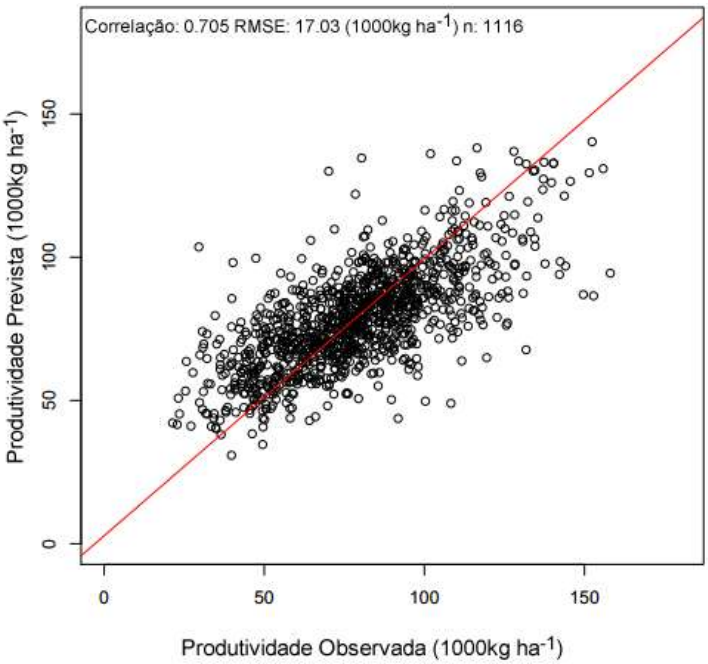

Figura 25 - Relação entre a produtividade da cana de açúcar observada e estimada pelo algoritmo gerado pela técnica do *Support Vector Machine*, considerando-se o conjunto de dados independentes para a base de dados da região de Bauru, SP

Tabela 6 – Indicadores estatísticos para avaliação dos modelos empregados na estimativa da produtividade da cana de açúcar, empregando-se o conjunto de dados independentes: coeficiente de correlação (r); índice de Willmott (d); índice de confiança (C); erro absoluto médio (EAM); raíz quadrada do erro médio (RMSE), e tempo de processamento dos modelos

Modelo	**r**	**d**	**C**	**EAM (t ha^{-1})**	**RMSE (t ha^{-1})**	**Tempo de Processamento (h)**
Random Forest	0,700	0,803	0,562	13,056	17,153	2,327
Gradiente Boosting	0,697	0,799	0,556	13,198	17,225	0,009
Support Vector Machine	0,705	0,814	0,574	12,838	17,030	0,040

Pode-se constatar também que, quando comparados com os resultados de todos os indicadores estatísticos das Tabelas 5 e 6, o desempenho preditivo de todos os modelos de mineração de dados foi superior na base de dados regionalizada, o que pode estar correlacionado com uma redução da variabilidade tanto das produtividades, quanto das variáveis que a condicionam, se comparado à análise do banco de dados integral. Conclui-se, assim, que a regionalização da modelagem, ou seja, a aplicação dos modelos de mineração de dados nas bases das regiões produtoras individuais, resultou em melhor desempenho preditivo que a modelagem aplicada na abordagem geral, que por sua vez abrange todas as unidades produtoras em todas as regiões do estado, e, com isso, uma variação muito maior dos fatores condicionantes da produtividade agrícola.

Na Figura 26 são apresentados os resultados da estimativa da produtividade por meio do modelo agrometeorológico da FAO para a região de Bauru, empregado para os mesmos blocos que foram utilizados para os modelos baseados em técnicas de mineração de dados. Como na estimativa pela abordagem geral, aqui o modelo agrometeorológico também apresentou desempenho pior do que o dos modelos de mineração de dados, com r = 0,45 e RSME = 32,77 t ha^{-1}. Isso pode ser explicado pelo fato de que tal tipo de modelo apenas contempla a interação da cultura com as condições meteorológicas, não levando em conta as variáveis associadas ao manejo

da cultura. Assim, este tipo de modelagem da produtividade não consegue capturar a grande variabilidade da produtividade observada a campo na região de análise.

