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

Caracterização do risco de infestação da broca da cana-de-açúcar no estado de São Paulo, com base em variáveis meteorológicas e técnicas de mineração de dados

Guilherme Antonio Buonadio

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

Piracicaba
2021

Guilherme Antonio Buonadio
Engenheiro Agrônomo

Caracterização do risco de infestação da broca da cana-de-açúcar no estado de São Paulo, com base em variáveis meteorológicas e 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
2021

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

Buonadio, Guilherme Antonio

Caracterização do risco de infestação da broca da cana-de-açúcar no estado de São Paulo, com base em variáveis meteorológicas e técnicas de mineração de dados / Guilherme Antonio Buonadio. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2021.

88 p.

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

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

DEDICATÓRIA

AOS MEUS PAIS E IRMÃ, MINHA QUERIDA ESPOSA E A MINHA LINDA FILHA BEATRIZ PELO APOIO INCONDICIONAL NESTA IMPORTANTE ETAPA DA MINHA VIDA.

AGRADECIMENTOS

À Deus

Por me dar sabedoria e renovar minha esperança a cada amanhecer. Por me dar força para alcançar meus objetivos e por me iluminar a cada passo nessa caminhada. Agradeço por sempre guiar meus passos e fornecer os caminhos para atingir a vitória. Sou muito grato por tudo que conquistei até hoje.

Aos meus pais Antonio Carlos e Maria Carlota

Por sempre me incentivarem, desde criança, o caminho do estudo e da dedicação e principalmente o fato de sempre continuar em evolução constante para alçar vôos mais altos, principalmente quando decidi fazer o mestrado. Por apoiarem minhas decisões e por me consolarem nos momentos mais difíceis. Pelo amor gratuito que não pode ser mensurado.

À minha querida irmã Maira e família

Por estarem ao meu lado me incentivando a buscar meus sonhos. Sempre orientando e dando conselhos, principalmente nas orientações para a escolha do curso de Engenharia Agronônomica.

A minha querida esposa Pamela

Por acreditar no meu esforço e na minha capacidade de passar em uma Universidade Pública. Por entender a pesada rotina de estudos e trabalho e sempre me apoiar. Por acreditar na busca constante do conhecimento. Por ser amiga, esposa e dar o maior presente de minha vida que é nossa querida filha Beatriz.

A minha querida filha Beatriz

Que veio ao mundo em um momento de muitas incertezas e desafios em meio a uma pandemia, tornando meus dias mais leves, felizes e muito mais completos. Presente de Deus em nossas vidas que nos propicia viver de uma forma mais intensa e aproveitar cada momento.

À Escola Superior de Agricultura "Luiz de Queiroz" – ESALQ/USP

Por me receber em 2008 na graduação em Engenharia Agronômica, e em 2018 no Mestrado em Engenharia de Sistemas Agrícolas. Além de todo o auxílio e respaldo em toda a minha trajetória profissional, agradeço muito a esta Escola por tudo que me forneceu de conhecimento, amizades e experiências.

Ao meu orientador Prof. Dr. Paulo Cesar Sentelhas (in memorian)

Por ter sido um brilhante professor de Meteorologia Agrícola na graduação e por ter me aceitado como seu aluno no mestrado mesmo sabendo de minha rotina de trabalho. Pelos ensinamentos durante esses anos, pelo apoio à pesquisa e ao projeto e por ter acreditado na possibilidade de associarmos academia com vida corporativa. Você foi um grande amigo que nunca esquecerei.

Ao Prof. Dr. Allan Panosso

Pela contribuição sem igual com o andamento do trabalho, principalmente pela disponibilidade de tempo e paciência. Sem a sua ajuda em todas as etapas e principalmente nas análises mais complexas este trabalho não seria possível e não teria avançado com tantos resultados positivos e contribuições ao setor.

Aos colegas do time agrícola da Usina São Martinho

Por acolherem a iniciativa do mestrado e propiciar meu crescimento profissional. Pelo acompanhamento ao longo da trajetória e ajuda na confecção e fornecimento dos dados, orientações e constantes aprendizados que contribuíram muito para o retorno desta pesquisa. Agradeço de coração a gerência agrícola e a todo time agrícola.

BIOGRAFIA

Engenheiro Agrônomo formado pela ESALQ\USP em 2012.

Pós-Graduação em Proteção de Plantas pela Universidade Federal de Viçosa – UFV em 2015.

MBA em Agronegócios pela ESALQ\USP em 2019.

Profissional com 10 anos de experiência no setor sucroenergético em empresas de grande porte, com expertise em pragas e doenças, planejamento agrícola (plantio, colheita e tratos culturais), desenvolvimento agronômico, experimentação agrícola, nutrição e adubação, aviação agrícola, produção de mudas pré-brotadas (mpb) qualidade agrícola, auditorias agrícolas e novas tecnologias.

EPÍGRAFE

“Sonhos determinam o que você quer.

Ação determina o que você conquista”

Aldo Novak

SUMÁRIO

RESUMO 10
ABSTRACT 11
1. INTRODUÇÃO 13
2. REVISÃO BIBLIOGRÁFICA 17
2.1. Broca da cana-de-açúcar 17
2.1.1. Importância da broca da cana-de-açúcar 17
2.1.2. Ciclo da praga 18
2.1.3. Impactos causados pela broca na cultura da cana-de-açúcar 19
2.1.4. Estratégias de controle da broca-da-cana 19
2.2. Fatores agrometeorológicos que afetam a produtividade da cana-de-açúcar 20
2.3. Modelagem da dinâmica populacional de pragas 21
2.3.1. Uso do método dos graus-dias para descrever o desenvolvimento de insetos 23
2.4. Mineração de Dados 23
2.4.1. Grupo das Árvores de Decisão 25
2.4.2. Support Vector Machine (SVM) 29
2.5. O uso de Mineração de Dados na agricultura 33
3. MATERIAL E MÉTODOS 35
3.1. Entendimento dos dados 35
3.2. Preparações dos dados 36
3.3. Correlação de Pearson 36
3.4. Análise de Componentes Principais 36
3.5. Modelos de estimação do índice de infestação da broca-da-cana 37
3.6. Avaliação dos modelos de estimação do índice de infestação da broca-da-cana 37
3.7. Determinação e mapeamento do risco de infestação da broca-da-cana 38
3.8. Distribuição do conhecimento 40
4. RESULTADOS 41
4.1. Análise da Componentes Principais 43
4.2. Análise da intensidade da infestação de broca-da-cana por variedade 45
4.3. Seleção e descrição das variáveis empregadas na mineração de dados 46
4.3.1. Intensidade da infestação final da broca da cana-de-açúcar 46

4.3.2. Temperatura Máxima 47
4.3.3. Temperatura Mínima 48
4.3.4. Temperatura média 49
4.3.5. Chuva 50
4.3.6. Umidade Relativa 51
4.3.7. Evapotranspiração Potencial 52
4.3.8. Número de dias com temperatura mínima ≥ 13 ºC 53
4.3.9. Número de dias com Temperatura máxima ≤ 35 ºC 54
4.3.10. Radiação solar 55
4.3.11. Déficit Hídrico 56
4.3.12. Excedente Hídrico 57
4.4. Determinação dos principais fatores determinantes do índice de infestação final utilizando-se diferentes técnicas de mineração de dados 58
4.4.1. *Random Forest* ou Floresta Aleatória 58
4.4.2. Boosting 59
4.4.3. Support Vector Machine 60
4.5. Modelos para a estimação do índice de infestação final de broca por meio de técnicas de mineração de dados 62
4.6. Determinação e mapeamento do risco de infestação da broca da cana-de-açúcar no estado de São Paulo 65
4.6.1 Variabilidade interanual do índice de infestação de broca da cana-de-açúcar 65
4.6.2 Variação espacial do risco de infestação de broca da cana-de-açúcar no estado de São Paulo 71
5.CONCLUSÕES 74
REFERÊNCIAS 77

RESUMO

Caracterização do risco de infestação da broca da cana-de-açúcar no estado de São Paulo, com base em variáveis meteorológicas e técnicas de mineração de dados

A broca da cana-de-açúcar (*Diatraea saccharalis*) é um inseto praga de grande importância, causando impactos expressivos nos canaviais se não for adequadamente controlada. A dinâmica populacional da broca da cana-de-açúcar é influenciada, basicamente, pelas condições ambientais, mais especificamente pela temperatura e pela chuva. sendo possível determinar a variação da população dessa praga por meio de modelos matemáticos com base em variáveis meteorológicas. Dessa forma, o objetivo deste estudo foi desenvolver modelos para estimação da dinâmica temporal e espacial de infestação da broca da cana-de-açúcar com base em variáveis meteorológicas, de modo a subsidiar estratégias de controle desta praga. Para tanto, foram utilizados dados de índice de infestação final de broca em oito safras (2012-2013 a 2019-2020), compreendendo o período de janeiro a dezembro, de unidades produtoras de cana-de-açúcar da região de Jaboticabal, SP, que compreende as seguintes localidades: Barrinha; Bueno de Andrade; Dumont; Guariba; Guatapará; Luis Antônio; Monte Alto; Motuca e Pradópolis, considerando-se os dados ao nível de gleba e em áreas com a ausência de controle químico e mínima intervenção com controle biológico. Além do índice de infestação final da broca, também foram utilizadas no estudo as seguintes variáveis meteorológicas: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com temperatura máxima ≤ 35°C (ndc35); número de dias com temperatura mínima ≥ 13°C (ndc13); umidade relativa (ur); precipitação (chuva); radiação solar (Rad); evapotranspiração potencial (etp); assim como variáveis do balanço hídrico climatológico de Thornthwaite e Mather (1955), para uma capacidade de água disponível de 125 mm (déficit hídrico (def_125), excedente hídrico (exc_125)). Os dados de infestação de broca e meteorológicos, compondo um banco de 2876 dados, foram analisados por meio de três técnicas de mineração de dados e aprendizado de máquinas: *Random Forest*; *Boosting* e *Support Vector Machine*, de modo a se obter as variáveis de maior influência sobre a ocorrência e proliferação da broca da cana-de-açúcar e, a partir delas, se confeccionar modelos de estimação da infestação da praga com base nos dados meteorológicos. Os modelos resultantes foram testados com dados independentes, utilizando-se o coeficiente de correlação (r), o índice de Willmott (D), o índice de confiança de Camargo (C), o erro médio (EM), o erro absoluto médio (EAM) e a raiz quadrada do erro médio (RMSE). Os modelos com maior grau de assertividade no treinamento foram o *Boosting* e o *Random Forest*, tendo o *Support Vector Machine* apresentado o pior desempenho. Resultado similar foi obtido na fase de teste com dados independentes. As variáveis de maior importância nos modelos de melhor desempenho foram chuva, umidade relativa e excedente hídrico. Para a confecção do mapa de risco para a ocorrência da broca da cana no Estado de São Paulo, empregou-se o modelo gerado pelo *Random Forest* (R^2 de 0,749, r de 0,865 e RMSE(%) de 0,536) e considerando-se 25 locais do estado de São Paulo para gerar o modelo linear para a espacialização dos dados. O mapa foi validado com dados de outros dez locais. Observou-se para o Estado de São Paulo um risco médio para a infestação da broca da cana-de-açúcar na região centro-oeste, ao passo que no extremo noroeste e nas regiões norte e centro-sul do estado o risco é alto. No sul do estado, assim como na faixa leste o risco é muito alto, contrastando com as regiões do Vale e do Pontal do Paranapanema onde o risco de infestação oscila entre baixo e muito baixo. A modelagem empregada neste estudo, apesar de ser exploratória e realizada com banco de dados pré-existente, possibilitou obter resultados que permitem contribuir para o manejo racional da broca da cana-de-açúcar no estado de São Paulo.

Palavras-chave: Modelos de aprendizado de máquina, Random forest, Boosting, Support vector machine

ABSTRACT

Characterization of the cane borer infestation risk in sugarcane fields of the state of São Paulo, Brazil, based on weather variables and data mining techniques

The cane borer is a insect pest influenced by environmental conditions and it is possible to determine its population dynamics through mathematical models based on meteorological variables. Thus, the objective of this study was to develop mathematical models to estimate the temporal and spatial dynamics of the cane borer based on meteorological variables, in order to support strategies to control this pest. For that, data of cane borer final infestation index of eight growing seasons (2012-2013, 2013-2014, 2014-2015, 2015-2016,2016-2017,2017-2018,2018-2019, and 2019-2020) were used, covering the period from January to December in the Jaboticabal region, which comprises nine locations: Barrinha; Bueno de Andrade; Dumont; Guariba; Guatapará; Luis Antônio; Monte Alto; Motuca; and Pradópolis, all located in the interior of the state of São Paulo, Brazil. The database was at farm level and in all fields assessed the cane did not receive chemical control and they had a minimum intervention with biological control. In addition to the cane borer infestation data, the following weathervariables were used in this study: maximum temperature (tmax); minimum temperature (tmin); average temperature (tmed); number of days with maximum temperature ≤ 35 °C (ndc35); number of days with minimum temperature ≥ 13 °C (ndc13); relative humidity (ur); precipitation (chuva); solar radiation (Rad); potential evapotranspiration (etp); and variables of the climatological water balance of Thornthwaite and Mather (1955), which were water deficit (def_125) and water surplus (exc_125). Infestation index and weather variables, totaling 2876 data, were analized through three techniques of data mining anda machine learning: *Random Forest, Boosting and Support Vector Machine*, for generating cane borer infestation estimation models. The resulting models were tested with independent data, using following indices and errors: correlation coefficient (r); Willmott index (D); Camargo confidence index (C); mean absolute error (MAE); and root mean square error (RMSE). The models with the highest degree of assertiveness in the training phase were Boosting and Random Forest, with Support Vector Machine presenting the worst performance. In the testing phase, the results were similar. The most important variables for the best performing models were rainfall, relative humidity, and water surplus. To make the cane borer infestation risk map for the State of São Paulo, the model generated by Random Forest was used (R^2 de 0,749, r de 0,865 e RMSE(%) de 0,536), and 25 locations were used to adjust a linear model for data spatialization, which was validated with data from other ten locations. The risk map allowed to observe that cane borer infestation has a medium risk in the midwest region of the state, whereas in the extreme nothwest and in the north and center-south of the state the risk is high, on the contrary of Vale and Ponta of Paranapanema Valley where the risk ranges from low to very low. The modeling approach used in the present study, depite being exploratory and carried out with a pre-existing database, allowed to obtain results that can contribute for a more rational cane borer control in the state of São Paulo.

Keywords: Machine learning models, Random forest, Boosting, Support vector machine

1. INTRODUÇÃO

A cana-de-açúcar é uma planta da família *Poaceae*, visto a sua classificação botânica e pertence ao gênero *Saccharum*. Hoje no Brasil temos seis espécies, o que origina a grande variabilidade genética dos híbridos comerciais cultivados no Brasil. Essas espécies são: *S. officinarum*; *S. robustum; S. spontaneum; S. barberi; S. sinensi; e S. edule* (MOZAMBANI, 2006). A cana-de-açúcar é uma planta C4, como a maioria das poáceas (gramíneas), assim chamada por formar compostos orgânicos com quatro carbonos durante a fotossíntese, além de apresentar alta taxa fotossintética e de eficiência na utilização e resgate de CO_2 (gás carbônico) da atmosfera (SEGATO et al., 2006).

Segundo dados da CONAB (2020), a produção de cana-de-açúcar no Brasil realizada na safra 2020/21, foi de 665,11 milhões de toneladas, apresentando incremento de 3,5 % em relação à safra anterior que foi de 642,72 milhões de toneladas. A área colhida foi estimada em 8,60 milhões de hectares, aumento de 1,9% se comparada com a safra 2019/20. Já para a região Centro-Sul a produção foi de 610,65 milhões de toneladas na safra 2020/21, em uma área de 7,70 milhões de hectares, incremento de 2,1% se comparada com a safra 2019/20. Isto torna o Brasil o maior produtor mundial de cana-de-açúcar, com pouco mais de um terço da área cultivada com cana-de-açúcar no mundo em solo brasileiro. Para obter melhores retornos econômicos no ciclo produtivo, cultiva-se a cana-de-açúcar em uma extensa diversidade de clima e solos, além do uso de diferentes variedades adaptada ao melhor ambiente de produção (UNICA, 2012).

A cana-de-açúcar é uma planta semi-perene, com seu ciclo de cultivo se estendendo por aproximadamente cinco anos (LEITE et al., 2008). Tal classificação ocorre já que após a operação de plantio a cultura permanece, em média, 5 anos (cortes) no campo. A "cana planta" ou cana de primeiro corte, é aquela originada de uma muda, que pode ser o próprio tolete ou mudas pré-brotadas (mpb). Após o plantio e realizado a primeira colheita na cultura (primeiro corte), o canavial rebrota e normalmente pode dar em média cinco cortes, sendo a cana nesse processo de sucessivas colheitas e rebrotas denominada como cana soca. A Figura 1 apresenta o ciclo fenológico da cultura da cana-de-açúcar, segundo Gascho e Shih (1983), a qual se aplica tanto à cana planta como à cana soca.

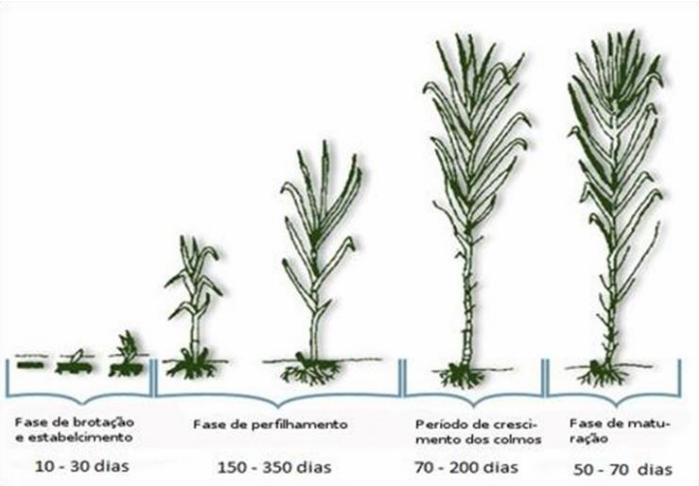


Figura 1. Estádios fenológicos da cana-de-açúcar e suas respectivas durações, em dias, do ciclo.

Fonte: Fonte: Gascho & Shih (1983).

De acordo com Landell et al. (2003a), o início das pesquisas com a cana-de-açúcar associou-se ao surgimento de algumas doenças e a expansão para áreas de cerrado em 1975, com o advento do Programa Nacional do Álcool - PROÁLCOOL, criado pelo governo federal (Decreto 76593 de 14/11/1975). Com a criação do PROÁLCOOL iniciou-se um novo ciclo de pesquisa que suportou a expansão da cultura no país. Em pouco tempo, a área plantada com a cana-de-açúcar triplicou, invadindo áreas consideradas menos aptas principalmente nas regiões de cerrados. Para enfrentar os novos desafios advindos dessa expansão iniciaram-se os programas de melhoramento genético do Instituto Agronômico de Campinas (IAC), Cooperativa de Produtores de Cana-de-Açúcar, Açúcar e Álcool do Estado de São Paulo (COPERSUCAR) e Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA) visando à obtenção de novas variedades, que surgiram no Brasil no começo do século passado. Atualmente, o programa de melhoramento da COPERSUCAR foi substituído pelo programa do Centro de Tecnologia Canavieira (CTC).

Sabe-se que o tempo de vida útil de uma variedade de cana-de-açúcar é limitado seja por degenerescência ou pela necessidade de alavancar a produtividade, dessa forma os lançamentos dos centros de pesquisa e melhoramento genético são constantes. King; Mungomery; Hugues (1965) atribuem que as principais causas da degenerescência são referentes à queda de fertilidade do solo e ao efeito acumulativo de danos provocados por doenças e pragas, já que a variedade não é capaz de sofrer alterações de potencial produtivo se não houver influências ambientais e dos sistemas de produção.