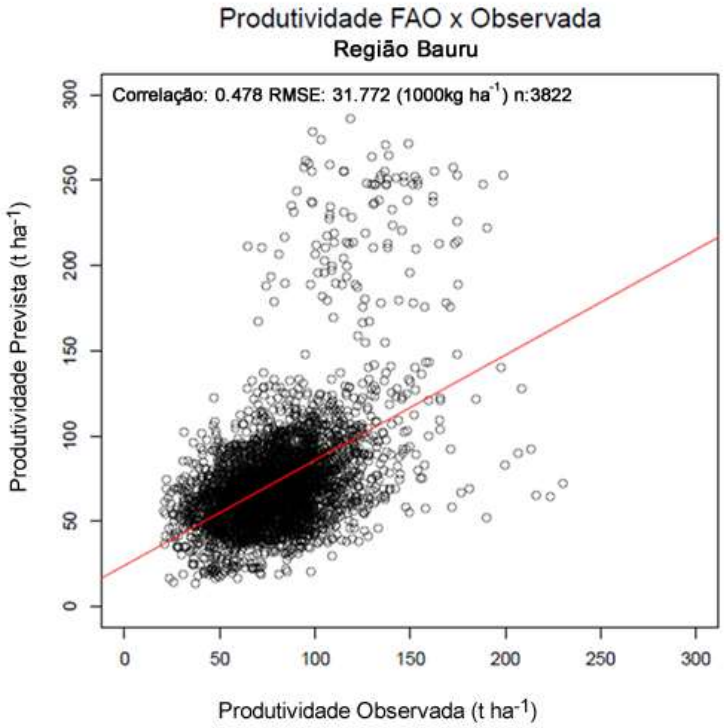

Figura 26 - Relação entre a produtividade da cana de açúcar observada e a estimada pelo modelo agrometeorológico da FAO, considerando-se o conjunto de dados independentes para a base de dados da região de Bauru

5 CONCLUSÕES

Conclui-se que, diante dos resultados obtidos que:

a) Foi possível elencar, bem como classificar por ordem de importância, diferentes fatores que afetam a produtividade da cana de açúcar em diferentes regiões do estado de São Paulo por meio da técnica de Mineração de Dados (MD), empregando-se bancos de dados operacionais das usinas de cana de açúcar. Dentre as variáveis testadas, o número de cortes (NCORTE) foi, para todos os modelos e em todas as regiões analisadas, a variável de maior importância;

b) Os modelos gerados por meio do emprego de técnicas de MD tiveram desempenho similar entre si, porém todos foram melhores do que o modelo agrometeorológico da FAO, ao nível de agrupamento de talhões. Não obstante, o desempenho da modelagem por MD foi aquém do que o desejável para sua aplicação operacional para a previsão de safra;

c) A regionalização da modelagem, ou seja, a aplicação dos modelos de mineração de dados na base de uma região produtora individual resultou em melhor desempenho que a modelagem aplicada na base de dados geral, que contempla todas as unidades produtoras em todas as regiões do estado.

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,** Basingstoke, v. 29, n. 13, p. 3753–3767, 2008.

ALLEN, R.G.; PEREIRA, L.S.; RAES, D.; SMITH, M. **Crop evapotranspiration:** guidelines for computing crop water requirements. Rome: FAO, 1998. 300 p. (FAO. Irrigation and Drainage Paper, 56).

ALMEIDA, T.I.R.; SOUZA FILHO, C.R.; ROSSETO, R. ASTER and Landsat ETM+ images applied to sugarcane yield. **International Journal of Remote Sensing,** Basingstoke, v. 27, n. 19, p. 4057-4069, 2006.

BAIER, W. Note on the terminology of crop-weather models. **Agricultural Meteorology**, Amsterdam, v. 20, p. 137-145, 1979.

BAILAY, R.A. Diseases. In: JAMES, G. **Sugarcane.** Oxford: Wiley Blackwell, 2004. p. 54-77.

BARROS, F.M.M. **Um sistema de recomendação de paginas web sobre a cultura da cana de açúcar.** 2013. 68 p. Dissertação (Mestrado em Engenharia Agrícola.) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2013.

BARROS, R.P.; ALMEIDA VIÉGAS, P.R.; SILVA, T.L.; SOUZA, R.M.; BARBOSA, L.; ALMEIDA VIÉGAS, R.; VASCONCELLOS BARRETTO, M.C.; SOARES DE MELO, A. Alterações em atributos químicos de solo cultivado com cana de açúcar e adição de vinhaça. **Pesquisa Agropecuária Tropical,** Goiânia, v. 40, n. 3, p. 341-346, jul./set. 2010.

BENVENUTI, F.A. **Relação de índices espectrais de vegetação com a produtividade da cana de açúcar e atributos edáficos.** 2005. 103 p. Dissertação (Mestrado em Engenharia Agrícola) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2005.

BERNARDES, M.S.; PRELLWITZ, W.P.V.; BRAGA JUNIOR, R.L.C.; SUGUITANI, C.; BEAUCLAIR, E.G.F.; CÂMARA, G.M.S. Equação para estimativa de produtividade dos sucessivos cortes associada ao ambiente de produção e manejo da cultura da cana de açúcar (*Saccharum spp.*). In: CONGRESSO NACIONAL DA STAB, 9., 2008, Maceió. **Anais...** Maceió: STAB, 2008. p. 628-631.

BERKELEY, I.S.N. **What is artificial intelligence?** Disponível em: <http://www.ucs.loisiana.edu/~isb9112/dept/phil341/wisai/WhatisAI.html>. Acesso em: 15 maio 2014.

BLACKBURN, F. **Sugarcane**. New York: Longman, 1984. 414 p.