De acordo com Sanguino (1988), no que se refere às doenças provocadas por fungos, bactérias, vírus e micoplasma, foram diagnosticadas no Brasil cerca de 40 dentre as 177 relacionadas em cultivo de cana-de-açúcar a nível mundial. Historicamente, são quatro as doenças mais importantes para a cultura da cana-de-açúcar, sendo elas: carvão; raquitismo das soqueiras; escaldaduras das folhas; e o mosaico, sendo todas elas basicamente controladas por meio do melhoramento genético de variedades (SANTOS, 2003). Por outro lado, as pragas (insetos e nematóides) e as plantas daninhas são os principais fatores limitantes da produção da cana-de-açúcar. Neste caso, faz-se necessário manejos com produtos químicos, que caso não sejam utilizados de forma correta poderão causar problemas ambientais, elevando o custo da produção.

Uma excelente estratégia para o controle de pragas na cultura, é a utilização de agentes de controle, principalmente, contra pragas de solo, sendo esta estratégia uma forma de manejo integrado (MIP), onde fariam parte de um conjunto de práticas, tais como: o controle biológico; a rotação de cultura; e o constante monitoramento de pragas. Em suma maioria, é utilizado o controle químico, fato este que acarreta uma série de impactos negativos à cultura e ao ambiente (MACHADO; HABIB, 2009).

De acordo com Gallo et al. (2002), os danos causados pelos insetos às plantas são variáveis, podendo ser observados em todos os órgãos vegetais. Dependendo da espécie e da densidade populacional da praga, do estádio de desenvolvimento, da estrutura vegetal atacada e da duração do ataque, os danos causados podem levar a prejuízos tanto quantitativos como qualitativos. Tais danos variam também em função das variedades utilizadas, das técnicas agronômicas empregadas e das características socioeconômicas da região produtora.

No caso da cultura da cana-de-açúcar, a produtividade é afetada por diversos fatores durante o ciclo, entre os quais as condições meteorológicas e de manejo são as principais. O maior fator de estresse ambiental (abiótico) para a cultura da cana-de-açúcar é a baixa disponibilidade de água no solo, a qual desencadeia o processo de déficit hídrico, que tem relação inversa com a produtividade (SANTOS; CARLESSO, 1998; MONTEIRO; SENTELHAS (2016). Com relação aos estresses bióticos, a ocorrência de pragas, dentre elas a broca (*Diatraea saccharallis*), é o de

maior preocupação entre os produtores de cana-de-açúcar, já que essa praga causa sérios prejuízos nos canaviais, afetando tanto a produtividade quanto a qualidade da matéria-prima destinada à indústria.

O Brasil perde anualmente cerca de 60 milhões de toneladas de cana-de-açúcar em decorrência do ataque de pragas. Considerando somente o que é destinado à produção de açúcar e etanol, isso corresponde a um prejuízo de aproximadamente 4,5 milhões de dólares (US$ 541 ha^{-1}) (OLIVEIRA et al., 2014). Uma das principais pragas responsáveis por parte desse prejuízo econômico é a espécie *Diatraea saccharalis* (Fabricius, 1794) (Lepidoptera: Crambidae), conhecida popularmente como broca-da-cana. A fase larval deste inseto trás inúmeros prejuízos visto que para se alimentar o mesmo abre galarias no interior da cana. Com isso, além da perda quantitativa de matéria prima ocasionada pela alimentação, essa praga favorece, também, a instalação de microrganismos como fungos e bactérias por meio dos orifícios que fazem ao penetrar nos colmos. Esses patógenos que atacam a cana em decorrência da infestação da broca afetam negativamente a qualidade dos produtos finais, como o açúcar e o etanol (DINARDO-MIRANDA, 2008; ROSSATO JÚNIOR et al., 2013)

A forma de encontrar a praga é iniciar intensos levantamentos para avaliar a presença desta praga no campo buscando massa de ovos, larvas pequenas nas folhas e larvas grandes dentro do colmo, sendo a observação dos dois primeiros estágios bem mais trabalhosa (SILVA et al., 2018). Nesse contexto, o desenvolvimento de modelos matemáticos para *D. saccharalis*, capazes de prever a ocorrência deste inseto em campo e, portanto, indicar o momento exato para a realização do controle pode trazer um enorme benefício.

Sendo assim, é importante entender com as condições do ambiente afetam a ocorrência dessa praga e, em função disso, estabelecer as regiões de maior risco e os momentos mais adequados para seu efetivo controle. Portanto, a hipótese deste estudo é que as condições meteorológicas afetam a dinâmica populacional e o grau de infestação da broca da cana-de-açúcar e que é possível se estimar isso considerando-se os bancos de dados provenientes das unidades produtoras por meio de técnicas de mineração de dados. Diante disso, este estudo teve os seguintes objetivos:

a) Obter, consistir e analisar o banco de dados de infestação de broca em diferentes regiões do Centro-Sul do Brasil;
b) Correlacionar o índice de infestação de broca com as variáveis meteorológicas;
c) Desenvolver modelos matemáticos ou de machine learning para estimação da dinâmica populacional da broca da cana-de-açúcar com base em variáveis meteorológicas e técnicas de mineração de dados;
d) Avaliar os modelos matemáticos para estimação da dinâmica populacional da broca da cana-de-açúcar com dados independentes;
e) Determinar o risco de ocorrência da broca da cana-de-açúcar em diferentes localidades do Estado de São Paulo com base no modelo de melhor desempenho e em séries históricas (30 anos), de modo a subsidiar as estratégias de controle dessa praga.

2. REVISÃO BIBLIOGRÁFICA

2.1. Broca da cana-de-açúcar

2.1.1. Importância da broca da cana-de-açúcar

De acordo com Nakano et al (2002), a broca-da-cana é umas das pragas mais importantes para a cultura da cana-de-açúcar, especialmente no Estado de São Paulo. A forma adulta dessa praga é uma mariposa de hábitos noturnos, a qual realiza a postura de seus ovos na face dorsal das folhas. Quando o canavial é jovem e não apresenta entrenós formados, a incidência da praga é menor, porém sabe-se que a broca-da-cana ocorre durante todo o ciclo da cultura (Pinto, 2006).

Apesar de o Brasil ocupar posição de destaque na produção e comercialização dos derivados de cana-de-açúcar, a cultura sofre com o ataque de diversos insetos-praga como brocas, cigarrinhas e besouros (PINTO et al., 2009), os quais causam danos diretos e indiretos, afetando negativamente a produtividade da cultura e a qualidade dessa matéria-prima. Os prejuízos ocasionados pela broca (*D. saccharalis*) são enormes pois o mais afetado é a produtividade da cultura visto que a praga fica no campo o ano todo. Os sintomas mais comuns são: amarelecimento das folhas, colmos menores, finos e secos, perda de peso da planta, redução do teor de sacarose, enraizamento aéreo dos colmos; danos na gema apical, em decorrência do ataque na região de crescimento, causando o sintoma conhecido por coração morto; e o favorecimento da ocorrência de patógenos como bactérias e fungos (BOTELHO; MACEDO, 2002; DINARDO-MIRANDA, 2008). Dentre estes microrganismos que se estabelecem na cultura da cana-de-açúcar em decorrência do ataque da broca, destacam-se as espécies *Colletotricum falcatum* (Went, 1893) e *Fusarium moniliforme* (Sheldon, 1904) que causam a chamada "podridão vermelha" e promovem a inversão da sacarose presente no colmo em levulose e glicose, que não são capazes de cristalizar, prejudicando, assim, a qualidade da matéria-prima para a indústria (BOTELHO; MACEDO, 2002; ROSSATO JÚNIOR et al., 2013).

Segundo Gallo et al. (2002), avaliando em torno de 100 colmos/talhão é possível checar no campo o índice de infestação final de broca (II) através da quantidade de internódios totais e internódios brocados, conforme a equação a seguir:

Intensidade de infestação *(II) = 100 . (nº de internódios brocados) / nº total de internódios*

Rossato Júnior (2012) também avaliou a intensidade de infestação (II) da broca na cultura da cana-de-açúcar e o efeito que essa praga causa na qualidade da matéria-prima e produção de açúcar e etanol. Esse autor observou que níveis elevados de intensidade de infestação (19,01% e 25,77%) causaram aumento de compostos fenólicos no caldo extraído bem como de fibra no colmo. Para Dinardo-Miranda et al. (2012), o índice mais utilizado para representar o dano que a broca causa à cultura da cana-de-açúcar é a intensidade de infestação, dado pela porcentagem de entrenós brocados. De acordo com Macedo e Lavorenti (2004), o índice da intensidade de infestação determina o impacto na porcentagem de entrenós brocados e na redução do teor de sacarose ocasionados pela broca.

2.1.2. Ciclo da praga

Há forte indício que a espécie *D. saccharalis* originou-se nas Américas (Central e do Sul). No hemisfério sul ocorreram os primeiros registros nas plantações de cana-de-açúcar (*Saccharum* spp.). É uma das principais pragas desta cultura (TERAN et al., 1983), embora também venha sendo tratada, recentemente, como uma séria ameaça para a cultura do milho (FIGUEIREDO et al., 2010). No Brasil, ocorrem apenas três dentre as 21 espécies pertencentes ao gênero *Diatraea*, entretanto, para a agricultura brasileira são de importância somente *D. saccharalis* e *D. flavipennella* (PINTO et al., 2009).

Segundo alguns autores como Melo e Parra (1988b) inferiram sobre o número de gerações possíveis para a espécie *D. saccharalis* por meio de estudos de exigência térmica em laboratório para as seguintes temperaturas: 20º, 22º, 25º, 30º e 32 ºC, e concluíram que no estado de São Paulo, em média, ocorrem cinco gerações por ano, sendo a última geração, que ocorre a partir do mês de abril, a de maior duração, em razão das menores temperaturas ao longo do período de desenvolvimento da praga, que coincide, neste caso, com o final do outono e início do inverno. Ainda de acordo com Melo e Parra (1988b), mais estudos sobre o número de gerações da broca da cana são necessários, já que este pode sofrer influência de outros, além dos graus dias, como o efeito da temperatura-base superior ou a presença de predadores e parasitoides, que podem influenciar no desenvolvimento dessa praga.

Segundo Geremias (2013), o limiar térmico inferior (temperatura base inferior = Tb) da *D. saccharalis* é de 14, 13 e 12 °C para as fases de ovo, lagarta e pupa, respectivamente. A temperatura ótima de desenvolvimento (Topt), estimadas pelos modelos de Lactin e Logan-6 foram de 30,2 a 30,7 ºC para a fase de lagarta e de 30,4 e 30,6 ºC para a fase de pupa, respectivamente. Para a fase de ovo, os valores de Topt permaneceram dentro do intervalo de 30 a 32 ºC. O limiar térmico superior de desenvolvimento (Tmáx) estimado por Geremias (2013) foi de 35,4 °C para a fase de lagarta, o qual ficou próximo da faixa de temperatura estimada experimentalmente, entre 36 e 38 ºC para fase de ovo e entre 32 e 34 ºC para a fase de pupa. Os limiares térmicos inferiores (ºC) e as constantes térmicas (Graus-dia) para as fases de ovo, lagarta e pupa foram, respectivamente, de: 14 ºC e 77 ºC; 13 ºC e 347 ºC.dia; 12 ºC e 139 ºC.dia. Esses valores, obtidos pelo modelo linear de desenvolvimento, diferem daqueles encontrados por Melo e Parra (1988b), o que pode estar associado às diferenças decorrentes da metodologia empregada na determinação da Tb e da constante térmica.

Este inseto tem as mariposas com cerca de 2,5 cm de envergadura e com asas anteriores e posteriores na cor amarelo-palha e branca, respectivamente, os ovos, cujo período de desenvolvimento embrionário varia de 4 a 12 dias, postos durante a noite na face inferior das folhas da cana e de forma imbricada com sobreposição parcial. Esses ovos possuem coloração amarelada e escurecem conforme o embrião se desenvolve. As larvas recém-eclodidas se alimentam, primeiramente, do parênquima da folha raspando-as até que atinjam o 2º ou 3º ínstar (PINTO et al.,2009). Posteriormente, perfuram o colmo e se alimentam formando galerias ascendentes em direção ao ponto de crescimento da planta. Muitas vezes, essas lagartas formam a galeria de modo circular, o que enfraquece o entrenó e facilita a quebra do colmo pelo vento (BOTELHO; MACEDO, 2002). As larvas permanecem no colmo até o 5º ínstar, cuja duração varia entre 20 e 79 dias e, deste modo, podendo haver em média cerca de cinco gerações anuais. No último ínstar, as lagartas atingem de 2,2 a 2,5 cm de comprimento e então fazem um orifício, antes de empuparem, cobrindo-o com fios de seda e serragem para que a saída do adulto seja facilitada após a emergência (TERAN et al., 1983; PINTO et al., 2009).

Conforme Pinto (2006), monitorar e levantar a praga no campo, principalmente no que tange a evolução da população de lagartas permite um bom monitoramento da mesma. O monitoramento ocorre em todo o ciclo da

cultura, desde o plantio até a colheita, sendo que é na colheita que a equipe de campo faz os levantamentos e calculada o II para observação de como foram as estratégias de controle e para se preparar para o planejamento para a safra seguinte.

2.1.3. Impactos causados pela broca na cultura da cana-de-açúcar

Os prejuízos provocados pela broca da cana-de-açúcar vêm sendo investigados por diversos autores. Estudos realizados na antiga Copersucar, com algumas variedades em cultivo no fim da década de 1990, revelaram que a cada 1% de intensidade de infestação da broca, as perdas podem chegar a 1,14% na produtividade de colmos, expressa em tonelada de cana por hectare (TCH), 0,42% em açúcar total recuperável (ATR) e 0,21% em etanol (ARRIGONI, 2002). Precetti et al. (1988), Gallo et al. (2002) e Campos e Macedo (2004) demonstram que para cada 1% de intensidade de infestação dessa praga ocorrem prejuízos de 0,25% na produção de açúcar, 0,20% na produção de álcool e 0,77% no peso da cana, o que foi confirmado pelos estudos de Sandoval e Senô (2010), que identificaram que as perdas relacionadas à infestação da broca-da-cana podem chegar a 35 kg de açúcar ha^{-1} e a 30 litros de álcool ha^{-1} com apenas 1% dos colmos infestados. Assim, o aumento da produtividade brasileira de cana-de-açúcar ao longo dos últimos anos se deveu, entre outros fatores, ao desenvolvimento de variedades mais tolerantes ao ataque da broca, o que diminuiu a intensidade de infestação de 10 para 2% em uma década (PINTO, 2006).

A broca acarreta em inúmeros danos à cultura da cana-de-açúcar como secamento e amarelecimento das folhas, brotação lateral, morte plantas atacadas, perda de TCH, prejuízo na matéria prima e qualidade do caldo, quebra de colmos em danos ocasionados pelas brocas na galeria da cana, além da ocorrência da podridão-de-fusarium e podridão vermelha, já citadas anteriormente. Dessa forma, dado grande impacto dessa praga para os canaviais e para as indústrias sucroenergéticas, o desenvolvimento de novas estratégias de seu monitoramento e manejo se fazem constantemente necessário.

2.1.4. Estratégias de controle da broca-da-cana

Atualmente, diversas estratégias são utilizadas para o controle da broca da cana-de-açúcar, sendo as mais utilizadas aquelas que empregam produtos químicos, biológicos e a utilização de feromônios. Com o avanço da praga e devido às pressões ambientais, o controle biológico cada vez mais é utilizado na agricultura para substituir produtos químicos (ODUM, 1988). De acordo com Braga et al. (2003) e Bug Agentes Biológicos (2009), o intuito do controle biológico é manter o equilíbrio no ambiente, mantendo as pragas em níveis aceitáveis graças a introdução de parasitóides, microorganismos e predadores naturais que vão acarretar em doença ou morte. O controle biológico da broca iniciou-se com as moscas, mas depois optou-se pelo uso do parasitoide *Cotésia flavipes*. Desde a sua primeira liberação em lavouras de cana-de-açúcar em Alagoas, em julho de 1974, observou-se uma melhoria no controle da broca. No estado de São Paulo o programa de controle biológico da broca com *C. flavipes* iniciou-se em 1977, sendo que em 1991 haviam sido liberados mais de 940 milhões de adultos desse parasitoide, o que resultou, em média para canaviais de 26 usinas, na redução da infestação de mais de 9% em 1980 para 3,17% em 1991 (CEBRASA, 2008). Além do controle biológico com cotésia, também é recomendado o uso de inimigos naturais predadores de ovos da broca-da-cana, como o parasitóide *Trichogramma galloi* **(**Hymenoptera: Trichogrammatidae) (Nakano et al., 2002).

Outra linha muito importante no manejo da praga e que vem ganhando muita expressão atualmente é a utilização de armadilhas de feromônios (NAKANO et al 2002). Nesta técnica de manejo costuma-se colocar no campo em pontos pré-determinados, normalmente 1 ponto para cada 50 hectares, armadilhas contendo fêmeas virgens com até 48 horas de idade, no interior de pequenas embalagens. Os machos são atraídos e ficam presos nas armadilhas para posterior avaliação e contagem que normalmente ocorre após 72 horas de sua instalação no campo. Além destas possibilidades também temos o controle químico, muito utilizado para o controle da forma larval da broca, antes que a mesma penetre no colmo da cana-de-açúcar, com diversos produtos registrados atualmente para este fim.

Segundo Papa (2007), a melhor estratégia de controle da broca-da-cana é a combinação harmônica dos diferentes métodos, como o químico, o varietal e biológico, levando a uma maior eficiência de controle dessa importante praga dos canaviais, portanto o manejo integrado de pragas (M.I.P), é ferramenta fundamental para o manejo das mais variadas pragas na agricultura atual.

2.2. Fatores agrometeorológicos que afetam a produtividade da cana-de-açúcar

A produtividade da cana-de-açúcar depende das condições ambientais, especialmente do total de energia solar, da temperatura do ar e da disponibilidade de água no solo. Marin et al. (2009) destacam que um bom crescimento vegetativo da cana-de-açúcar depende de um período de alta radiação solar, além de temperaturas elevadas e alta disponibilidade de água no solo. A cana-de-açúcar, por ser uma planta com metabolismo C4, possui uma maior taxa fotossintética do que as plantas C3, o que indica que alta incidência de radiação solar, dias longos, com temperaturas elevadas e com boa disponibilidade de água no solo levam a um maior perfilhamento e, consequentemente, a produtividades mais elevadas. Por outro lado, em condições de dias curtos, com baixa irradiância solar e temperaturas amenas, o perfilhamento é menos intenso (MONTEIRO, 2011) e as canas se tornam mais longas e finas (RODRIGUES, 1995), mesmo sob condições de boa disponibilidade de água no solo.

A temperatura do ar afeta os 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), enquanto 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).

Além das variáveis já citadas, a água desempenha papel crucial na produtividade da cultura da cana-de-açúcar. A demanda de evapotranspiração é o que define a necessidade de água da cultura. 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.

Um dos principais fatores ecológicos que influencia tanto direta quanto indiretamente nos insetos é a temperatura. A temperatura ótima em que os insetos têm mais rápido desenvolvimento e se reproduzem mais é em torno de 25 ºC. Entre 15 ºC e 38 ºC encontramos a temperatura limiar mínima e máxima para a maioria das espécies (SILVEIRA NETO et al., 1976). Segundo Teran (1979), os fatores que mais modificam as flutuações populacionais das pragas são as condições climáticas, especialmente a temperatura e a chuva. Em estudo realizado sobre a ocorrência de *Diatraea spp* na cultura do sorgo, Lyra Neto et al. (1990) observaram que a chuva foi a variável ambiental que mais se correlacionou com a flutuação dessa praga. Já Botelho et al. (1978) observaram que dentre os fatores climáticos que afetam a broca-da-cana, a chuva e a amplitude térmica explicaram 43,3% da flutuação populacional da praga Além disso, após diversos estudos foi observado por Botelho et al. (1993) que a broca apresenta dois picos populacionais, sendo o mais pronunciado em setembro no início das águas, e o segundo em fevereiro, com a menor prevalência ocorrendo no mês de junho, época em que ocorrem baixas precipitações na região centro-sul do Brasil (BOTELHO et al., 1993).