BOCCA, F.F. **Produtividade de cana de açúcar:** caracterização dos contextos de decisão e utilização de técnicas de mineração de dados para modelagem. 2014. 86 p. Dissertação (Mestrado em Engenharia Agrícola) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2014.

BONESSO, D. **Estimação dos parâmetros do Kernel em um classificador SVM na classificação de imagens hiperespectrais em uma abordagem multiclasse.** 2013. 108 p. Dissertação (Mestrado em Sensoriamento Remoto) – Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

BRAGA, A.P.; LUDERMIR, T.B.; CARVALHO, A.C.P.L.F. **Redes neurais artificiais:** teoria e aplicações. 2. ed. Rio de Janeiro: LTC, 2007. 226 p.

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

CÂMARA, G.M.S. Ecofisiologia da cultura da cana de açúcar. In: _____. **Produção de cana de açúcar.** Piracicaba: FEALQ, 1993. p. 31-64.

CAMARGO, A.P.; SENTELHAS, P.C. Avaliação do desempenho de diferentes métodos de estimativa de evapotranspiração potencial no Estado de são Paulo, Brasil. **Revista Brasileira de Agrometeorologia**, Santa Maria, v. 5, n. 1, p. 89-97, 1997.

CAMILO, C.O; SILVA, J.C. **Mineração de dados:** conceitos, tarefas, métodos e ferramentas: relatório técnico. Goiânia: Universidade Federal de Goiás, Instituto de Informática, 2009. 28 p.

CARDOZO, N.P. **Modelagem da maturação da cana de açúcar em função de variáveis meteorológicas.** 2012. 199 p. Dissertação (Mestrado em Física do Ambiente Agrícola) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2012.

CARDOZO, N.P.; SENTELHAS, P.C. Climatic effects on sugarcane ripening under the influence of cultivars and crop age. **Scientia Agricola**, Piracicaba, v. 70, p. 449-456, 2013.

CARUANA, R.; KARAMPATZIAKIS, N.; YESSENALINA, A. An empirical evaluation of supervised learning in high dimensions. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 25., 2008, Helsinki. **Proceedings…** Helsinki: ACM, 2008. p. 96-103.

CARVALHO, A.P.L.F.; BRAGA, A.P.; LUDERMIR, T.B. Fundamentos de redes neurais artificiais. In: ESCOLA DE COMPUTAÇÃO, 11., 1998, Rio de Janeiro. Rio de Janeiro: UFRJ. 1998. 246 p.

CENTRO DE TECNOLOGIA CANAVIEIRA. **Pragas e doenças da cana de açúcar:** Roguing. 2013. Disponível em: <http://www.ctcanavieira.com.br/downloads/CadernetaPragas2013.pdf>. Acesso em: 13 out. 2015.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0:** step-by-step data mining guide. Illinois: SPSS, 2000. 78 p.

CHINCHULUUN, A.; XANTHOPOULOS, P.; TOMAINO, V.; PARDALOS, P.M. Data Mining techniques in agricultural and environmental sciences. **International Journal of Agricultural and Environmental Information Systems,** Hershey, v. 1, n. 1, p. 26-40, Jan./June 2010. Disponível em: <http://www.academia.edu/1769745/Data_Mining_Techniques_in_Agricultural_and_Environmental_Sciences>. Acesso em: 18 set. 2015.

COLETI, J.T. Técnica cultural de plantio. In: PARANHOS, S.B. (Coord.). **Cana de açúcar:** cultivo e utilização. Campinas: Fundação Cargill, 1987. p. 284-328.

DELGADO ROJAS, J.S. **Modelo agrometeorológico para estimativa dos efeitos de deficiência hídrica na produtividade agro-industrial da cana de açúcar.** 1998. 74 p. Dissertação (Mestrado em Agrometeorologia) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1998.

DI GIROLAMO NETO, C. **Desenvolvimento e avaliação de modelos de alerta para a ferrugem do cafeeiro.** 2013. 155 p. Dissertação (Mestrado em Engenharia Agrícola– Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2013.

DIAS, F.L.F.; ROSSETTO, R. Calagem e adubação da cana de açúcar. In: SEGATO, S.V.; PINTO, A.S.; JENDIROBA, E.; NÓBREGA, J.C.M. de (Org.). **Atualização em produção de cana de açúcar.** Piracicaba: CP2, 2006. p. 107-119.

DOORENBOS, J.; KASSAN, A.H. **Yield response to water.** Rome: FAO, 1979. 193 p. (FAO. Irrigation and Drainage Paper, 33).

DOURADO-NETO, D.; TERUEL, D.A.; REICHARDT, K.; NIELSEN, D.R.; FRIZZONE, J.A.; BACCHI, O.O.S. Principles of crop modeling and simulation: I. Uses of mathematical models in agricultural sciences. **Scientia Agricola**, Piracicaba, v. 55, p. 46-50, 1998.