Diversos parâmetros influem no desenvolvimento dos insetos, organismos ectotérmicos, mas um dos principais é a temperatura (JAROSIK et al., 2004). A temperatura, um fator abiótico afeta diversos processos metabólicos, desenvolvimento, reprodução e mortalidade dos insetos. (BOMMIREDYY et al., 2004; DE CONTI et al., 2010; AMARASEKARE; SIFUENTES, 2012). Em diversos trabalhos uma forma de correlacionar desenvolvimento dos insetos e temperatura é através do uso de modelagem e validação dos dados, seja laboratorial ou a campo. Alguns dos parâmetros de extrema importância são a taxa de desenvolvimento (dias^{-1}), a constante térmica (NAVA et al., 2005; YUKAWA et al., 2016), os limiares térmicos inferior e superior (Temperatura base inferior e superior) e a taxa intrínseca de crescimento. A taxa intrínseca de crescimento permite determinar a faixa térmica favorável à espécie, ou seja, estipular os limiares de temperatura em que a população é capaz de se recuperar de baixas densidades e manter a população viável (AMARASEKARE; COUTINHO, 2013). Em alguns trabalhos realizados anteriores, alguns destes parâmetros foram estudados e estimados (MELO; PARRA, 1988; GEREMIAS, 2013).

2.3. Modelagem da dinâmica populacional de pragas

A sinergia entre a agrometeorologia e a ocorrência de pragas e doenças tem rendido uma série de sistemas de manejo dessas enfermidades, os quais melhoram a eficácia do controle, tornando-os economicamente viáveis. Além disso, esses sistemas vêm sendo utilizados para avaliar o risco climático para a ocorrência desses microrganismos e estabelecer as zonas de escape para essas enfermidades. Para muitas culturas em que a resistência genética às pragas e às doenças não é suficientemente alta, adota-se o controle químico para manutenção da produtividade, o qual deverá levar em consideração o melhor momento de controle com base nas condições ambientais (GILLESPIE E SENTELHAS, 2008).

Os modelos matemáticos são considerados uma forma de representar um sistema real (RAVINDRAN et al., 1987). Esses modelos descrevem, de acordo com um objetivo pré-estabelecido, os processos que ocorrem em um determinado ambiente por meio de algoritmos (JONES, 1989; SMERAGE, 1985). Os modelos são como abstrações

da realidade (HANNON; RUTH, 2001). Tais abstrações consistem em analisar um sistema e extrair a essência dele para, posteriormente, descrevê-lo e representá-lo por meio de equações matemáticas (LIMA et al., 2009). Portanto, segundo Hannon e Ruth (2001), modelos eficientes identificam mudanças e como elas impactam os processos subsequentes além de ajudar na previsibilidade futura e aplicação das descobertas geradas.

Os esforços para desenvolver modelos que descrevam a dinâmica da população dos insetos, especialmente aqueles de importância agrícola, têm ganhado destaque nos últimos anos. Por dependerem das condições ambientais para regular suas taxas metabólicas, esses microrganismos podem ter seu desenvolvimento descrito em função das características do ambiente, especialmente as condições meteorológicas. Assim, pode-se fazer uso desses modelos para avaliar o desenvolvimento e a dinâmica da população desses insetos tanto no cenário climático presente como futuro, já que estudos dessa natureza são mais importantes para as pragas das culturas, as quais interferem na produtividade das culturas agrícolas (SCHOWALTER, 2006). Nesse sentido, Sutherst e Maywald (1985) desenvolveram um software denominado CLIMEX, o qual prevê a distribuição de espécies de insetos/pragas de acordo com as condições climáticas, possibilitando assim antever bem como prever o risco de ocorrência de novas espécies em áreas agrícolas onde elas ainda não ocorrem.

A modelagem, como citada, é uma ferramenta que pode ser utilizada em sistemas agrícolas auxiliando na tomada de decisão em manejos. Por meio dela é possível estabelecer relações complexas, testar hipóteses, reformular delineamentos experimentais (MCKINION, 1992), entender as dinâmicas populacionais de pragas e inimigos naturais (SMERAGE, 1985, GUTIERREZ, 1996), simular a relação entre fatores abióticos e bióticos (SCOMPARIN, 2009). Assim, um modelo resulta da combinação de diferentes fatores, de natureza biológica, física ou da combinação de ambas, podendo incluir, ainda, fatores econômicos e sociais (JONES, 1989), podendo também ser essas variáveis qualitativas e quantitativas (LIMEIRA, 2010).

Uma das variáveis ambientais mais importantes na modelagem de microrganismos é a temperatura do ar, a qual afeta sobremaneira a dinâmica populacional de muitos deles, como, por exemplo, os insetos, que são organismos ectotérmicos (HADDAD e PARRA, 1984). Apesar da taxa de desenvolvimento dos insetos ser dependente da temperatura do ar, algumas espécies que habitam ambientes extremos são capazes de regular a temperatura corpórea, em alguns graus, por meio de respostas comportamentais ou fisiológicas (SCHOWALTER, 2006). Fabre et al. (2006), estudou a influência do ambiente externo no desenvolvimento de insetos (modelagem para simular dinâmica populacional de afídeos) observando alta relação destes em função da temperatura, além de variáveis como fotoperíodo e termoperíodo. Existem modelos matemáticos que foram criados para estudar a relação de controle entre *D. saccharalis* e *T. galloi (*AMBROSANO et al., 1996). O modelo levou em conta diversos fatores que afetam a relação entre broca-parasitóide-ambiente. Os resultados foram interessantes e foi considerado um trabalho válido para orientar o manejo da broca com a utilização deste parasitóide.

Portanto, de diversos trabalhos em literatura a temperatura é um dos fatores abióticos mais relevantes, seguido pelo fotoperíodo, parâmetro também crucial no ciclo dos insetos e que afeta diretamente a duração dos mesmos. Parra et al. (1983) determinaram o efeito do fotoperíodo para o desenvolvimento de larvas, pupas e adultos de *D. saccharalis*. Segundo os autores, o fotoperíodo tem uma forte relação com a temperatura do ar e que a fase em que a broca encontra-se como larva é mais impactada do que a fase em que encontra-se no estágio de pupa.

2.3.1. Uso do método dos graus-dias para descrever o desenvolvimento de insetos

O graus-dias é um método bastante conhecido para descrever o desenvolvimento de insetos (PEREIRA et al., 2002), que em suma é uma relação entre a temperatura do meio em que o inseto se encontra e o seu desenvolvimento. Este método demonstra a relação que existe entre o desenvolvimento de animais ectotérmicos com as condições meteorológicas sendo, portanto, muito utilizado como uma ferramenta em estudos de ecologia, de insetos em geral, mas principalmente para o manejo e previsão de ocorrência e proliferação de pragas das culturas agrícolas (HIGLEY et al., 1986) para se determinar, por exemplo, emergência, desenvolvimento de folhas e caules, floração, entre outros (MILLER et al., 2018).

O requerimento térmico necessário ao desenvolvimento dos organismos, supostamente proposto por Réaumur no século XVIII, pode ser expresso por meio da constante térmica (K). Esta constante, expressa em graus-dia (GD, ^{o}C dia), revela que o tempo de desenvolvimento relacionado à temperatura é constante (SILVEIRA-NETO et al., 1976).

$$K = d \, . \, (Tmed - Tb)$$

em que: K = constante térmica; d = duração do tempo de desenvolvimento; Tmed = temperatura média do ambiente; e Tb = temperatura base inferior do organismo.

Segundo Haddad et al. (1999), como os insetos são dependentes da temperatura do ar para o rápido desenvolvimento, é necessário considerar a temperatura base abaixo do qual o inseto não se desenvolve (Tb). Dessa maneiro o estudo com a praga é muito importante e estudos em laboratório com temperaturas constantes são necessários para a determinação dos limites térmicos inferior e superior dos insetos.

Geremias (2013) utilizou o modelo dos graus-dia com o objetivo de determinar a faixa ótima de temperatura e a constante térmica para a *D. saccharalis* em condições de laboratório e, posteriormente, em condições de campo, a fim de, validar o modelo obtido. Segundo esse autor, a constante térmica obtida em laboratório por meio deste modelo pode ser utilizada em estudos de campo desde que os dados de temperatura tenham a sua origem bem estabelecida, ou seja, sejam obtidos próximos à área de interesse e com equipamento que possibilite a determinação da real temperatura média do ar. Em seu trabalho, Geremias (2013) empregou dados de estações convencionais e automáticas, e observou que os dados desta última foram os mais adequados por não superestimarem ou subestimarem a temperatura média diária.

2.4. Mineração de Dados

Atualmente, em função do enorme volume de informações geradas e prontamente disponíveis, há uma grande dificuldade para a análise e interpretação desses bancos de dados, utilizando-se ferramentas Big Data para tais análises. De acordo com Witten et al. (2011), a evolução rápida da geração de informações tem elevado o volume de dados armazenados ao redor do mundo, o que, segundo esses autores, se duplica a cada 20 meses. Assim, isso produz uma grande lacuna entre gerar e interpretar as informações. Por estes fatores as técnicas de mineração e aprendizado de máquinas ajudam no entendimento destes grandes volumes de dados. Diversas áreas como medicina, finanças e agropecuária podem usufruir da MD e reunir informações e conhecimentos (GALVÃO; MARIN, 2009).

De acordo com Camilo e Silva (2009), existem diversos processos que descrevem as fases e atividades da MD, sendo o mais utilizado o modelo CRISP-DM (*Cross Industry Standard Process for Data Mining*), que consiste das seguintes fases (CAMILO; SILVA, 2009):

a) Fase de compreender o contexto (*business understanding*): Há um planejamento inicial do projeto que deverá seguir as premissas da mineração de dados, bem como entender quais os objetivos e necessidades do mesmo;

b) Fase de entendimento dos dados (*data understanding*): O banco de dados fornece uma gama acentuada de variações, fontes e origens. Portanto nessa fase, gasta-se tempo para aproximação e entendimento dos dados visando conhecer mais detalhes e possibilitar identificar problemas com antecedência;

c) Fase de preparação dos dados (*data preparing*): Nesta fase, após um bom trabalho com o banco de dados, é possível gerar um banco final de dados, que será utilizado nas análises futuras. Este banco precisa de adequações para que fique completo e confiável

d) Fase de modelagem (*modeling*): Nesta fase aplica-se de fato as técnicas de mineração de dados, sendo que cada problema tem o seu contexto específico, demandando banco de dados apto para cada exigência;

e) Fase de avaliação (*evaluation*): Nesta fase é importante validar, testar e revisar os processos e entender se o objetivo foi alcançado para tomar decisões assertivas no futuro;

f) Fase de Implantação (*deployment*): Feito a avaliação é importante implantar os sistemas e distribuir conhecimento ao longo do processo transmitindo ao usuário final da melhor forma.

Como se pode observar na Figura 2, as seis fases do processo de MD 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.

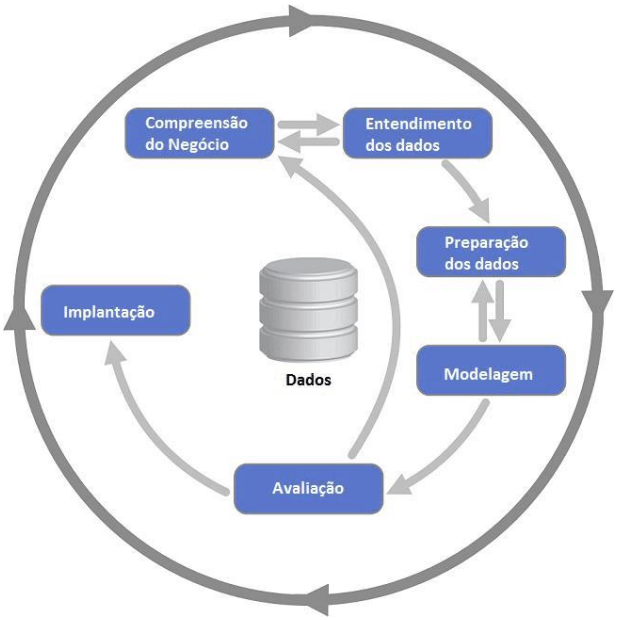


Figura 2. Fases do processo de mineração de dados, denominada de CRISP DM (CHAPMAN et al., 2000).

Segundo Han et al. (2011), as tarefas da mineração de dados são divididas em preditivas (predições encima de dados) ou descritivas (caracterização dos dados). As descrições dessas tarefas são apresentadas a seguir:

a) Tarefas preditivas

As tarefas preditivas são divididas em duas etapas:

Classificação: a classificação pode ser definida como um processo que visa encontrar modelos com objetivo de descrever e distinguir classes. Esses modelos são gerados com base na análise de uma base de dados de treinamento, onde os dados são conhecidos. A partir disso, 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 artificiais (HAN et al apud MARTINS, 2020).

Regressão: a análise de regressão é uma metodologia estatística usada principalmente para predições numéricas. Enquanto a classificação estima fatores categóricos (discretos), os modelos de regressão estimam funções de valores contínuos (HAN et al apud MARTINS, 2020).

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). Técnica muito aplicada quando se precisa analisar informações médicas e também quando se quer traçar o perfil de compras dos clientes (BARROS, 2013).

Agrupamento ou Clusterização: Neste caso não temos análises de dados caracterizados, no agrupamento (cluster) o objetivo é analisar dados sem o conhecimento prévio de seus agrupamentos em classes. *Clusters* são gerados de forma que os objetos dentro de um grupo 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 apud MARTINS, 2020).

Portanto, cada tarefa tem sua especificidade e diferentes técnicas podem ser utilizados, não tendo certo ou errado, melhor ou pior e sim as mais adequadas para cada caso e resolução do problema em questão (VIEIRA, 2014). Dentre as diversas técnicas disponíveis para realizar a tarefa de predição, pode-se destacar as denominadas de *Boosting*, *Random forest*, ambas pertencentes ao grupo das Árvores de Decisões, e o *Support Vector Machine*. A seguir são apresentadas breves descrições dessas técnicas.

2.4.1. Grupo das Árvores de Decisão

Os grupos de árvore de decisão têm em sua construção e concepção uma forma mais simples de entendimento podendo ser utilizados tanto em tarefas preditivas quanto em descritivas. No contexto das MD, as árvores de decisão são ferramentas bastante utilizadas sendo denominados árvores de classificação (variável dependente categórica) e árvores de regressão (variável contínua) (SPINDEL, 2008). Nos sistemas estudados as árvores de decisão captam relações não lineares, na maioria das vezes em que o uso de abordagens lineares, como a regressão, fica restrito (JAMES et al., 2013). Para tanto, é utilizado 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 numa árvore, porém em sentido inverso (WITTEN et. al., 2011), conforme apresentado na Figura 3. Seu ciclo de construção começa com um único nó de decisão, situado mais acima na árvore, chamado de nó inicial ou raiz. Nos nós internos, ou nós 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 continua

sendo executada até que não haja mais ganhos expressivos na capacidade preditiva do modelo. Assim, o dado entrante é submetido a sucessivas avaliações e, ao final, um ou mais modelos localizados nos nós terminais, ou folhas, são combinados para gerar o resultado final (SPINDEL, 2008; TACONELI, 2008).

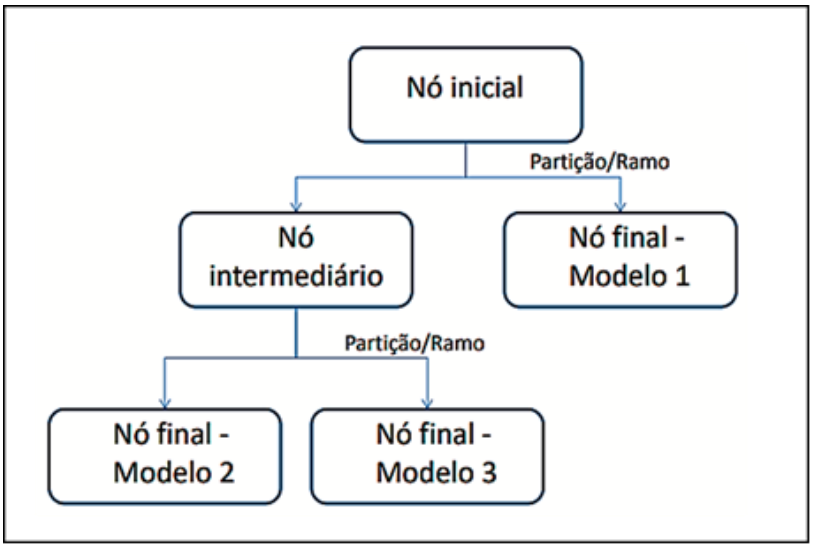


Figura 3. Representação esquemática de uma árvore de decisão.

Em árvores de regressão e classificação deve-se buscar uma árvore de menor tamanho (número de nós finais) possível, 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 adotando-se o procedimento denominado de poda, o qual 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*

A técnica *Random Forest* se baseia na aplicação de *bagging* em árvores de decisão, com uma importante extensão - além de amostrar os registros, o algoritmo também amostra as árvores. Nas árvores de decisão tradicionais, para determinar como criar uma sub-repartição de uma repartição A, o algoritmo escolhe uma variável e um ponto de divisão por meio da minimização de um critério a ser escolhido. Entretanto, no caso das florestas aleatórias (ou *Random Forest*), em cada estágio do algoritmo, a escolha de uma variável é limitada a um subconjunto aleatório de variáveis. Dessa forma, comparado com o algoritmo de árvore básico, o algoritmo de *Random Forest* adiciona mais dois passos: o *bagging* e a amostragem *bootstrap* das variáveis em cada divisão (BRUCE; BRUCE, 2019).

O primeiro passo é o *bagging* (ou *bootstrap* agregador), utilizado quando buscamos uma resposta Y e temos P variáveis preditoras, ou seja, Y é função de X1, X2, ..., Xp com n registros. *Bagging* é como o algoritmo básico de

agrupamentos, exceto pelo fato de que, em vez ajustar diversos modelos aos mesmos dados, cada modelo é ajustado a uma reamostra *bootstrap*, seguindo os seguintes passos abaixo (BRUCE; BRUCE, 2019):

1. Inicializar M (o número de modelos a serem ajustados) e n (o número de registros a escolher), com a condição n < M. Ajustar a interação m=1;
2. Tirar uma reamostra *bootstrap* (ou seja, com reposição) de n registros dos dados de treinamento para formar uma subamostra Y_m a X_m (a *bag*);
3. Treinar um modelo usando Y_m e X_m para criar um conjunto de regras de decisão $\hat{f}_m(X)$;
4. Incrementar o contador do modelo m = m + 1. Se m ≤ M, voltar ao Passo 1.

No caso em que $\hat{f}_m$ prevê a probabilidade Y = 1, a estimativa *bagged* é dada por:

$$\hat{f}_m = \frac{1}{M}\left((\hat{f}_1(X) + \hat{f}_2(X) + \cdots \hat{f}_m(X)\right) \qquad (2)$$

A segunda etapa é a amostragem *bootstrap* (com reposição) dos registros;

1. Tirar uma subamostra *bootstrap* (com reposição) dos registros;
2. Para a primeira divisão, amostrar p < P variáveis aleatoriamente sem reposição;
3. Para cada uma das variáveis amostradas $X_{j\,(1)}, X_{j\,(2)}, \ldots X_{j\,(p)}$, aplicar o algoritmo da divisão:

a) Para cada valor $s_{j\,(k)}$ de $X_{j\,(k)}$:

- Dividir os registros na repartição A com $X_{j\,(k)} < s_{j\,(k)}$ como uma repartição, e os registros restantes em que $X_{j\,(k)} \geq s_{j\,(k)}$ como outra repartição
- Meça a homogeneidade de classes dentro de cada sub-repartição de A

b) Selecionar o valor de $s_{j\,(k)}$ que produzam máxima homogeneidade de classes dentro da repartição.

4. Selecionar a variável $X_{j\,(k)}$ e o valor da divisão $s_{j\,(k)}$ que produzam máxima homogeneidade de classes dentro da repartição;
5. Prossiga à próxima divisão e repita os passos anteriores, começando com o passo 2;
6. Continuar com divisões adicionais seguindo o mesmo procedimento até que a árvore tenha crescido;
7. Voltar ao passo 1, retirar outra amostra *bootstrap*, e recomeçar todo o processo.

Para se escolher o número de variáveis que se deve amostrar em cada passo utiliza-se a regra $\sqrt{P}$, em que P é o número de variáveis preditoras.

A potência do algoritmo de *Random Forest* se mostra quando se constrói modelos preditivos para dados com muitas características e registros, uma vez que ele tem a habilidade de determinar automaticamente quais variáveis preditoras são importantes e descobrir relacionamentos complexos entre elas, correspondentes a termos de

interação. Entretanto, da mesma forma que muitos algoritmos estatísticos de aprendizado de máquina (*Machine Learning*), *Random Forest* pode ser considerado um algoritmo de "caixa-preta", ou seja, com manivelas para ajustar o funcionamento dessa caixa. Essas "manivelas" são chamadas de hiperparâmetros, os quais dizem respeito aos parâmetros que devem ser ajustados antes de se desenvolver um modelo. Esses hiperparâmetros não são otimizados como parte do processo de treinamento e, diferentemente dos modelos estatísticos tradicionais, os hiperparâmetros para *Random Forest* não exigem escolhas (por exemplo, a escolha dos preditores a serem usados em um modelo de regressão) e são mais críticos, especialmente para evitar sobreajuste (BRUCE; BRUCE, 2019). Os dois hiperparâmetros mais importantes para *Random Forest* são:

- ***Nodesize*** - tamanho mínimo para os nós terminais (folhas na árvore). O padrão é 1 para árvores de classificação e 5 para árvores de regressão.
- ***Maxnodes*** - número máximo de nós em cada árvore de decisão. Não existe um limite padrão e árvores maiores deverão ser ajustadas conforme as restrições de *nodesize*.

Além destes temos alguns outros de grande importância: min_n, mtry, trees, tree_depht.

a) Boosting

O *Boosting* é uma técnica geral usada para criar um agrupamento de modelos e foi desenvolvido na mesma época que o *Bagging*. Assim como o *Bagging*, o *Boosting* também é usado em árvores de decisão, entretanto, apesar de suas similaridades, o *Boosting* assume uma abordagem com muito mais detalhes. Como resultado, enquanto o *Bagging* pode ser feito relativamente com poucos ajustes, o *Boosting* exige muito mais cuidados em sua aplicação (BRUCE; BRUCE, 2019).

Em modelos de regressão linear, os resíduos costumam ser examinados para ver se o ajuste pode ser melhorado. O *Boosting* leva esse conceito muito além e ajusta uma série de modelos com cada modelo sucessivo ajustado para minimizar o erro dos anteriores. Algumas variantes do algoritmo costumam ser usadas: *Adaboost*; *Boosting* gradiente e *Boosting* gradiente estocástico. O último, *Boosting* gradiente estocástico, é o mais geral e amplamente usado, uma vez que incorpora a reamostragem de registros e colunas em cada rodada. A ideia básica por trás dos diversos algoritmos de *Boosting* é essencialmente a mesma. Porém, o mais fácil de entender é o *Adaboost*, descrito a seguir (BRUCE; BRUCE, 2019):

1. Inicializar M, o número máximo de modelos a serem ajustados, e ajustar o contador de iterações m = 1. Inicializar os pesos de observação por meio da Eq.3:

$$w_i = \frac{1}{N} \tag{3}$$

   para $i = 1,2, \dots, N$. Inicializar o modelo de agrupamento $\hat{F}_0 = 0$;

2. Treinar um modelo usando $\hat{f}_m$ utilizando os pesos de observações $w_1, w_2, \dots, w_N$, que minimize o erro ponderado e_m definido pela soma dos pesos para as observações classificadas erroneamente;

3. Adicionar o modelo ao grupo $\hat{F}_m$, dado pela equação a seguir:

$$\hat{F}_m = \hat{F}_{m-1} + \alpha_m \times \hat{f}_m \tag{4}$$

em que α_m é dado por:

$$\alpha_m = \frac{log1 - e_m}{e_m} \tag{5}$$

4. Atualizar os pesos $w_1, w_2, \dots, w_N$ de forma que eles sejam aumentados para as observações que eram mal classificadas. O tamanho do aumento depende de α_m, com valores maiores de α_m levando a maiores pesos;
5. Incrementar o contador de modelo $m = m + 1$. Se $m \leq M$, voltar ao passo 1.

Em geral, a estimativa *Boosted* é dada pela seguinte equação:

$$\hat{F} = \alpha_1 \times \hat{f}_1 + \alpha_2 \times \hat{f}_2 + \cdots + \alpha_M \times \hat{f}_M \tag{6}$$

Ao aumentar os pesos para as observações mal classificadas, o algoritmo força os modelos a treinarem mais pesadamente nos dados que tiveram um mal desempenho. Dessa forma, o fator α_m garante que os modelos com menor erro tenham maior peso. No âmbito do *Random Forest*, o *Boosting* gradiente estocástico inclui a aleatoriedade ao algoritmo visto acima por meio da amostragem de observações e variáveis preditoras em cada estágio (BRUCE; BRUCE, 2019).

O software de domínio público mais usado para *Boosting* é o *XGBoost,* uma implementação do *Boosting* gradiente estocástico. Trata-se de uma implementação computacional eficiente com muitas opções e está disponível como um pacote para as principais linguagens de software de ciência de dados. Na linguagem R, por exemplo, o *XGBoost* está disponível como o pacote *XGboost* (BRUCE; BRUCE, 2019).

A função *XGboost* tem muitos parâmetros que devem ser ajustados, sendo os dois mais importantes o *subsample* (controla a fração de observações que devem ser amostradas em interação) e o *eta* (fator de encolhimento aplicado em α_m no algoritmo de *Boosting*). Usar o *subsample* faz o *Boosting* agir como *Random Forest*, exceto pelo fato de a amostragem ser feita sem reposição. Já o parâmetro de encolhimento *eta* é útil para prevenir o sobreajuste por meio da redução da mudança de pesos, já que uma mudança menor nos pesos significa que o algoritmo é menos propenso a sobreajustar o conjunto de treinamento (BRUCE; BRUCE, 2019).

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). Supondo-se que seja necessário escolher, em um plano bidimensional, um classificador que separe os círculos dos triângulos na ilustração abaixo (Figura 4). Na Figura 4a tem-se um classificador com um erro

considerável, em função de sua baixa capacidade em distinguir pontos próximos pertencentes às diferentes classes. Sua taxa de acerto é bem inferior à função da Figura 4c, 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 da Figura 4a é muito mais simples do que o da Figura 4c. Um classificador mais complexo que o da Figura 4a e não tão sobre-ajustado quanto o da Figura 4c seria o classificador da Figura 4b. Ele tem uma complexidade intermediária e classifica bem grande parte dos exemplos (LORENA; CARVALHO, 2003).

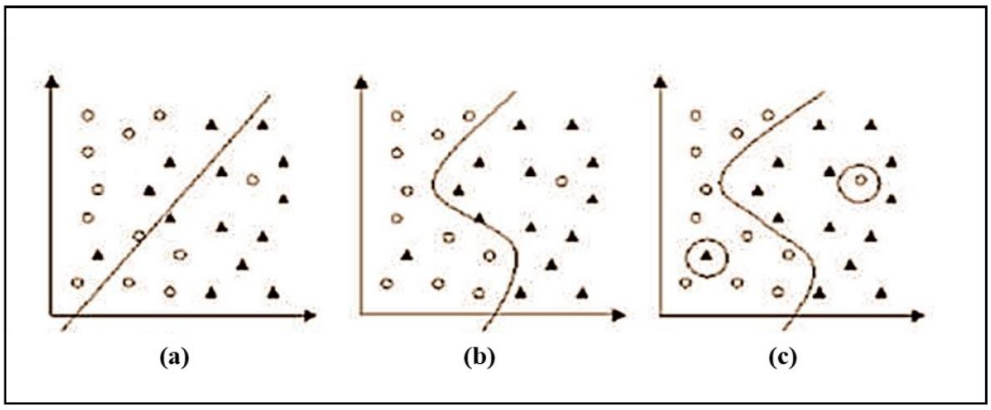


Figura 4. Exemplo de conjunto de dados de treinamento pertencentes a duas classes (círculos e triângulos) e classificados segundo três diferentes hipóteses pela técnica SVM: (a) com erro considerável; (b) com complexidade intermediária; e (c) com elevada complexidade. Adaptado de Lorena e Carvalho (2003).

As SVMs foram desenvolvidas com base na Teoria do Aprendizado de Máquina 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 para os de teste. Originalmente, as SVMs 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 5, 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).

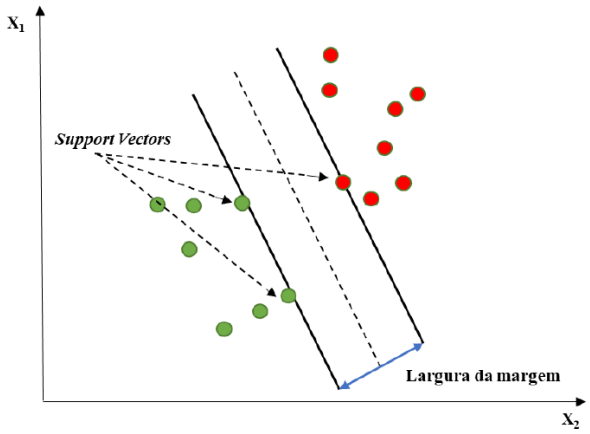


Figura 5. Representação esquemática de observações separadas por hiperplano com respectivas margens. Adaptado de Sayad (2010).

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 $yi = +1, -1$, 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 otimização (HASTIE, 2009; SAYAD, 2010):

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

Como a solução que minimiza a função objetivo consiste em um problema convexo, então existe um valor mínimo global. A Figura 6 representa o problema de otimização de uma SVM para classificação linear.

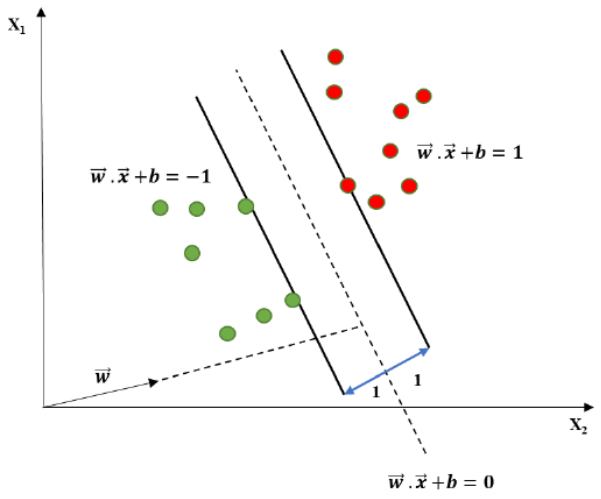


Figura 6. Representação esquemática do problema de otimização de uma SVM para a classificação linear. Adaptado de Sayad (2010).

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. Nessas situações, as SVM procuram encontrar um hiperplano que maximize as margens ao mesmo tempo 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 $yi = +1, -1$ é possível por meio da adição de uma variável de penalização ξ_i. A Figura 7 representa um exemplo da aplicação de uma SVM no caso em que os dados não são linearmente separáveis.

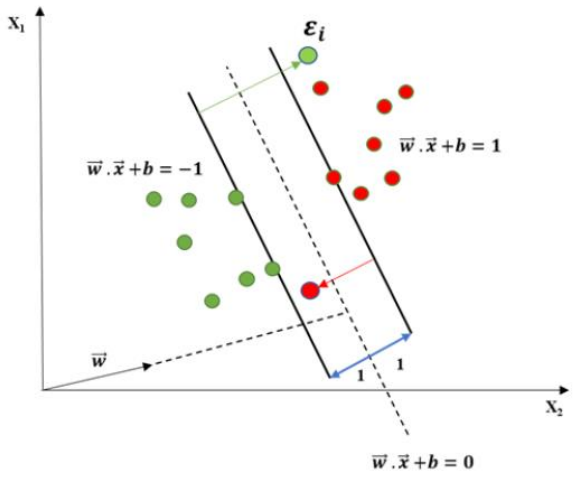


Figura 7. Representação esquemática do problema de otimização de uma SVM para o caso em que o conjunto de dados não é linearmente separável. Adaptado de Sayad (2010).

No caso em que os dados não são linearmente separáveis, a variável ξ_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 otimização (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 de as 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, 1999; 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).

2.5. O uso de Mineração de Dados na agricultura

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 envolvendo técnicas de mineração de dados foram os que mostraram maior potencial para utilização na tomada de decisão referente à adoção de medidas de controle da 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

identificaram diversas variáveis que a cana-de-açúcar, tanto do ambiente quanto de manejo, de acordo com sua importância hierárquica na variabilidade das produtividades, 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 apenas pelas localidades.

Nonato (2010) desenvolveu modelos preditivos para identificar áreas cultivadas com cana-de-açúcar em imagens do satélite LANDSAT 5 no estado de São Paulo, 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.

Hammer (2016), analisando dados de produtividade, manejo agrícola e dados meteorológicos de 18 usinas de cana-de-açúcar no estado de São Paulo, ordenou as variáveis que mais condicionam a produtividade da cana-de-açúcar, de acordo com a sua importância, e desenvolveu modelos matemáticos para a estimação da produtividade de cana-de-açúcar, com base em técnicas de mineração de dados (*Random Forest, Boosting e Support Vector Machines*). Esse autor observou que dentre as variáveis que afetam a produtividade da cana o número de cortes foi o que teve maior peso, seguido da idade do corte e de variáveis ambientais relativas às condições meteorológicas e dos ambientes de produção (solo). Esse autor também observou que os modelos gerados tiveram desempenho mediano, mas melhor do que o modelo da zona agroecológica da FAO.

Martins (2020) em seu trabalho descreve que para determinar e definir regiões com alto potencial de produção de fibras de qualidade superior é importante conhecer os fatores que condicionam a qualidade da fibra de algodão. Entendendo a grande importância de trabalhar com matérias superiores e visando incentivar a produção de fibras de algodão de melhor qualidade, os objetivos de seu estudo foram identificar e ordenar os fatores que afetam a qualidade da fibra de algodão, de acordo com sua importância, bem como o desenvolvimento de modelos matemáticos de estimação para os índices de qualidade da fibra de algodão por meio de técnicas de mineração de dados (MD). Para tanto, dados das variáveis meteorológicas e de índices de qualidade da fibra de algodão de 32 cultivares brasileiras foram submetidos à análises de correlação de Pearson (para uma abordagem geral); à análise de agrupamento ("clusterização"); e à três técnicas de MD (*Random Forest*; *Boosting*; e *Support Vector Machine*). De modo geral, os modelos gerados por meio do emprego de técnicas de MD tiveram desempenho melhor do que as análises de correlação feitas a nível de abordagem geral e ao nível de agrupamento. O modelo *Bossting* foi ligeiramente melhor do que os modelos *Random Forest* e o *Support Vector Machine*. As variáveis preditoras que foram comuns às três técnicas para os índices de qualidade da fibra avaliados variaram, porém estiveram ligadas à radiação solar, temperatura e chuva.

3. MATERIAL E MÉTODOS

A descrição dos procedimentos utilizados no estudo foi considerada utilizando as etapas que constituem o processo CRISP DM (*Cross Industry Standard Process for Data Mining*) já citada anteriormente no trabalho.

Além disso, 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: *Random Forest* (RF); *Boosting (*XGB); e *Support Vector Machine* (SVM).

3.1. Entendimento dos dados

Para o trabalho foram utilizados dados de índice anual de infestação final de broca da cana-de-açúcar (*D. saccharalis*) de oito safras (2012-2013, 2013-2014, 2014-2015, 2015-2016, 2016-2017, 2017-2018, 2018-2019, 2019-2020), para o período de janeiro a dezembro, de área de cultivo da região de Jaboticabal que compreende nove micro-regiões, sendo elas: Barrinha; Bueno de Andrade; Dumont; Guariba; Guatapará; Luis Antônio; Monte Alto; Motuca; e Pradópolis, no interior do estado de São Paulo. Foram utilizados os bancos de dados ao nível de gleba, considerando-se apenas as áreas em que não houve a realização de controle químico e sem a intervenção de controle biológico por meio da liberação da vespa *Cotesia Flavipes*. Esses dados correspondem às avaliações feitas em 30 variedades de cana-de-açúcar. Além da variável índice de infestação final da broca-da-cana (II), também foram utilizados no trabalho as seguintes variáveis meteorológicas e de balanço hídrico, na escala diária: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); precipitação pluviométrica (chuva); irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125). A etp foi estimada com base no método de Penman-Monteith (ALLEN et al, 1998), enquanto as variáveis def_125 e exc_125 foram estimadas por meio do balanço hídrico climatológico de Thornthwaite e Mather (1955), considerando-se uma capacidade de água disponível no solo (CAD) de 125 mm, a qual engloba grande parte das características dos solos da região de estudo. Os dados meteorológicos empregados neste estudo foram obtidos junto às bases de dados do Instituto Nacional de Meteorologia (INMET), do Centro Integrado de Informações Agrometeorológicas (CIIAGRO), do Instituto Agronômico de Campinas (IAC), e do Sistema Integrado de Gerenciamento de Recursos Hídricos do Departamento de Água e Energia Elétrica do Estado de São Paulo (SIGRH-DAEE-SP), os quais foram interpolados para a obtenção dos dados das regiões de interesse.

O Índice de infestação final de broca (II), corresponde ao período em que o canavial ficou no campo referente à colheita do ano anterior até a data de colheita do ano vigente, sendo mensurado a campo no momento da colheita pelo auditor da usina em questão da seguinte forma:

Intensidade de infestação *(II) = 100 . (nº de internódios brocados) / nº total de internódios*

Como exemplo, para o cálculo do período de estudo a nível de gleba, considerou-se o momento da colheita de determinada gleba no mês atual subtraído da data de colheita ocorrida no ano anterior, dessa forma temos o período que compreende o intervalo dos dados do estudo que correspondem majoritariamente, em média, de 10 a 16 meses para aproximadamente 2.900 glebas, que é o banco de dados de II.

Quanto as variáveis meteorológicas e do balanço hídrico, os dados foram computados para as respectivas safras (2012-2013 a 2019-2020) e de acordo com o período em que a cultura permaneceu no campo, ou seja, entre a data do corte do canavial do ano anterior e o corte do ano vigente. As variáveis ndc35 e ndc13 foram calculadas utilizando-se os dados de temperatura máxima e mínima presentes no banco de dados, para o período referente à gleba em estudo. Além dessas variáveis, também foram empregados os dados de graus-dia (GD), importante variável para estudos com pragas, sendo esses calculados diariamente pela diferença entre a tmed e a temperatura basal inferior (tb) para o desenvolvimento da praga. Apesar da importância desta variável em estudos de ecologia de insetos, esta apresentou alta colinearidade com a tmed, optando-se, assim, por manter neste estudo apenas a tmed, em função da simplicidade de sua obtenção.