EMBRAPA. CNPTIA. **Agência de informação da Embrapa:** cana de açúcar. Disponível em: <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_33_711200516717.html>. Acesso em: 15 maio 2014.

EVERINGHAM, Y.; SEXTON, J.; SKOCAJ, D.; BAMBER, G.I. Accurate prediction of sugarcane yield using a random forest algorithm. **Agronomy for Sustainable Development**, Paris, v. 27, p. 1-9, 2016.

FAO. **FAOSTAT:** food and agricultural commodities production / countries by commodity. 2015. Disponível em: <http://faostat3.fao.org/browse/rankings/countries_by_commodity/E>. Acesso em: 13 out. 2015.

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**, Amsterdam, v. 112, p. 149-157, 2009.

FREUD, Y.; SCHAPIRE, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences,** New York, v. 55, n. 1, p. 119-139, 1997.

GALVÃO, N.D.; MARIN, H.F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, São Paulo, v. 22, n. 5, p. 686-690, 2009.

GARCÍA ALONSO, C.R.; TORRES JIMÉNEZ, M.; HERVÁS MARTÍNEZ, C. Income prediction in the agrarian sector using product unit neural networks. **European Journal of Operational Research**, Cordoba, n. 204, p. 355-365, 2009.

GARSIDE, A.L.; BELL, M.J.; ROBOTHAM, B.G.; MAGAREY, R.C.; STIRLING G.R. Managing yield decline in sugarcane cropping systems. **International Sugar Journal**, Glamorgan, v. 107, p. 16-26, 2005.

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**, Amsterdam, v. 95, p. 156-170, 2006.

GOUVÊA, J.R.F. **Mudanças climáticas e a expectativa de seus impactos na cultura da cana de açúcar na região de Piracicaba, SP.** 2008. 98 p. Dissertação (Mestrado em Física do Ambiente Agrícola) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2008.

HAN, J.; KAMBER, M. **Data mining:** concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2011. 744 p.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning:** prediction, inference and data mining. 2nd ed. Berlin: Springer Verlag, 2009. 745 p.

HAYKIN, S. **Neural networks:** a comprehensive foundation. 2nd ed. New Jersey: Prentice Hall, 1999. 842 p.

HOLDEN, G.; MC GUIRE, T. **Irrigation of sugarcane manual.** Indooroopilly: BSES Limited, 1998. 59 p.

HOOGENBOOM, G. Contribution of agrometeorology to the simulation of crop production and its applications. **Agricultural and Forest Meteorology Research**, Griffin, v. 103, p. 137-157, 2000.

INMAN-BAMBER, N.G. A growth model for sugarcane based on a simple carbon balance and the CERES-maize water balance. **South African Journal for Plant and Soil.** Mount Edgecombe, v. 8, p. 93-99, 1991.

_____. Effect of age and season on components of yield of sugarcane in South Africa. **Proceedings of the South African Sugar Technologists Association,** Mount Edgecombe, v. 68, p. 23-27, 1994.

IRELAND, C.R. **Experimental statistics for agriculture and horticulture.** Oxfordshire: CAB International, 2010. 360 p.

JAMES, G.; HASTIE, T.; WITTEN, D.; TIBSHIRANI, R. **An introduction to statistical learning:** with applications in R. London: Springer, 2013. 426 p.

KEATING, B.A.; ROBERTSON, M.J.; MUCHOW, R.C.; HUTH, N.I. Modeling sugarcane production system I. Development and performance of the sugarcane module. **Field Crops Researcher,** Amsterdam, v. 61, n. 3, p. 253-271, 1999.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, Los Angeles, v. 28, n. 5, p. 1-26, 2008.

KUMAR, A.; KANNATHASAN, N. A survey on data mining and pattern recognition techniques for soil data mining. **International Journal of Computer Science Issues,** Chennai, v. 8, n. 3, p. 422-428, 2011.

LAIME, E.M.O.; FERNANDES, P.D.; OLIVEIRA, D. C. S.; FREIRE, E.A. Possibilidades tecnológicas para a destinação da vinhaça: uma revisão. Revista Trópica – Ciências Agrárias e Biológicas, v. 5, n. 3, p. 16-29, 2011.

LAWES, R.A.; LAWN, R.J.; WEGENER, M.K.; BASFORD, K.E. Understanding and managing the late time of ratooning effect on cane yield. **Proceedings of the Australian Society of Sugar Cane Technologists,** Red Hook, v. 24, 2002a.

LAWES, R.A.; Mc DONALD, L.M.; WEGENER, M.K.; BASFORD, K.E; LAWN,R.J. Factors affecting cane yield and commercial cane sugar in the Tully district. **Australian Journal of Experimental Agriculture,** Collingwood, v. 42, p. 473-480, 2002b.