3.2. Preparação dos dados

Inicialmente, o banco de índice de infestação de broca e dos dados meteorológicos e do balanço hídrico continha 25 variáveis, as quais foram processadas em uma análise preliminar no software R para identificar aquelas com maior grau de importância. Assim, apenas 12 variáveis (II, tmax, tmin, tmed, ndc35, ndc13, ur, chuva, Rad, etp, def_125 e exc_125) foram utilizadas para a confecção dos modelos, as quais foram compiladas e unificadas em uma planilha de Excel® para manipulação e posterior uso no software R, onde foi aplicado tratamento rigoroso para a retirada de *outliers*. Após esse tratamento, o banco de dados final continha 2876 dados a nível de gleba, considerando-se oito safras e 12 variáveis, fato que justificou o uso de MD para análise e geração de conhecimento a partir deles.

3.3. Correlação de Pearson

O coeficiente de correlação de Pearson (r) ou coeficiente de correlação produto-momento mede o grau da relação linear entre duas variáveis quantitativas. É um índice adimensional com valores situados ente -1,0 e 1.0, que reflete a intensidade de uma relação linear entre dois conjuntos de dados. Esse coeficiente apresenta as seguintes interpretações:

$r = 1 \rightarrow$ indica correlação perfeita e positiva entre as duas variáveis;

$r = -1 \rightarrow$ indica correlação perfeita e negativa entre as duas variáveis;

$r = 0 \rightarrow$ indica que não há correlação linear entre as duas variáveis, o que, no entanto, não significa que não possa haver correlação "não linear". Assim, o resultado $r = 0$ deve ser investigado por outros tipos de relação.

Visando analisar a importância de cada variável utilizada nos algoritmos de mineração de dados, as mesmas foram correlacionadas entre si e com o índice de infestação da broca-da-cana (II) gerando uma matriz de correlação. Essa matriz de correlação foi reelaborada no ambiente R, sendo que quanto maior a correlação maior a intensidade das cores, sendo azul para as correlações positivas e vermelha para as correlações negativas.

3.4. Análise de Componentes Principais

Uma segunda análise empregada ao conjunto de dados foi a de componentes principais, a qual teve como objetivo resumir as informações relevantes contidas nas p variáveis ($p = 12$), em um conjunto menor de variáveis

latentes ortogonais que são denominados componentes principais (autovetores), gerados por combinações lineares das variáveis originais a partir dos autovalores da matriz de covariância.

Cada par de componentes principais gerou uma representação bidimensional do espaço amostral original denominado "*biplot*" na qual é possível explicar a estrutura de variáveis direcionando feixes das variáveis nas regiões de máxima variabilidade. Foram considerados os componentes principais (CP's) cujos autovalores foram superiores à unidade (KAISER, 1958). Os coeficientes das funções lineares, que definem os CP's, foram utilizados na interpretação de seu significado, usando o sinal e o tamanho relativo dos coeficientes como uma indicação do peso a ser atribuído para cada variável. Essa análise também foi processada no ambiente R.

3.5. Modelos de estimação do índice de infestação da broca-da-cana

Os modelos de estimação do índice de infestação da broca-da-cana foram gerados a partir da execução de diferentes técnicas de MD no ambiente R. As técnicas empregadas neste estudo foram: *Random Forest* (RF); *Boosting* (XGB); e *Support Vector Machine* (SVM), as quais já foram devidamente descritas anteriormente. Como interface para a aplicação dessas técnicas, foi utilizado o pacote *Classification And Regression Training* (*Caret*), o qual 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 MD, como por exemplo, o número de árvores ideal no RF, bem como a caracterização das variáveis preditoras mais importantes dos modelos. Além disso, o pacote *Caret* proporciona uma visualização mais amigável dos resultados (KUHN, 2008). Foi realizada uma abordagem geral, em que os algoritmos acima foram aplicados na base integral de dados, que contém os dados de todas as regiões e safras analisadas. Com base nisso, os dados foram organizados para que fossem submetidos ao processo de treinamento dos algoritmos de MD e posteriormente validados.

3.6. Avaliação dos modelos de estimação do índice de infestação da broca-da-cana

Para se elaborar a modelagem por meio das técnicas de MD, inicialmente separou-se 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 MD, 30% dos dados foram selecionados aleatoriamente da base original. A análise do desempenho dos modelos foi efetuada pela comparação entre os índices de infestação da broca-da-cana final estimados pelos modelos e os índices de infestação observados a campo. O desempenho dos modelos foi avaliado considerando-se os seguintes índices estatísticos: coeficiente de correlação (r); coeficiente de determinação (R^2); índice de concordância "D" de Willmott (1981), 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 do índice de infestação; 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.

3.7. Determinação e mapeamento do risco de infestação da broca-da-cana

Para a determinação do risco de infestação da broca-da-cana foram selecionados 25 locais no Estado de São Paulo (Tabela 1), para os quais foi utilizado o modelo de estimação de II de melhor desempenho nas fases de treinamento e teste. Para cada um desses 25 locais foram obtidos os dados meteorológicos de um período de 30 anos (1990 a 2019), considerando-se todas as 11 variáveis necessárias para a estimação de II. Esses dados foram obtidos junto às mesmas bases já mencionadas anteriormente (INMET, CIIAGRO, SIGRH-DAEE-SP) e interpolados para as regiões de interesse. O modelo de infestação da broca-da-cana foi gerado anualmente resultando em 25 valores de II. Com base nesses dados foi estabelecida a frequência relativa de anos com infestações acima do limiar considerado pelas empresas sucro-energéticas como definidor da necessidade de controle químico, ou seja, 2%, para cada uma das 25 localidades da Tabela 1.

Tabela 1. Localidades utilizadas para a determinação do risco de infestação de broca-da-cana no Estado de São Paulo e do modelo linear para espacialização da infestação em função das coordenadas geográficas e da altitude.

Local	Lat	Long	Alt (m)
Araçatuba	-21,2115	-50,4261	355
Igarapava	-20,0382	-47,7479	584
Sertãozinho	-21,1434	-48,0070	556
Araraquara	-21,7946	-48,1766	675
Barra Bonita	-22,4994	-48,5523	468
Piracicaba	-22,7250	-47,6476	524
Elias Fausto	-23,0435	-47,3743	584
Maracaí	-22,6142	-50,6687	421
Batatais	-20,8916	-47,5856	870
Lins	-21,6794	-49,7435	419
Jaboticabal	-21,2554	-48,3224	615
Araras	-22,3577	-47,3849	640
Ourinhos	-22,9831	-49,8569	483
Motuca	-21,5083	-48,1534	615
Buritizal	-20,1913	-47,7051	843
Santa Rosa do Viterbo	-21,4703	-47,3621	667
Ariranha	-21,1895	-48,7882	587
Itapira	-22,4363	-46,8222	639
Pradópolis	-21,3565	-48,0633	529
Patrocínio Paulista	-20,6391	-47,2821	742
Orindiuva	-20,1795	-49,3487	440
Riberão Preto	-21,1767	-47,8208	531
Brotas	-22,2844	-48,1275	645
Votuporanga	-20,4203	-49,9783	499
Olímpia	-20,7374	-48,9155	518

A essa frequência (FII>2%) se denominou de risco de infestação da broca-da-cana, o qual foi classificado da seguinte forma: Baixo (≤ 30%); Moderado (30,1 a 60%); Alto (60,1 a 90%); e Muito Alto (> 90%).

Os dados de FII>2% desses 25 locais foram correlacionados, por meio de uma regressão linear múltipla, com as coordenadas geográficas (latitude – Lat e longitude – Long) e com a altitude (Alt), de modo a gerar um modelo linear para a espacialização dessa variável:

$$\mathbf{FII>2\% = \mathit{a} + \mathit{b}.Lat + \mathit{c}.Long + \mathit{d}.Alt + \mathit{e}.Lat.Long + \mathit{f}.Lat.Alt + \mathit{g}.Long.Alt + \mathit{h}.Lat^2 + \mathit{i}.Long^2 + \mathit{j}.Alt^2}$$

em que: *a* é o coeficiente linear da equação de regressão múltipla, enquanto os coeficientes *b* a *j* representam os coeficientes angulares da regressão, respectivamente, para cada uma das variáveis consideradas (Lat, Long e Alt) e para o produto ou quadrado delas.

Para a confecção do mapa de risco de infestação da broca-da-cana empregou-se o mesmo procedimento adotado por Yamada e Sentelhas (2014) Monteiro e Sentelhas (2014) e Sentelhas et al. (2016). Primeiramente, no sistema de informação geográfica ARCGIS 10.6.1, os mapas de latitude (Lat) e longitude (Long), juntamente com o de altitude (Alt), este último obtido das imagens do *Shuttle Radar Topography Mission* (SRTM), foram confeccionados. Por meio da ferramenta *Spatial Analyst – Options* e *Spatial Analyst – Raster Calculator* transformou-se as equações de Lat e Long em mapas, que juntamente com o mapa de Alt, proveniente do SRTM, permitiu-se empregar a álgebra de mapas para o cálculo da FII>2% por meio do modelo linear acima apresentado e da ferramenta *Spatial Analyst – Raster Calculator*. Com isso foi possível se calcular o FII>2% para cada pixel de 90 × 90 m, o que gerou assim o mapa temático da referida variável para o estado de São Paulo.

O desempenho do modelo linear, e consequentemente do mapa temático, foi avaliado considerando-se os dados independentes de 10 localidades bem distribuídas no estado de São Paulo (Tabela 2).

Tabela 2. Localidades utilizadas para a validação do modelo linear para espacialização do risco de infestação de broca-da-cana no Estado de São Paulo.

Local	Lat	Long	Alt (m)
Assis	-22,6610	-50,3995	560
Franca	-20,5418	-47,4197	1002
Palestina	-20,3916	-49,4327	543
Santa Fé do Sul	-20,2119	-50,9271	386
Itapeva	-23,9676	-48,9010	684
Andradina	-20,8939	-51,3837	401
Jaú	-22,2963	-48,5587	530
Pirassununga	-21,9960	-47,4268	625
José Bonifácio	-21,0396	-49,6910	430
Mococa	-21,4708	-47,0006	594

Assim, comparou-se os valores de FII>2% gerado pelo mapa com os valores calculados de acordo com o modelo de MD de melhor desempenho. Os índices estatísticos empregados nesta análise foram os mesmos citados anteriormente para avaliar o modelo de MD de melhor desempenho, ou seja: R^2; r; D RMSE; EAM; EM e C.

3.8. Distribuição do conhecimento

A distribuição do conhecimento será feita por meio da divulgação dos resultados deste estudo junto aos produtores, consultores e empresas sucro-energéticas, que necessitam de informações para um melhor manejo da broca-da-cana no estado de São Paulo. A partir do mapa de risco de infestação da broca-da-cana confeccionado neste estudo, os principais envolvidos com a produção de cana-de-açúcar no estado de São Paulo terão uma melhor noção das áreas mais sujeitas à ocorrência dessa praga, auxiliando assim no seu manejo racional.

4. RESULTADOS E DISCUSSÃO

A Figura 8 apresenta a matriz de correlação de Pearson para todas as variáveis envolvidas no presente estudo, elaborada pelo software R em que o grau de correlação obedece a escala de cor, sendo azul para as correlações positivas e vermelha para as negativas.

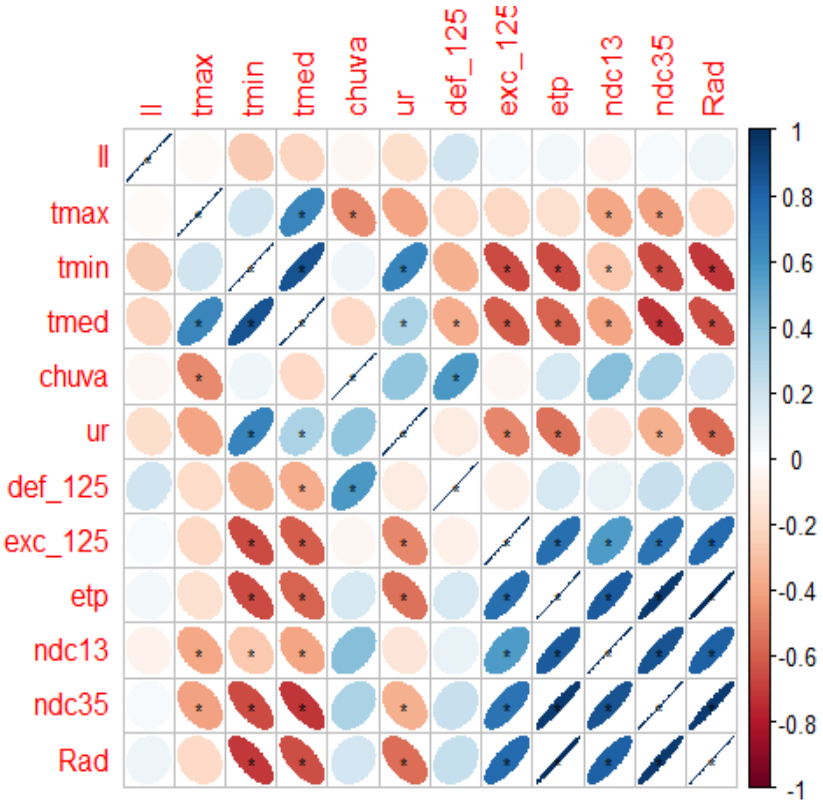


Figura 8. Diagrama de correlações entre as variáveis meteorológicas, de balanço hídrico independentes e o índice de infestação da broca da cana-de-açúcar à nível de significância de 5% ($p<0,05$). Variáveis: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125). *Valores de r significativos a um nível de 5% de probabilidade pelo teste t.

Como pode-se notar na Figura 8, as maiores correlações encontradas com o índice de infestação (II) foram para as variáveis independentes: tmin, com correlação negativa; e def_125, com correlação positiva. Isso significa que à medida que a temperatura mínima aumenta o índice de infestação final diminuiu, enquanto para o déficit hídrico, quanto maior essa variável, maior o índice de infestação final da praga. Em seguida, com correlações menores, mas relativamente maiores que as demais variáveis, aparecem a Rad, etp, ndc35 e exc_125. Observa-se que mesmo sem correlação linear com interação significativa ao nível estatístico para a variável II, é possível explicar por meio da abordagem dos modelos uma porcentagem significativa da variabilidade de II, o que demonstra a complexidade da interação entre essas variáveis.

Como observado na Tabela 3, a correlação mais negativa entre o II e as variáveis meteorológicas se deu com a tmin ($r = -0,255$), seguida pelas variáveis tmed, ur, ndc13, chuva e tmax, também com correlações negativas.

Isso significa que à medida que essas variáveis aumentam, há redução do II. Mesmo sabendo-se que a Correlação de Pearson não é um indicativo da relação causa-efeito entre duas variáveis, e sim, da intensidade da relação entre elas, pode-se inferir que essas variáveis demostram ter papel importante na ocorrência da broca-da-cana. Para as variáveis que apresentam correlação positiva com II, o def_125 foi o que apresentou maior correlação (r = 0,196), seguido pela Rad, etp, ndc35, exc_125.

Tabela 3. Coeficientes de correlação de Pearson (r) das correlações das variáveis meteorológicas e de balanço hídrico independentes com o índice de infestação final da broca da cana-de-açúcar (II). Variáveis: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125).

Variável	r
tmin	-0,255
tmed	-0,210
ur	-0,180
ndc13	-0,067
chuva	-0,043
tmax	-0,027
def_125	0,196
Rad	0,077
etp	0,059
ndc35	0,037
exc_125	0,036

A importância das variáveis meteorológicas para a ocorrência da broca-da-cana vem sendo destacada em diversos estudos. De acordo com Silveira Neto (1976) há picos da população de mariposas nos períodos mais úmidos, já que tal condição afeta os processos de oviposição, fisiologia, desenvolvimento e longevidade dos insetos. Parra et al. (1999) estudando a viabilidade dos ovos e posturas de *D. saccharalis* observaram que a viabilidade reduz muito em umidades menores que 60% e menor viabilidade (eclosão) em ambientes totalmente secos. Esses ovos sofrem muito em condições de baixa umidade pois tem aerópilas, que para obter oxigênio do ar os expõem a perda de água. Segundo Carbognin (2016), a maior ocorrência de adultos se dá no período das chuvas, principalmente, entre janeiro e abril. Apesar disso, esses autores não identificaram correlação entre os fatores climáticos e a flutuação populacional da broca-da-cana para as localidades de Pradópolis e Valparaíso, enquanto em Quirinópolis, onde a umidade foi fator abiótico limitante para o crescimento populacional desta espécie, certa correlação foi observada. Além da correlação positiva ($p<0,05$) entre a infestação de broca e as variáveis meteorológicas em Quirinópolis, o estudo também identificou correlação positiva ($p<0,01$) entre a umidade e o número de mariposas capturadas.

Embora esses autores retratem a importância da umidade para os insetos (BERTELS, 1970; BOTELHO et al., 1977; MELO et al., 1984), de modo geral, este fator não foi significativo para a maioria das áreas. Da mesma forma, as demais variáveis ambientais não apresentaram correlação, inclusive a temperatura, embora, saiba-se que esta é fundamental para o desenvolvimento dos insetos (HADDAD et al., 1999) e interfere no número de gerações de *D. saccharalis* (MELO; PARRA, 1988b). Segundo Teran (1979), os fatores edáficos e biológicos, também, são responsáveis por alterar a dinâmica populacional da broca-da-cana. Botelho et al. (1977) estudaram oito variáveis

meteorológicas (amplitude térmica, temperaturas mínimas e máximas do ar, chuva, velocidade do vento, umidade relativa média e temperaturas máximas e mínimas do solo) e verificaram essas explicaram cerca de 41% da dinâmica populacional de *D. saccharalis*.

4.1. Análise da Componentes Principais

A Figura 9 apresenta a análise de componentes principais, a qual foi empregada para identificar dentre as variáveis meteorológicas e de balanço hídrico analisadas quais estão mais relacionadas à intensidade de infestação final de broca-da-cana (II).

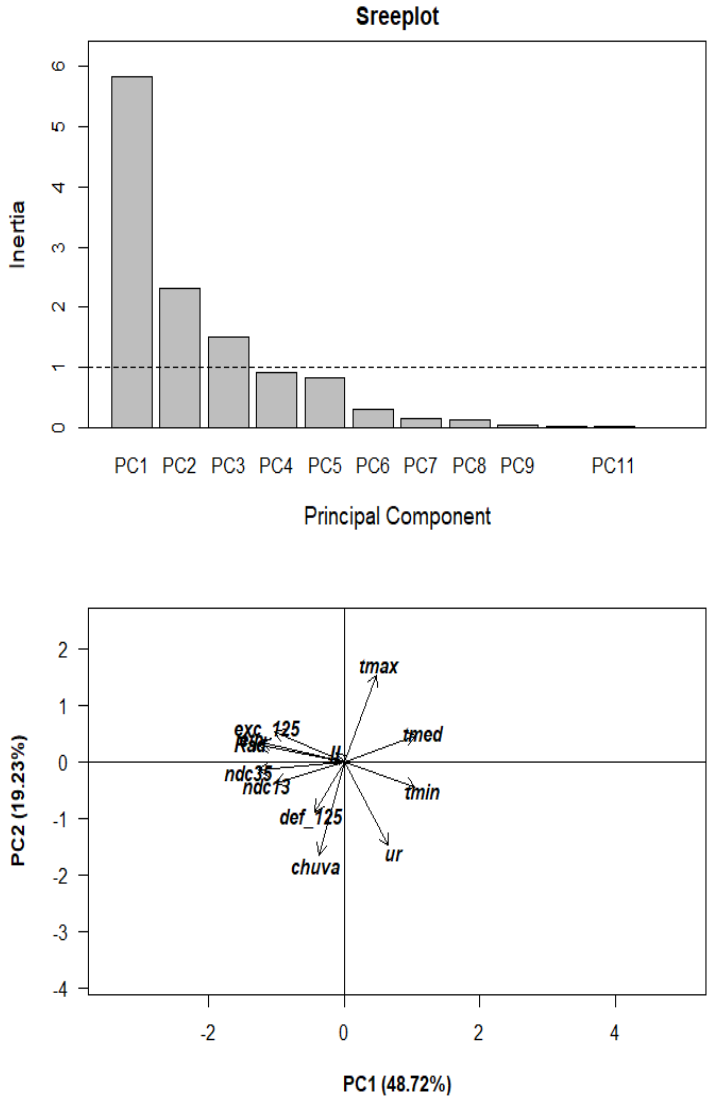


Figura 9. Diagrama *biplot* de componentes principais entre as variáveis meteorológicas, de balanço hídrico e o índice de infestação da broca da cana-de-açúcar. Variáveis: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125).