LEME, E.J.A. Uso e tratamento de resíduos agroindustriais no solo. In: CÂMARA, G.M.S. **Produção de cana de açúcar.** Piracicaba: FEALQ, 1993. p. 31-64.

LIMA, C.A.M. **Comitê de máquinas:** uma abordagem unificada empregando máquina de vetores-suporte. 2004. 314 p. Tese (Doutorado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e Computação, Universidade Estadual de Campinas, Campinas, 2004.

LORENA, A.C.; CARVALHO, A.C.P.L.F. **Introdução às máquinas de vetores suporte (support vector machines).** São Carlos: USP, ICMC, 2003. 58 p. (Relatórios Técnicos, 192).

MACCHERONI, W.; MATSUOKA, S. Manejo das principais doenças da cana de açúcar. In: SEGATO, S.V.; PINTO, A.S.; JENDIROBA, E; NÓBREGA, J.C.M. de (Org.). **Atualização em produção de cana de açúcar.** Piracicaba: CP2, 2006. p. 239-256.

MARIN, F.R.; PELLEGRINO, G.Q.; ASSAD, E.D.; PINTO, H.S.; ZULLO JUNIOR, J. Cana-de-açúcar. In: MONTEIRO, J.E.B.A. **Agrometeorologia dos cultivos:** o fator meteorológico na produção agrícola. Brasília: INMET, 2009. p. 111-130.

MARIN, F.R.; JONES, J.W.; ROYCE, F.; SUGUITANI, C; DONZELI, J.L.; PALLONE FILHO, W.J.P.; NASSIF, D.S.P. Parametrization and evaluation of predictions of CSM-CANEGRO for Brazilian sugarcane. **Agronomy Journal**, Madison, v. 103, p. 100-110, 2011.

McDONALD, L. The effect of time of ratooning on sugarcane growth in the Burdekin. **Proceedings of the Australian Society of Sugar Cane Technologists,** Red Hook, v. 28, 2006.

MEIRA, C.A.A. **Processo de descoberta de conhecimento em bases de dados para a análise e o alerta de doenças de culturas agrícolas e sua aplicação na ferrugem do cafeeiro.** 2008. 198 p. Dissertação (Doutorado em Engenharia Agrícola) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2008.

MONTEIRO, L.A. **Modelagem agrometeorológica como base para a definição de ambientes de produção para a cultura da cana de açúcar no estado de SP.** 2011. 121 p. Dissertação (Mestrado em Física do Ambiente Agrícola) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2011.

_____. **Sugarcane yield gap in Brazil:** a crop modelling approach. 2015. 131 p. Thesis (Doctor in Agricultural Systems Engineering) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2015.

MONTEIRO, L.A.; SENTELHAS, P.C. Potential and actual sugarcane yields in Southern Brazil as a function of climate conditions and crop management. **Sugar Tech,** New Delhi, v. 16, n. 3, p. 264-276, 2014. DOI 10.1007/s12355-013-0275-0.

MOURA, A.M. **Geoprocessamento na gestão e planejamento urbano**. Belo Horizonte: Ed. da Autora, 2003. 294 p.

MUCHERINO, A.; PAPAJORGJI, P.J.; PARDALOS, P.M. **Data mining in agriculture.** Gainesville: Springer, 2009. 274 p.

MÜLLER, K.R., MIKA, S., RÄTSCH, G., TSUDA, K., SCHÖLKOPF, B. An introduction to kernel-based learning algorithms. **IEEE Transactions on Neural Networks**, New York, v. 12, n. 2, p. 181–201, 2001.

NASSIF, D.S.P. **Parametrização e avaliação do modelo CSM-CANEGRO para variedades brasileiras de cana de açúcar.** 2010. 92 p. Dissertação (Mestrado em Física do Ambiente Agrícola) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2010.

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. **Pesquisa Agropecuária Brasileira,** Brasília, v. 47, n. 3, p. 311-318, 2010.

NEVES, M.F.; TROMBIN, V.G.; CONSOLI, M. **Mapeamento e quantificação do setor sucroenergético 2008.** Ribeirão Preto: Markestrat; Fundace; UNICA, 2009. 34 p.

NONATO, R.T. **Aplicação de mineração de dados na identificação de áreas cultivadas com cana de açúcar em imagens de sensoriamento remoto no Estado de São Paulo.** 2010. 128 p. Dissertação (Mestrado em Engenharia Agrícola) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2010.

O´LEARY, G.J. A review of three sugarcane simulation models with respect to their prediction of sucrose yield. **Field Crops Research,** Amsterdam, v. 68, p. 97-111, 2000.

OMETTO, J.C. **Equação para estimativa de evapotranspiração potencial, sua aplicação no cálculo das necessidades hídricas e do rendimento agro-industrial da cana de açúcar na região de Piracicaba (SP).** 1974. 129 p. Dissertação (Livre-Docência) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1974.

ORLANDO FILHO, J.; MACEDO, N.; TOKESHI, H. Seja o doutor de seu canavial. **Arquivo do Agrônomo POTAFOS**, Piracicaba, n. 6, p. 1-16, 1994.

ORLANDO FILHO, J.; SILVA, G.M.A.; LEME, E.J.A. Utilização agrícola dos resíduos da agroindústria canavieira. In: INSTITUTO DO AÇÚCAR E DO ÁLCOOL. Programa Nacional de Melhoramento da cana-de-açúcar. **Nutrição e adubação da cana de açúcar no Brasil.** Piracicaba, 1983. p. 229-264.

ORNELLA, L.; TAPIA, E. Supervised machine learning and heterotic classification of maize (*Zea mays L.*) using molecular marker data. **Computers and Electronics in Agriculture**, Rosario, v. 74, p. 250-257, 2010.

PALMA NETO, L.G.; NICOLETTI, M.C. **Introdução às redes neurais construtivas.** São Carlos: Edufscar, 2005. 192 p.

PEREIRA, A.R.; ANGELOCCI, L.R.; SENTELHAS, P.C. **Agrometeorologia:** fundamentos e aplicações práticas. Guaíba: Ed. Agropecuária, 2002. 478 p.

PICOLI, M.C.A. **Estimativa da produtividade da cana de açúcar utilizando agregados de redes neurais artificiais: estudo de caso Usina Catanduva.** 2007. 90 p. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2007.

PRADO, H. **Pedologia Fácil – Aplicações na agricultura.** Piracicaba, 2008. 148 p.

PROCÓPIO, S.O.; SILVA, A.A.; VARGAS, L.; FERREIRA, F.A. **Manejo de plantas daninhas na cultura da cana de açúcar.** Viçosa: Universidade Federal de Viçosa, 2003. 150 p.

RAMBURAN, S.; ZHOU, M.; LABUSCHAGNE, M. Interpretation of genotype x environment interactions of sugarcane: identifying significant environmental factors. **Field Crops Research**, Amsterdam, v. 124, p. 392-399, 2011.

RAMBURAN, S.; WETTERGREEN, T.; BERRY, S.D.; SHONGWE, B. Genetic, environmental and management contributions to ratoon decline in sugarcane. **Field Crops Research**, Amsterdam, v. 146, p. 105-112, 2013.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: LIU, L.; ÖZSU, M.T. (Ed.). **Encyclopedia of database systems.** New York: Springer US, 2009. p. 532-538.

RODRIGUES, J.D. **Fisiologia da cana de açúcar**. Botucatu: UNESP, 1995. 99 p.

ROLIM, J.C. Instalação da cultura e sistemas de produção. In: CÂMARA, G.M.S. **Produção de cana de açúcar.** Piracicaba: FEALQ, 1993. p. 31-64.

SANTOS, D.L. **Zoneamento da favorabilidade climática para a ocorrência da ferrugem alaranjada da cana de açúcar nas principais regiões produtoras do Brasil e da Austrália.** 2013. 100 p. Dissertação (Mestrado Engenharia de Sistemas Agrícolas) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2013.

SAYAD, S. **Support vector machines.** Toronto: University of Toronto, 2010. Disponível em: <http://chem-eng.utoronto.ca/~datamining/Presentations/SVM.pdf>. Acesso em: 10 abr. 2016.

SCARPARI, M.S. **Modelos para a previsão da produtividade da cana de açúcar *(Saccharum spp.)* através de parâmetros climáticos.** 2002. 79 p. Dissertação (Mestrado em Fitotecnia) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2002.

SHAPIRE, R.E. The strength of weak learnability. **Machine Learning,** Boston, n. 5, p. 197-227, 1990.

SILVA, T.G.F.; MOURA, M.S.B.; ZOLNIER, S.; CARMO, J.F.A.; SOUZA, L.S.B. Biometria da parte aérea da cana soca irrigada no submédio do Vale do São Francisco. **Revista Ciência Agronômica,** Fortaleza, v. 43, n. 3, p. 500-509, jul./set. 2012.

SINGELS, A.; JONES, M.; VANDERBERG, M. **DSSAT v.4.5 CSM-CANEGRO:** sugarcane plant module; scientific documentation. Mount Edgecombe: South African Sugarcane Research Institute, International Consortium for Sugarcane Modeling, 2008. 34 p.

SPINDEL, B.R. **Análise e desenvolvimento de sistema de estimação de modelos da classe STAR-Tree.** 2008. 94 p. Dissertação (Mestrado em Engenharia Elétrica) - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

TACONELI, C.A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia.** 2011. 99 p. Dissertação (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2011.

TAVARES SOUZA, P.C; WILHELM, V.E. Uma revisão bibliográfica dos modelos agrometeorológicos. **Tuiuti: Ciência e Cultura,** Curitiba, n. 42, p. 141-150, 2009.

THORNLEY, J.H.M.; FRANCE, J. **Mathematical models in agriculture:** quantitative methods for the plant, animal and ecological sciences. Wallingford: CABI, 2007. 924 p.