Por meio da Figura 9, observa-se que a primeira componente principal explica 48,72% da variabilidade dos dados e a segunda componente explica 19,23%, totalizando aproximadamente 68% da relação entre as variáveis, o que demonstra a importância delas na definição do II. A intercorrelação entre as variáveis e a intensidade de infestação é demonstrada pelos ângulos entre as setas e a importância dessas variáveis pelo comprimento das setas. O grau com que a variável se correlaciona com o eixo é determinado pelo ângulo formado entre determinada seta e o eixo de ordenadas (MELÉM et al., 2008; LYRA et al., 2010). Assim, observa-se que as variáveis que estão no mesmo quadrante e, consequentemente, no mesmo sentido correlacionam-se positivamente, e que as variáveis que estão em sentidos inversos correlacionam-se negativamente. Por exemplo, as principais variáveis com maior relação com o II % são ndc35, Rad, etp, exc_125 e ndc13, já que essas estão no mesmo sentido e, portanto, correlacionadas positivamente. As variáveis tmin, tmed, ur e tmax apresentaram relação negativa com II. As demais variáveis apresentam um grau menor de relação, o que demonstra menor importância delas na definição do II. Na Tabela 4 são apresentados os valores absolutos das componentes principais para cada variável, seus autovalores e a porcentagem de variância explicada. As informações contidas na Tabela 4 confirmam o que já foi explicado acima, ou seja, que as variáveis com sinais opostos sendo antagônicas e as com sinais de mesmo sentido expressam uma ação conjunta na variável resposta. Dessa forma, observa-se o grande peso das seguintes variáveis, que expressam ação conjunta: ndc35; Rad; etp; exc_125; e ndc13. Já as variáveis que apresenta uma relação antagônica são: tmin; tmed; ur; e tmax.

Tabela 4. Correlação entre cada componente principal (PC1, PC2) e entre cada variável do estudo entre índice de infestação final da brica da cana-de-açúcar (II) e as variáveis meteorológicas: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125).

	PC1	**PC2**
Auto valores	2.39	1.48
Variância explicada (%)	48.72	19.23
Correlações		
II	-0,12	0,10
chuva	-0,26	-0,83
def_125	-0,32	-0,43
tmax	0,39	0,7
ur	0,49	-0,74
ndc13	-0,76	-0,2
exc_125	-0,76	0,25
tmed	0,79	0,18
tmin	0,79	-0,23
etp	-0,93	0,15
Rad	-0,96	0,13
ndc35	-0,96	-0,08

4.2. Análise da intensidade da infestação de broca-da-cana por variedade

A Figura 10 apresenta o II da broca-da-cana para cada uma das 30 variedades consideradas neste estudo. Observa-se que para esse grupo de variedades há uma grande variação para uma mesma variedade entre as localidades avaliadas, os campos amostrados e as diferentes safras. Observa-se para os 2876 dados do estudo, que para as 30 variedades de cana-de-açúcar de maior importância há variações no II, o que indica que os materiais avaliados apresentam diferentes susceptibilidades à broca, o que dificulta o manejo. Quanto às variedades resistentes, diversos trabalhos foram realizados no Brasil, como os de Gallo (1954), Amaral e Arruda (1964) e Macedo et al. (1977), demonstrando que algumas variedades apresentam menor II que outras. Derneika e Lara (1991) concluíram que as variedades SP71-345, SP71-6113 e SP71-1081 comportaram-se como moderadamente resistentes à broca, enquanto as variedades CP51-22, SP71-3146 e SP71-5574 foram altamente suscetíveis.

Utilizar variedades resistentes no campo é uma alternativa bastante viável para a redução no controle químico, menor problemas com as pragas e uma boa combinação com diferentes formas de manejo, como por exemplo o controle biológico. Porém foi observado que a resistência de insetos e o controle biológico podem apresentar conflitos, o que demanda um esforço adicional para estudos destas práticas a campo. Temos alguns exemplos de resistência extrínseca e de inter-relações negativas observados por Boethel e Eikenbary (1986) e Lara (1991).

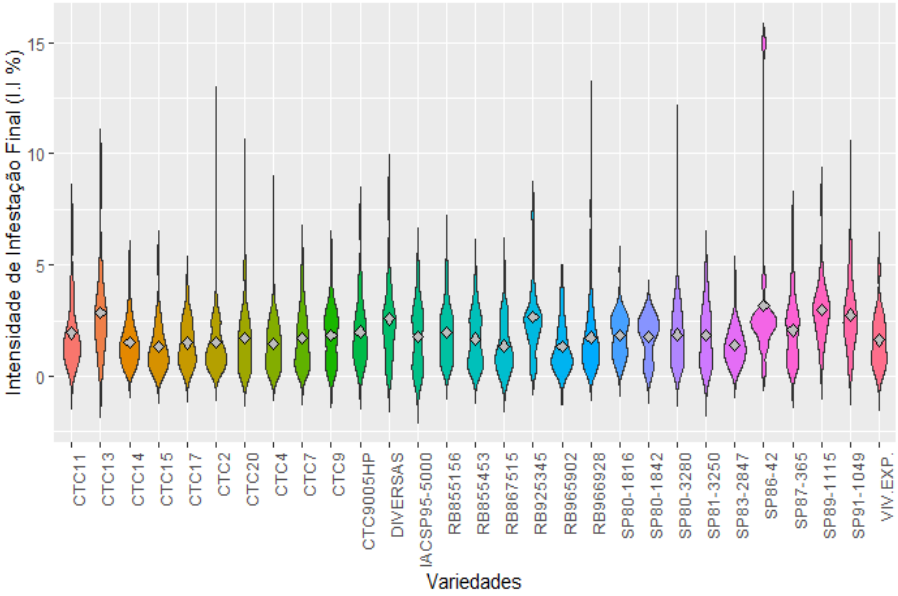


Figura 10. Intensidade de infestação final (II) da broca da cana-de-açúcar para as 30 variedades consideradas neste estudo nas diferentes localidades, campos amostrados e safras.

Na Figura 11 observa-se a evolução do II ao longo das oito safras (2012-2013 a 2019-2020) e nove localidades consideradas. Cada uma dessas localidades representam uma área de aproximadamente 10.000 hectares com diversas fazendas que compõem a região. Observa-se também que devido à baixa quantidade de dados, a região de Monte Alto foi retirada da análise e modelagem, pois os dados desta cidade se mostraram insuficientes para gerar a variabilidade exigida em estudo de mineração de dados. Para todas as localidades foi possível se observar uma ampla variação no II, com tendência de valores mais elevados no início e final das safras analisadas, ficando as safras

intermediárias com valores mais baixos. Tal comportamento diferenciado ocorre devido às diferenças existentes entre os locais, principalmente, das condições meteorológicas.

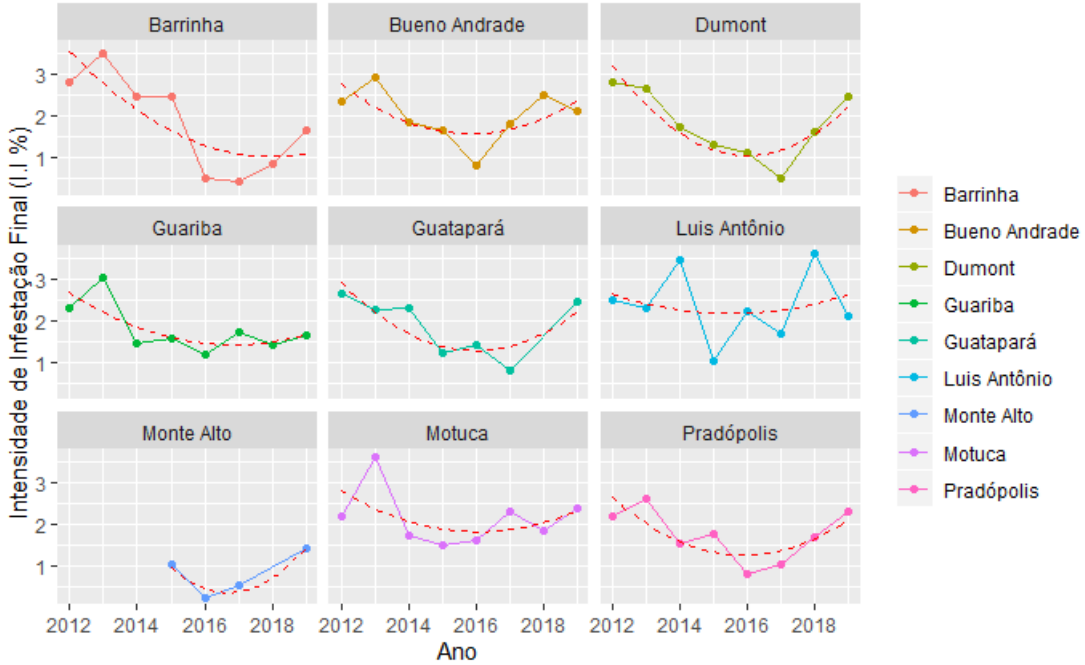


Figura 11. Intensidade de infestação final (II) da broca da cana-de-açúcar nas diferentes localidades e safras (2012-2013 a 2019-2020).

4.3. Seleção e descrição das variáveis empregadas na mineração de dados

4.3.1. Intensidade da infestação final da broca da cana-de-açúcar

Observa-se na Figura 12 que a intensidade da infestação final da broca da cana-de-açúcar (II) variou entre as localidades analisadas, com algumas delas apresentando amplitude bastante elevada, com destaque para Barrinha, Guatapará e Luis Antônio. Quanto à distribuição interanual, observa-se também grande diversidade dos dados e frequência com maior variação no II para as safras 2012-2013, 2014-2015 e 2017-2018. As safras 2012-2013 e 2013-2014 apresentaram um índice médio de II maior do que nas demais safras, fato este ocorrido principalmente devido às poucas chuvas que ocorreram nestes anos. É notável que o II tem grande dependência das variáveis climáticas, sendo essas a principal fonte de variação entre as localidades e safras analisadas. Ainda na Figura 12, observa-se que o II apresenta uma distribuição assimétrica, na qual prevalecem as infestações mais baixas, entre 0 e 3%.

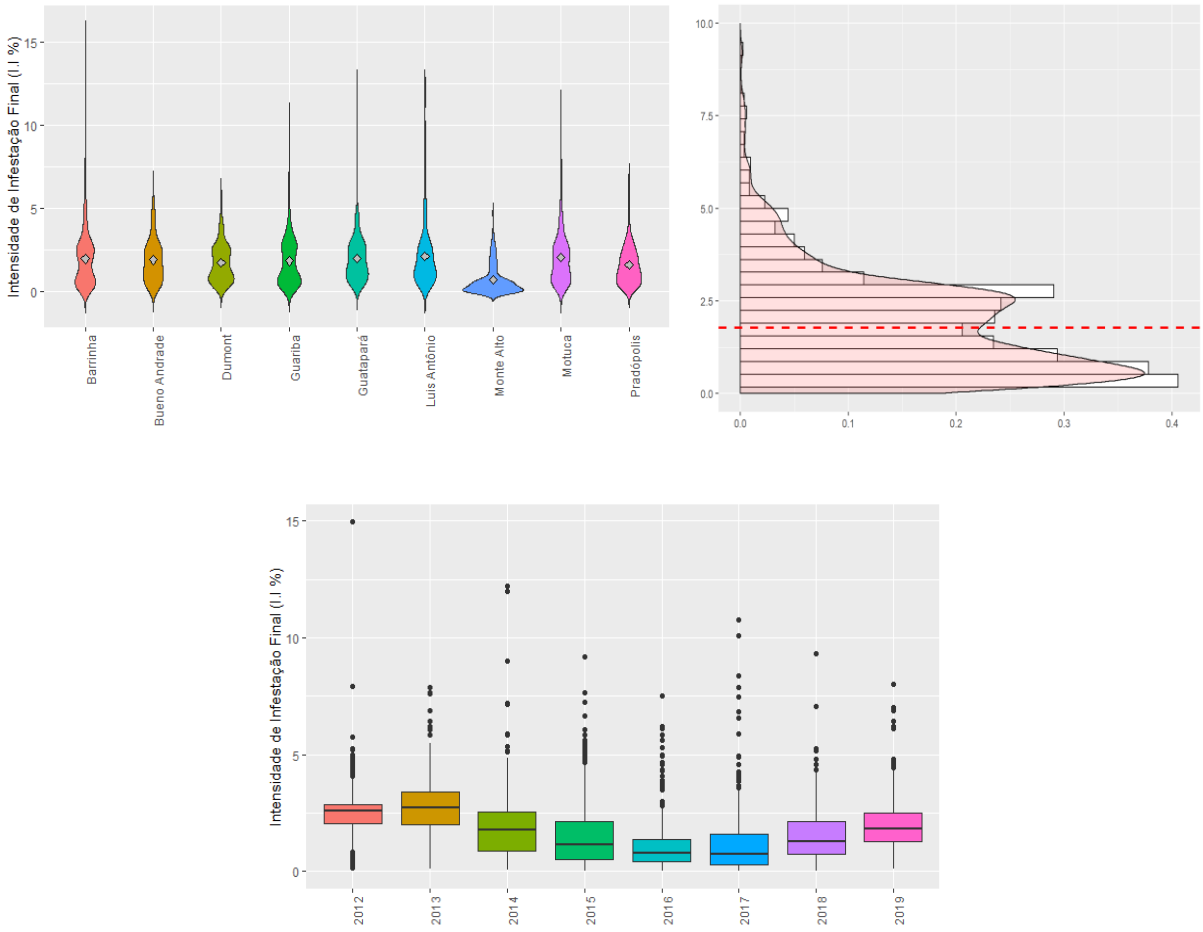


Figura 12. Intensidade de infestação final (II) da broca da cana-de-açúcar nas localidades e safras avaliadas, no estado de São Paulo.

4.3.2. Temperatura Máxima

Os dados da Figura 13 apresentam a variação da temperatura máxima (tmax) nas localidades e safras em que se avaliou o índice de infestação da broca da cana-de-açúcar no estado de São Paulo. Observa-se que a tmax variou entre as localidades e as safras, sendo Barrinha, Guariba e Pradópolis aquelas com maior variação. Quanto à distribuição anual, observa-se que as safras de 2018-19 e 2019-2020 forma as que apresentaram as maiores temperaturas máximas. Nos demais anos, os valores foram próximos, sendo a safra 2017-2018 aquela com o menor valor. Ainda na Figura 13, observa-se que a tmax apresenta distribuição normal, com mediana de 29,4 ºC na região analisada.

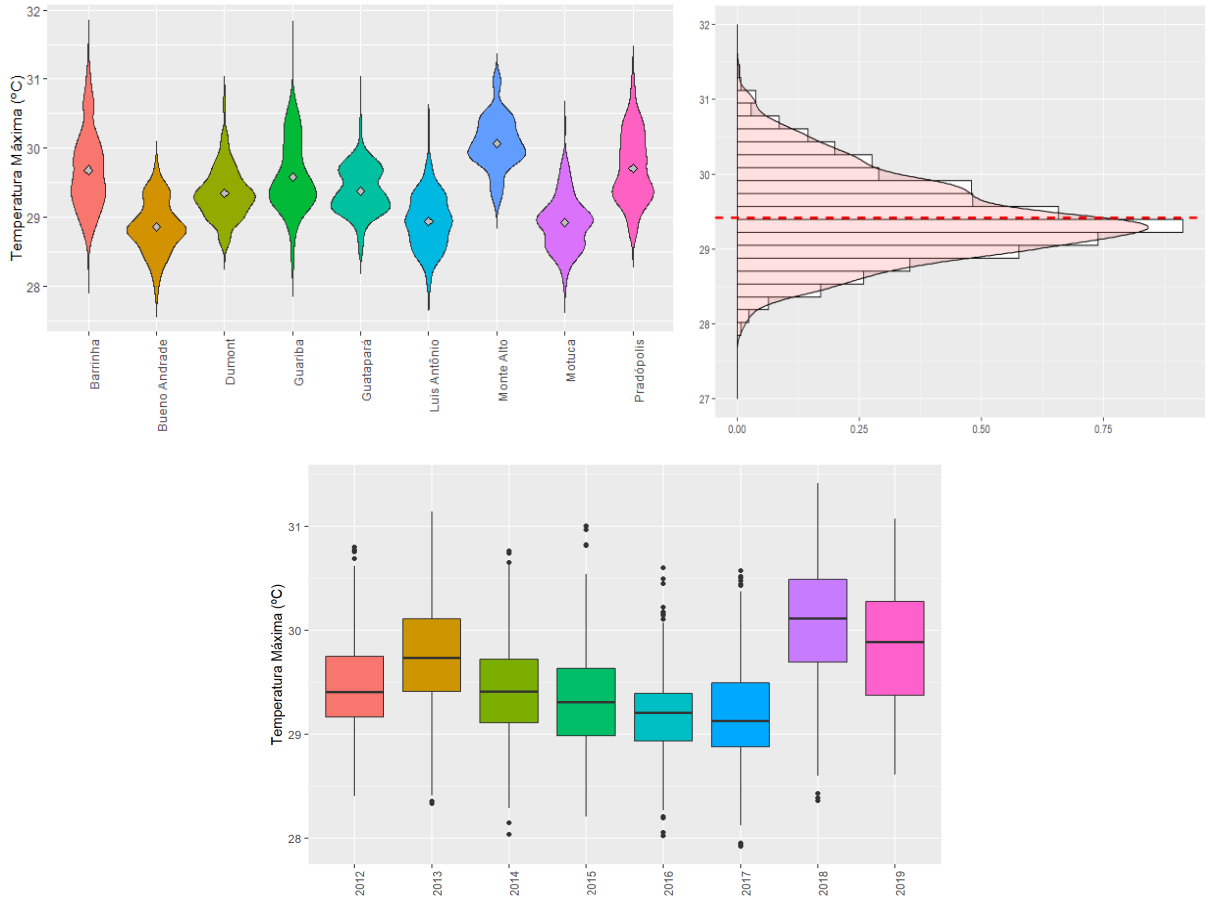


Figura 13. Temperatura máxima nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.3. Temperatura Mínima

Os dados da Figura 14 apresentam a variação da temperatura mínima (tmin) nas localidades e safras em que se avaliou o índice de infestação da broca da cana-de-açúcar no estado de São Paulo. Observa-se que a tmin variou entre as localidades e as safras, tanto em termos da média como da variação interanual. Dentre as localidades analisadas, Monte Alto é a que apresentou a maior mediana, porém a menor variação interanual. No entanto, como esta localidade apresentava um menor volume de dados ela foi desconsiderada da fase de modelagem. Por outro lado, em Dumond foi onde ocorreu a maior variação interanual da tmin, porém com a tmin média sendo muito próxima das demais localidades. Dentre as safras analisadas, as de 2015-2016 e de 2016-2017 foram as que apresentaram os maiores valores, seguidas pela safra 2019-2020. Por outro lado, na safra 2012-2013 a tmin foi a mais baixa dentre os anos analisados. Ainda na Figura 14, observa-se que a tmin apresenta distribuição normal, porém com dois picos de frequência, sendo um em torno de 16,3 ºC e outro em torno de 17,5 ºC.