THORNTHWAITE, C.W.; MATHER, J.R. **The water balance.** New Jersey: Drexel Institute of Technology, 1955. 104 p. (Publications in Climatology, v. 8, n. 1).

UNIÃO DA INDÚSTRIA DE CANA DE AÇÚCAR. **Moagem de cana de açúcar e produção de açúcar e etanol:** safra 2012/2013. Disponível em: <http://www.unicadata.com.br>. Acesso em: 07 mar. 2014.

VIANNA, M.S. **Risco climático para a cultura da cana de açúcar e estratégias de manejo de irrigação complementar para a sua minimização em diferentes regiões brasileiras.** 2014. 177 p. Dissertação (Mestrado em Engenharia de Sistemas Agrícolas) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2014.

VIANNA, M.S.; SENTELHAS, P.C. Performance of DSSAT CSM-CANEGRO under operational conditions and its use in determining the ‘saving irrigation’ impact on sugarcane crop. **Sugar Tech,** New Delhi, v. 18, n. 1, p. 75-86, 2015.

VIBHA, L.; HARSHAVARDHAN, G.M.; PRASHANTH, S.J.; DEEPA SHENOY, P; VENUGOPAL, K.R.; PATNAIK, L.M. A hybrid clustering and classification technique for soil data mining. In: IET-UK INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY IN ELECTRICAL SCIENCES, 2007, Chennai. **Proceedings…** Chennai: ICTES, 2007. p. 1090-1095.

VIEIRA, F.D. **Modelos baseados em técnicas de mineração de dados para suporte à certificação racial de ovinos.** 2014. 88 p. Dissertação (Mestrado em Engenharia Agrícola) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2014.

WILLMOTT, C.J. On the validation of models. **Physical Geography,** Norwich, v. 2, p. 184-194, 1981.

WITTEN, I.H.; FRANK, E.; HALL, M.A. **Data mining:** practical machine learning tools and techniques. San Francisco: Morgan Kaufmann Publ., 2011. 630 p.

APÊNDICE

Apêndice

Variáveis independentes relacionadas no diagrama de correlações (Figura 10).

NCORTE	Número de cortes: 1, 2, 3, 4, 5
IDADE	Idade do canavial em número de meses entre os cortes
ESPAÇAMENTO	Espaçamento de plantio adotado
TEMPMEDIA-F1	Temperatura media dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
TEMPMEDIA-F2	Temperatura media dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
TEMPMEDIA-F3	Temperatura media dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
TEMPMEDIA-F4	Temperatura media dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
TEMPMEDIA-F5	Temperatura media dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
TEMPMEDIA_T	Temperatura media de todas fases de crescimento (vide Tabelas 2 e 3).
URMEDIA-F1	Umidade relativa do ar (%) media dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
URMEDIA-F2	Umidade relativa do ar (%) media dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
URMEDIA-F3	Umidade relativa do ar (%) media dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
URMEDIA-F4	Umidade relativa do ar media dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
URMEDIA-F5	Umidade relativa do ar (%) media dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
URMEDIA_T	Umidade relativa do ar (%) media de todas as fases de crescimento (vide Tabelas 2 e 3).
RADGLOBAL-F1	Irradiância solar extraterrestre (MJ m^{-2} dia^{-1}) acumulada dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
RADGLOBAL-F2	Irradiância solar extraterrestre (MJ m^{-2} dia^{-1}) acumulada dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
RADGLOBAL-F3	Irradiância solar extraterrestre (MJ m^{-2} dia^{-1}) acumulada dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
RADGLOBAL-F4	Irradiância solar extraterrestre (MJ m^{-2} dia^{-1}) acumulada dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).

RADGLOBAL-F5	Irradiância solar extraterrestre (MJ m^{-2} dia^{-1}) acumulada dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
RADGLOBAL_T	Irradiância solar extraterrestre (MJ m^{-2} dia^{-1}) acumulada de todas as fases de crescimento (vide Tabelas 2 e 3).
VELVENTO-F1	Velocidade do vento (m s^{-1}) média dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
VELVENTO-F2	Velocidade do vento (m s^{-1}) média dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
VELVENTO-F3	Velocidade do vento (m s^{-1}) média dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
VELVENTO-F4	Velocidade do vento (m s^{-1}) média acumulada dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
VELVENTO-F5	Velocidade do vento (m s^{-1}) média acumulada dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
VELVENTO_T	Velocidade do vento (m s^{-1}) média de todas as fases de crescimento (vide Tabelas 2 e 3).
INSOLAÇÃO-F1	Insolação (h dia^{-1}) media dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
INSOLAÇÃO-F2	Insolação (h dia^{-1}) media dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
INSOLAÇÃO-F3	Insolação (h dia^{-1}) media dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
INSOLAÇÃO-F4	Insolação (h dia^{-1}) media dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
INSOLAÇÃO-F5	Insolação (h dia^{-1}) media dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
INSOLAÇÃO_T	Insolação (h dia^{-1}) media de todas as fases de crescimento (vide Tabelas 2 e 3).
CHUVA-F1	Precipitação (mm) acumulada dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
CHUVA-F2	Precipitação (mm) acumulada dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
CHUVA-F3	Precipitação (mm) acumulada dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
CHUVA-F4	Precipitação (mm) acumulada dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
CHUVA-F5	Precipitação (mm) acumulada dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
CHUVA_T	Precipitação (mm) acumulada de todas as fases de crescimento (vide Tabelas 2 e 3).
ETP-F1	Evapotranspiração potencial (mm) acumulada dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).