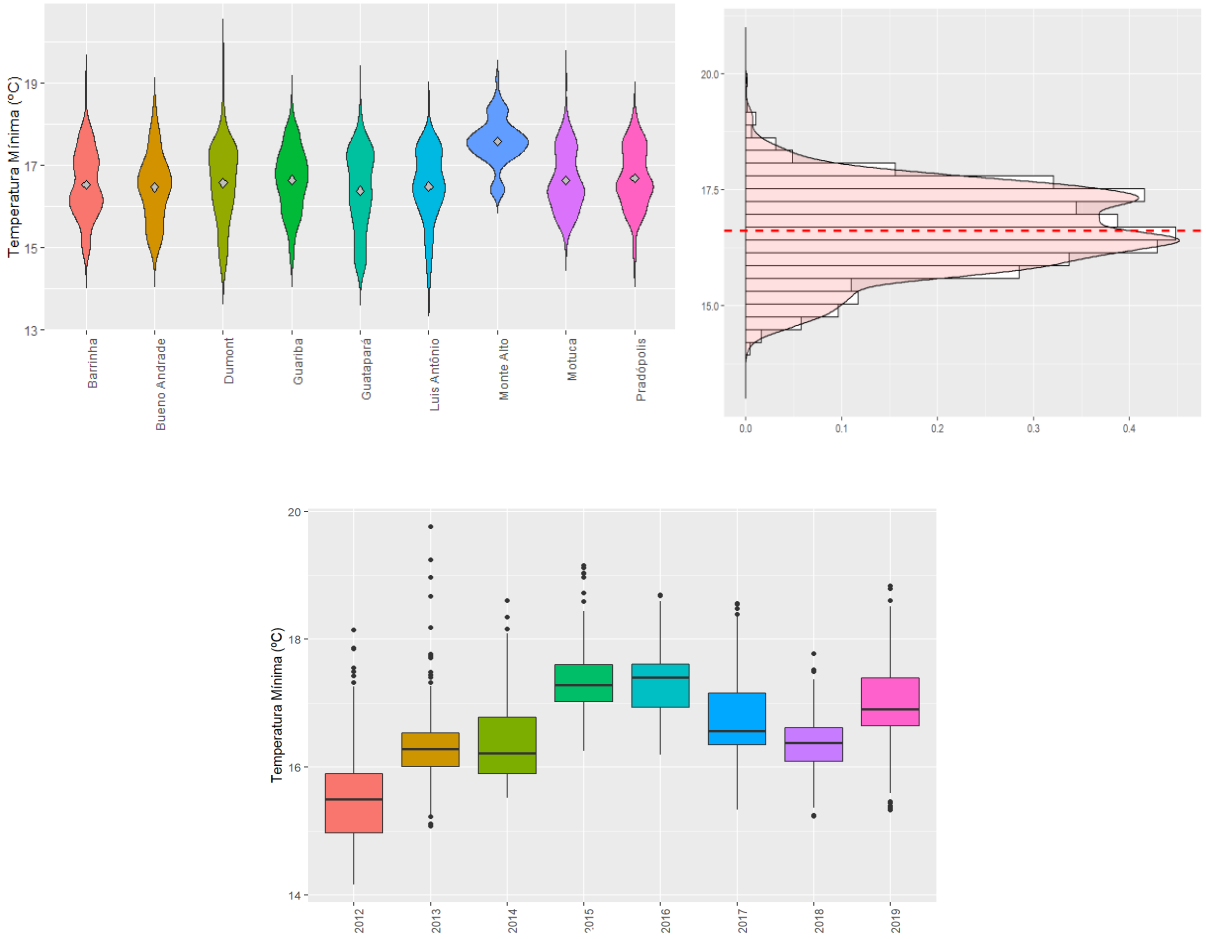


Figura 14. Temperatura mínima nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.4. Temperatura média

Os dados da Figura 15 apresentam a variação da temperatura média (tmed) nas localidades e safras em que se avaliou o índice de infestação da broca da cana-de-açúcar no estado de São Paulo. Observa-se que a tmed, assim como a tmax e tmin, variou entre as localidades e as safras, tanto em termos da média como da variação interanual. Dentre as localidades analisadas, Monte Alto, em razão principalmente da tmin, foi a que apresentou a maior mediana, porém a menor variação interanual. No entanto, como esta localidade apresentava um menor volume de dados ela foi desconsiderada da fase de modelagem. Por outro lado, em Dumond foi onde ocorre a maior variação interanual da tmed, porém com o valor médio sendo muito próximo dos apresentados pelas demais localidades. Dentre as safras analisadas, as de 2015-2016, 2016-2017 e 2019-2020 foram aquelas com os maiores valores médios, enquanto na safra 2012-2013 a tmed foi a mais baixa dentre os anos analisados. Ainda na Figura 15, observa-se que a tmed apresenta distribuição normal, com a mediana em torno dos 23,0 ºC.

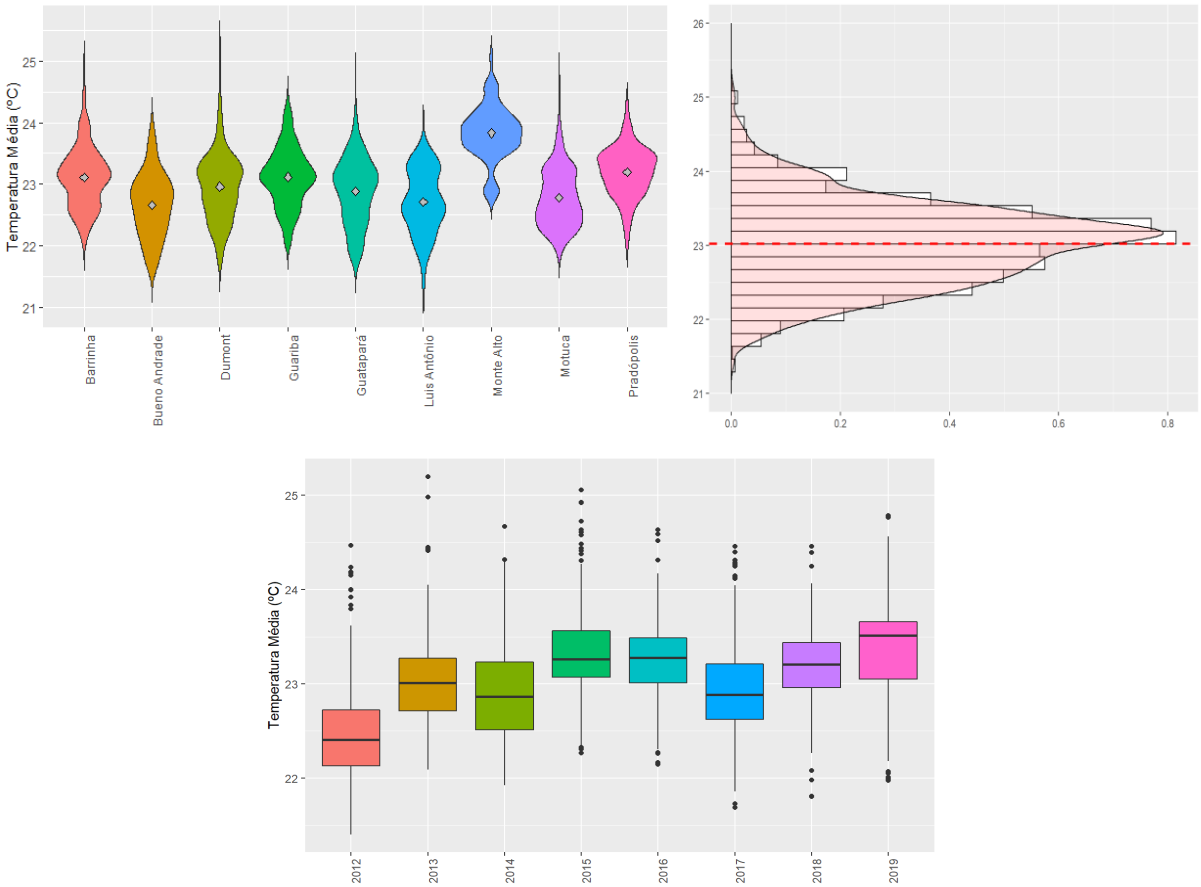


Figura 15. Temperatura média nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.5. Chuva

Os dados da Figura 16 apresentam a variação da chuva anual nas diferentes localidades e safras analisadas. Observa-se uma grande heterogeneidade entre as localidades, com menor variabilidade interanual em Barrinha, Guatapará e Pradópolis. Com relação à variação dos dados entre os anos na região estudada, envolvendo o agrupamento de todas as regiões, observa-se que o ano de 2016 foi o mais chuvoso, com uma mediana acima dos 1600 mm, enquanto 2018 foi o menos chuvoso, com a mediana ficando abaixo dos 1000 mm. Ainda na Figura 16, observa-se que a chuva anual apresenta distribuição normal, ligeiramente assimétrica, com a mediana em torno de 1300 mm.

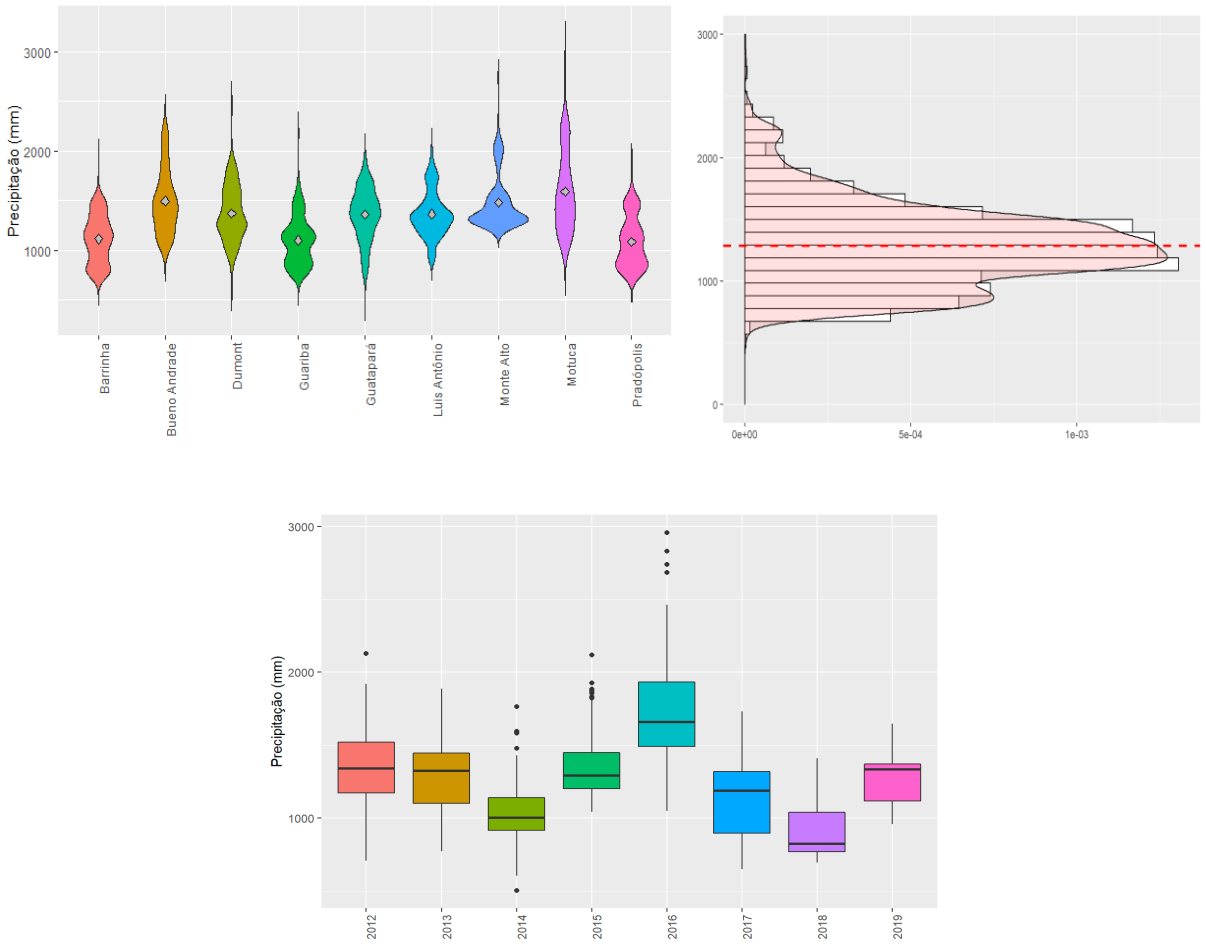


Figura 16. Chuva anual nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.6. Umidade Relativa

Os dados da Figura 17 apresentam a variação da umidade relativa média (ur) nas localidades e safras em que se avaliou o índice de infestação da broca da cana-de-açúcar no estado de São Paulo. Observa-se que a ur, assim como ocorreu para a tmin, apresentou distribuição próxima da normalidade, porém com dois picos por volta de 69 e 73%. As variações entre localidades e temporal foram menos expressivas do que para as demais variáveis. A variabilidade entre safras mostrou uma maior ur no ano de 2016, seguido dos anos de 2017 e 2015, enquanto 2012 foi o ano com a menor ur.

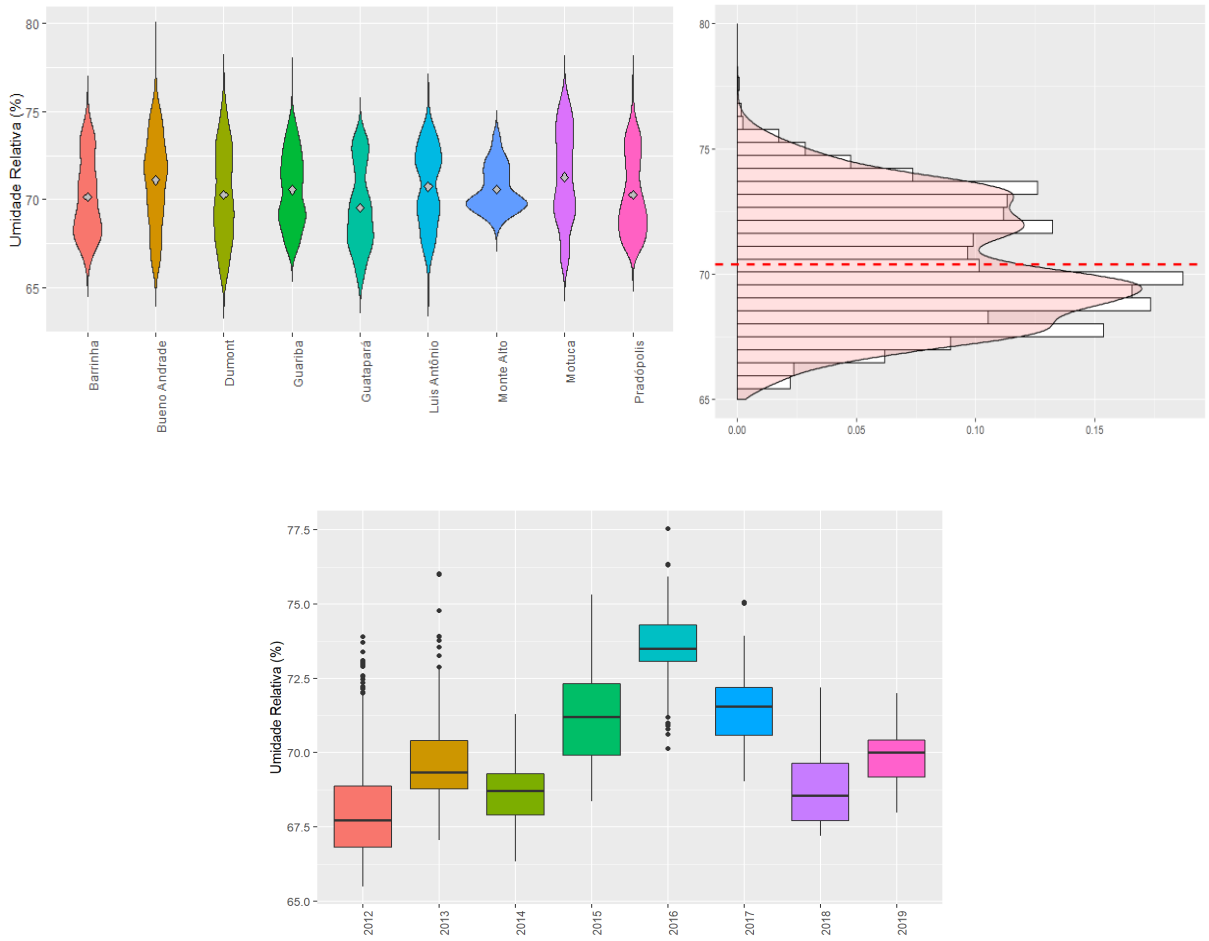


Figura 17. Umidade relativa nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.7. Evapotranspiração Potencial

Os dados da Figura 18 apresentam a variação da evapotranspiração potencial anual (etp) nas diferentes localidades e safras analisadas. Observa-se uma grande homogeneidade entre as localidades, com menor variabilidade interanual em Luiz Antonio e Monte Alto. No entanto, como Monte Alto apresentava um menor volume de dados ela foi desconsiderada da fase de modelagem. Com relação à variação dos dados entre os anos na região estudada, também se observa pouca variação nos valores de etp, havendo, no entanto, uma mediana um pouco mais elevada no ano de 2012 e menor no ano de 2016, com valores oscilando entre 1250 e 1500 mm. Ainda na Figura 18, observa-se que a etp anual apresenta distribuição normal, com pico máximo em torno dos 1300 mm.