ETP-F2	Evapotranspiração potencial (mm) acumulada dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
ETP-F3	Evapotranspiração potencial (mm) acumulada dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
ETP-F4	Evapotranspiração potencial (mm) acumulada dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
ETP-F5	Evapotranspiração potencial (mm) acumulada dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
ETP_T	Evapotranspiração potencial (mm) acumulada de todas as fases de crescimento (vide Tabelas 2 e 3).
ARM-F1	Armazenamento de água disponível no solo (mm) médio dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
ARM-F2	Armazenamento de água disponível no solo (mm) médio dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
ARM-F3	Armazenamento de água disponível no solo (mm) médio dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
ARM-F4	Armazenamento de água disponível no solo (mm) médio dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
ARM-F5	Armazenamento de água disponível no solo (mm) médio dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
ARM_T	Armazenamento de água no solo (mm) médio de todas as fases de crescimento (vide Tabelas 2 e 3).
ETR-F1	Evapotranspiração real (MJ m^{-2} dia^{-1}) acumulada de todos os decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
ETR-F2	Evapotranspiração real (MJ m^{-2} dia^{-1}) acumulada de todos os decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
ETR-F3	Evapotranspiração real (MJ m^{-2} dia^{-1}) acumulada de todos os decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
ETR-F4	Evapotranspiração real (MJ m^{-2} dia^{-1}) acumulada de todos os decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
ETR-F5	Evapotranspiração real (MJ m^{-2} dia^{-1}) acumulada de todos os decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
ETR_T	Evapotranspiração real (MJ m^{-2} dia^{-1}) acumulada de todas

	as fases de crescimento (vide Tabelas 2 e 3).
DEF-F1	Déficit hídrico (mm) acumulado dos decêndios da fase de crescimento 1 (vide Tabelas 2 e 3).
DEF-F2	Déficit hídrico (mm) acumulado dos decêndios da fase de crescimento 2 (vide Tabelas 2 e 3).
DEF-F3	Déficit hídrico (mm) acumulado dos decêndios da fase de crescimento 3 (vide Tabelas 2 e 3).
DEF-F4	Déficit hídrico (mm) acumulado dos decêndios da fase de crescimento 4 (vide Tabelas 2 e 3).
DEF-F5	Déficit hídrico (mm) acumulado dos decêndios da fase de crescimento 5 (vide Tabelas 2 e 3).
DEF_T	Déficit hídrico (mm) acumulado de todas as fases de crescimento (vide Tabelas 2 e 3).
ETR_ETC-F1	Relação entre evapotranspiração real e evapotranspiração da cultura, como medida do déficit hídrico relativo para a fase de crescimento 1 (vide Tabelas 2 e 3).
ETR_ETC-F2	Relação entre evapotranspiração real e evapotranspiração da cultura, como medida do déficit hídrico relativo para a fase de crescimento 2 (vide Tabelas 2 e 3).
ETR_ETC-F3	Relação entre evapotranspiração real e evapotranspiração da cultura, como medida do déficit hídrico relativo para a fase de crescimento 3 (vide Tabelas 2 e 3).
ETR_ETC-F4	Relação entre evapotranspiração real e evapotranspiração da cultura, como medida do déficit hídrico relativo para a fase de crescimento 4 (vide Tabelas 2 e 3).
ETR_ETC-F5	Relação entre evapotranspiração real e evapotranspiração da cultura, como medida do déficit hídrico relativo para a fase de crescimento 5 (vide Tabelas 2 e 3).
ETR_ETC_T	Relação entre evapotranspiração real e evapotranspiração da cultura, como medida do déficit hídrico relativo para todas as fases de crescimento (vide Tabelas 2 e 3).
CALCARIO	Aplicação de calcário (% da área do bloco em questão)
FOSFATO	Aplicação de fosfato (% da área do bloco em questão)
GESSO	Aplicação de gesso (% da área do bloco em questão)
INSETICIDA	Aplicação de inseticida (% da área do bloco em questão)
VINHAÇA	Aplicação de vinhaça (% da área do bloco em questão)
consolidado$TCHREAL	Produtividade observada da cana de açúcar (TCH)