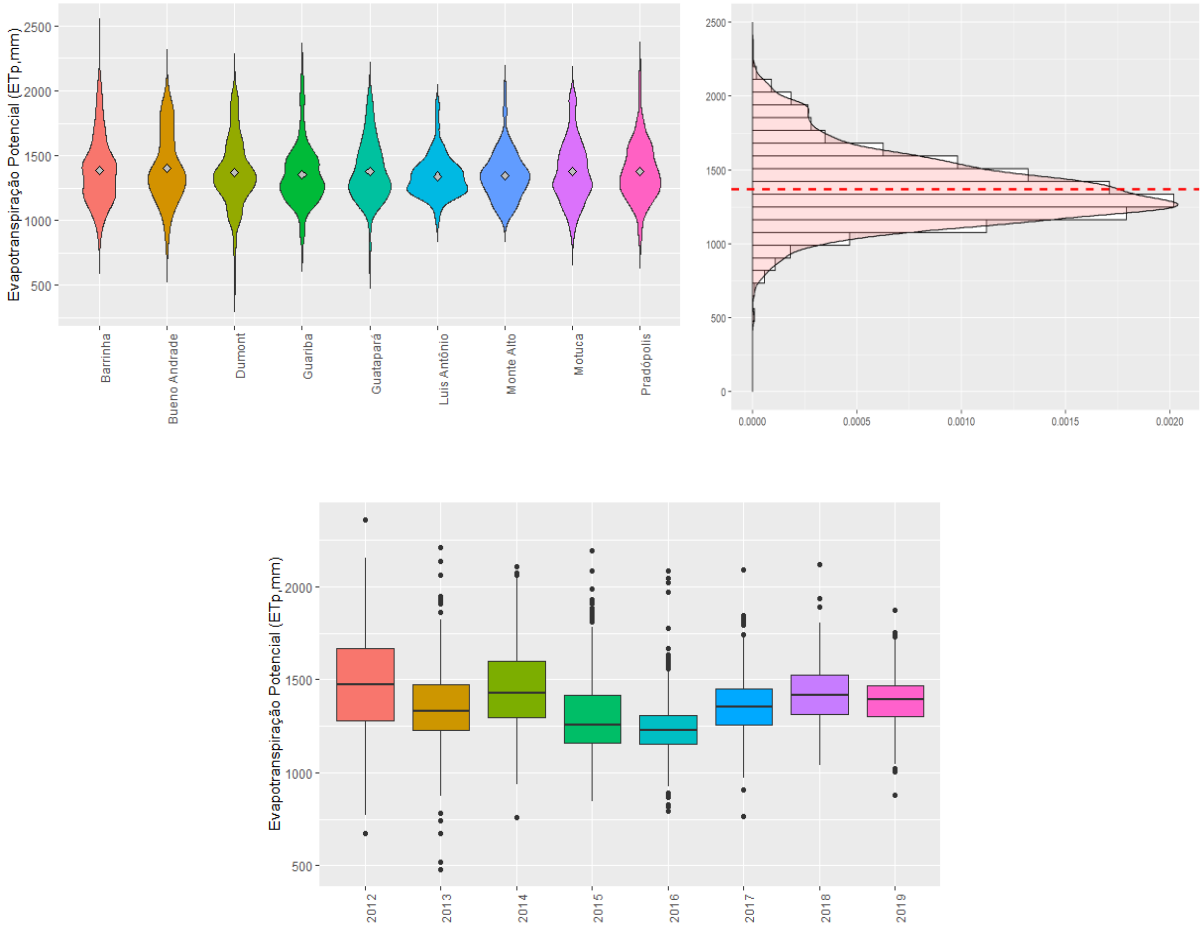


Figura 18. Evapotranspiração potencial nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.8. Número de dias com temperatura mínima ≥ 13 ºC

Observa-se na Figura 19 que o número de dias com temperatura mínima ≥ 13 ºC (ndc13) dentro do ciclo da cana apresentou uma distribuição similar entre as localidades e anos analisados, com um valor médio em torno dos 300 dias. Áreas com ciclo maior correspondem ao período em que o canavial permaneceu no campo até o momento do corte. Dentre os anos analisados, os que apresentaram maior ndc13 foram os anos de 2014, 2015 e 2016. Quanto à distribuição apresentada pelos dados de ndc13, essa foi tipicamente normal, com um ligeiro desvio, sendo a classe de maior frequência um pouco da mediana.

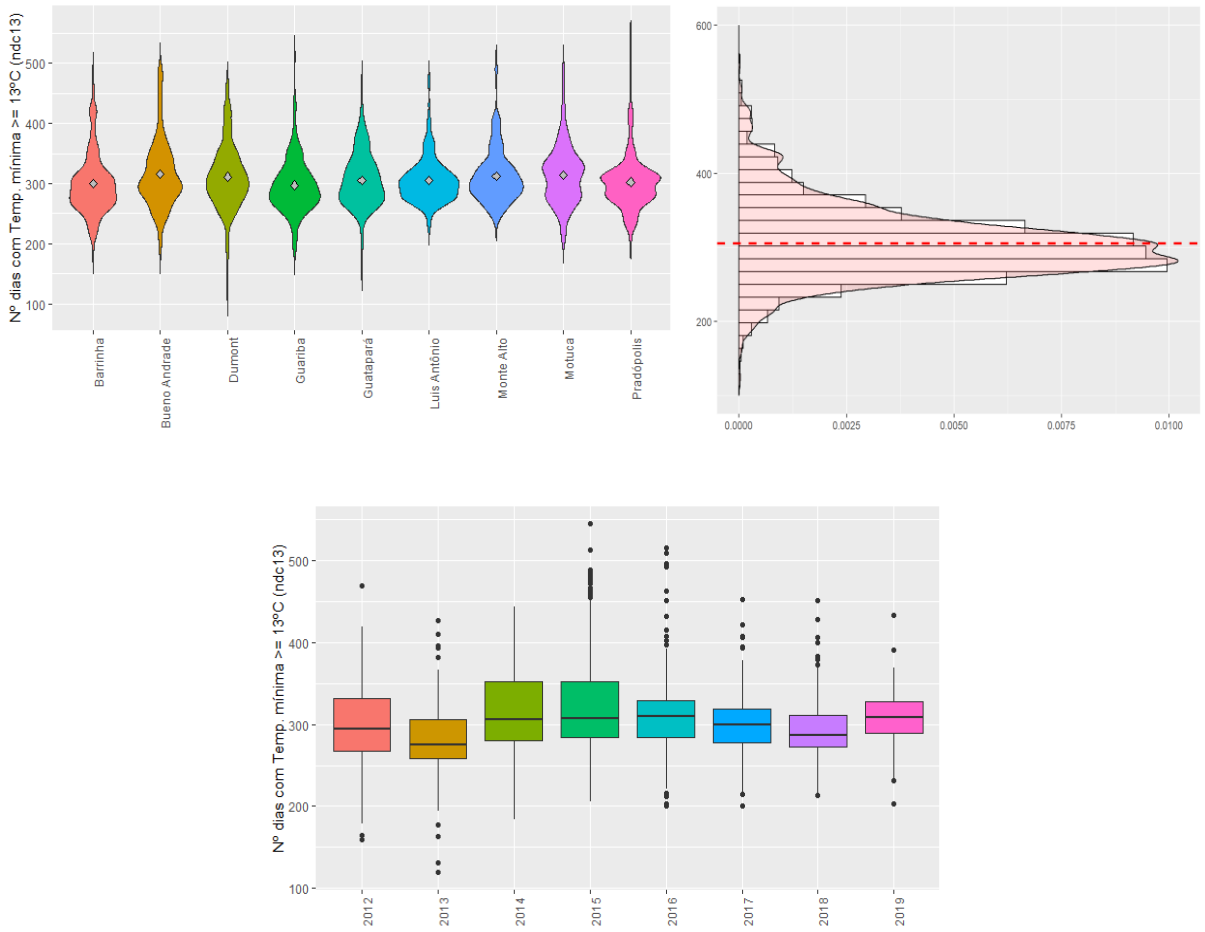


Figura 19. Número de dias com temperatura mínima ≥ 13 ºC nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.9. Número de dias com Temperatura máxima ≤ 35 ºC

Observa-se na Figura 20 que o número de dias com temperatura máxima ≤ 35 ºC (ndc35) dentro do ciclo da cana apresentou uma distribuição similar entre as localidades e anos analisados, com um valor médio em torno dos 360 dias. Áreas com ciclo maior correspondem ao período em que o canavial permaneceu no campo até o momento do corte. Esse valor elevado se deve ao fato de que muitos dos canaviais avaliados apresentavam ciclos de 12 a 18 meses. Dentre os anos analisados, os que apresentaram maior ndc35 foram os anos de 2012, 2013 e 2017, porém, apenas ligeiramente acima dos demais. Quanto à distribuição apresentada pelos dados de ndc35, assim como observado para o ndc13, essa foi tipicamente normal, com um ligeiro desvio, sendo a classe de maior frequência um pouco da mediana.

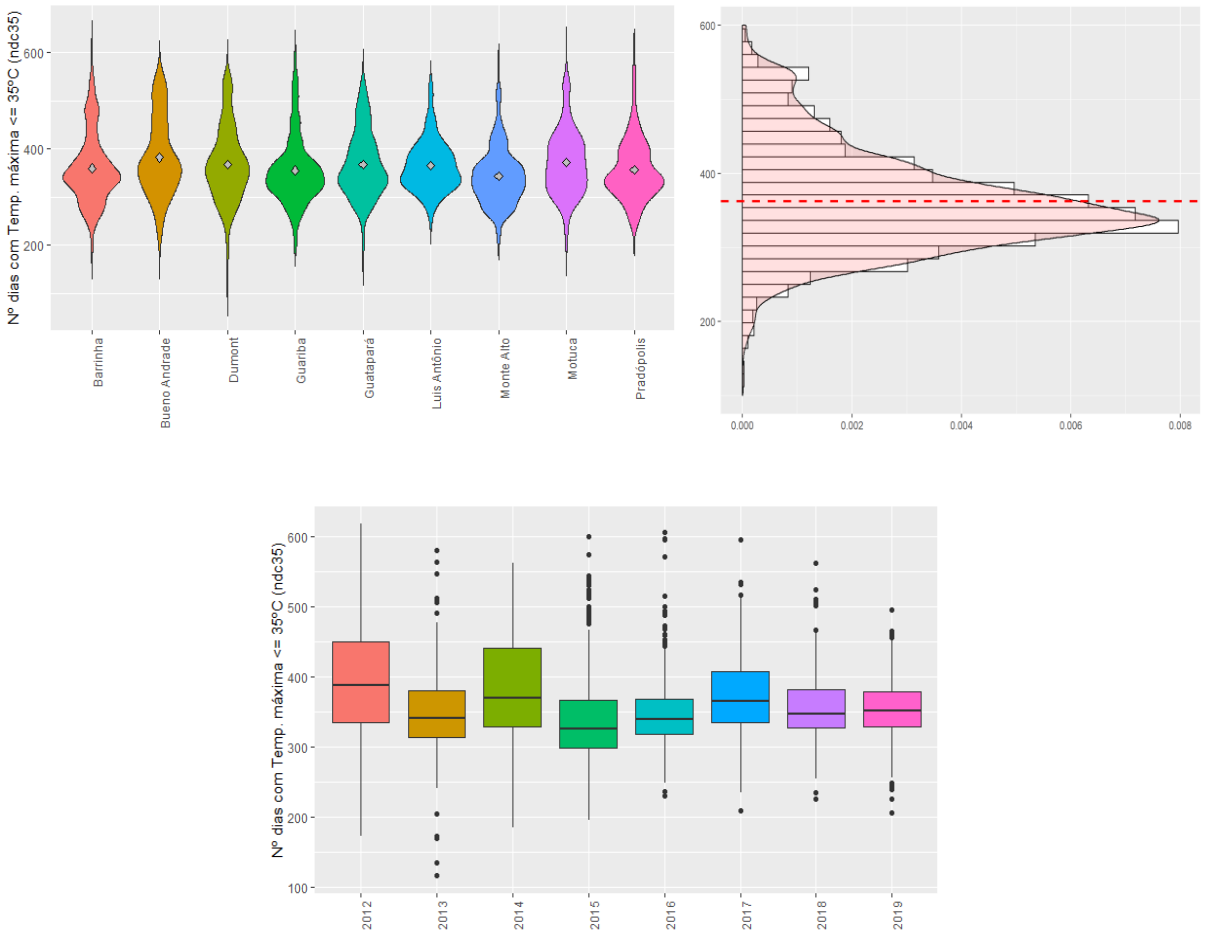


Figura 20. Número de dias com temperatura máxima ≤ 35 ºC nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.10. Radiação solar

A Figura 21 apresenta os valores de radiação solar (Rad) total ao longo do ciclo dos canaviais nas localidades e safras analisadas. Quando avaliada a variação entre localidades, dado o fato que todas essas se encontram na mesma macrorregião, há pouca variação entre elas. No entanto, quando se analisa as diferentes safras, observa-se que há uma maior mediana e variação dos valores nos anos de 2012 e 2013 e menor no ano de 2019. Quanto à distribuição geral, observa-se uma distribuição normal com ligeiro desvio para os valores mais baixos, com a mediana sendo da ordem de 6400 MJ m^{-2} $ciclo^{-1}$.

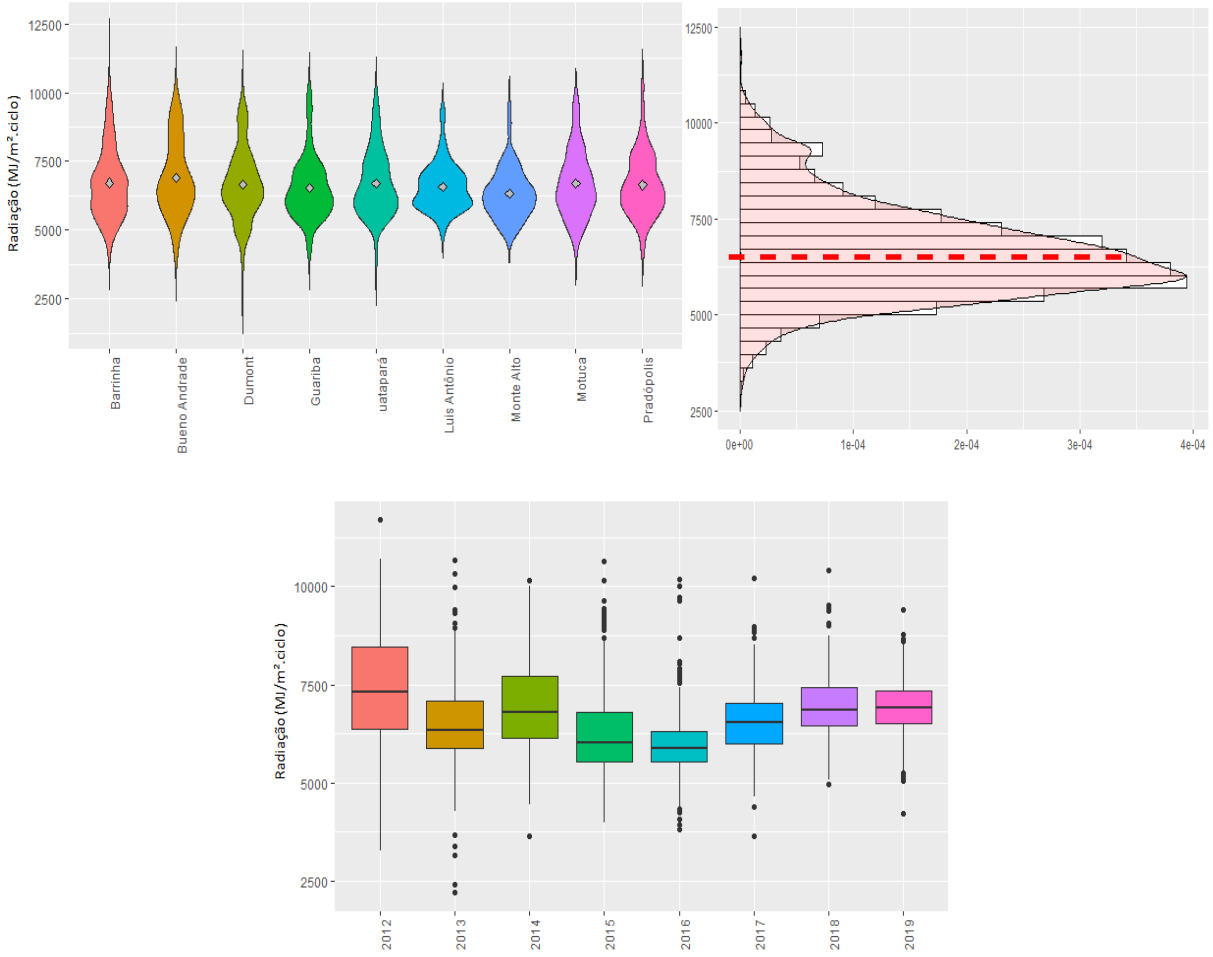


Figura 21. Radiação solar acumulada ao longo do ciclo do canavial nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.11. Déficit Hídrico

A Figura 22 apresenta os valores de déficit hídrico (def_125) total ao longo do ciclo dos canaviais nas localidades e safras analisadas. Quando avaliada a variação entre localidades, observa-se maiores variações para as localidades de Luis Antônio, Monte Alto e Pradópolis com déficit hídrico médio em torno de 380 mm. Quando se analisou as diferentes safras, observou-se grandes variações nos dados, sendo que há uma maior mediana e variação dos valores nos anos de 2012 e 2013 e menor no ano de 2014. Quanto à distribuição geral, observa-se grandes desvios, com picos de déficit hídrico em torno de 200 mm, com a mediana sendo da ordem de 400 mm no ciclo.

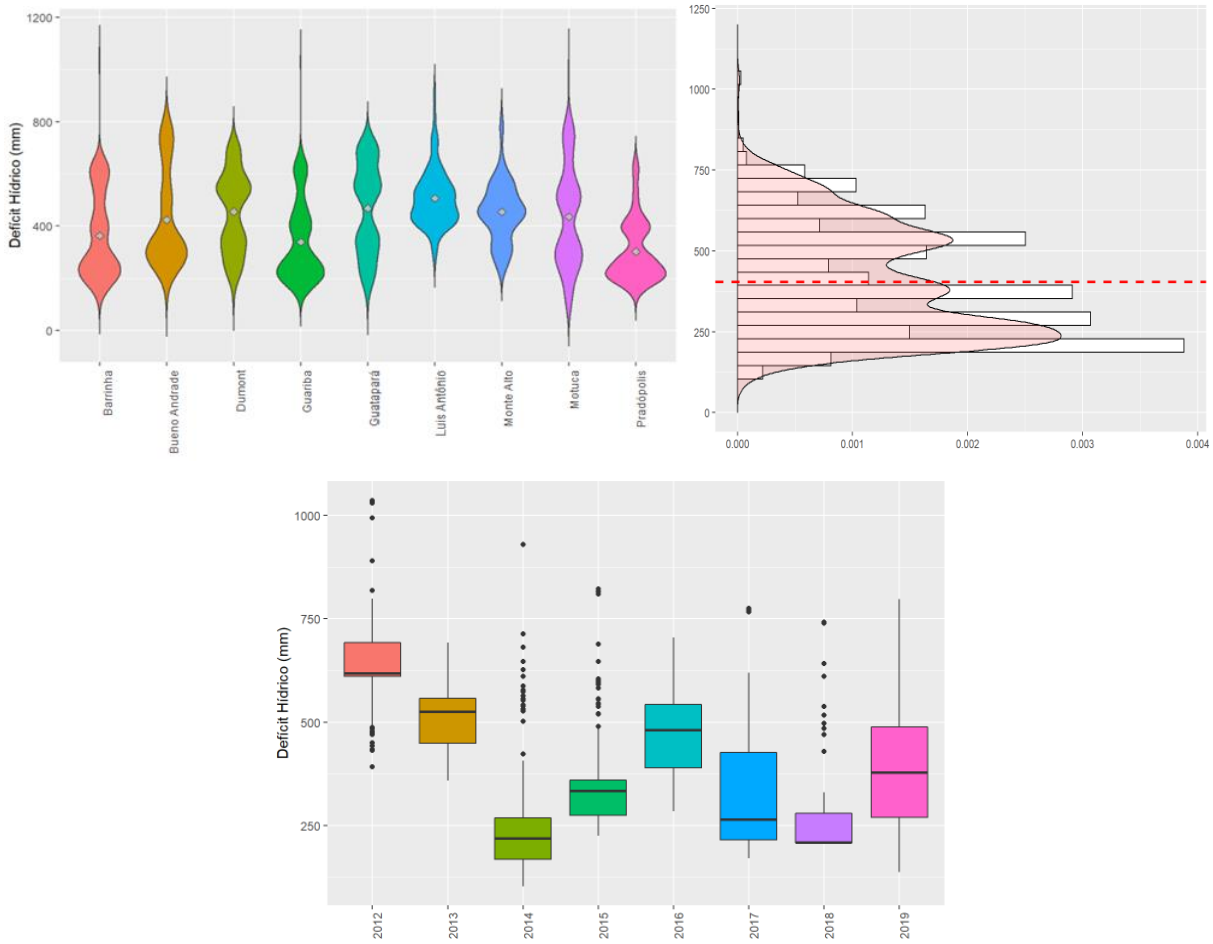


Figura 22. Déficit hídrico acumulada ao longo do ciclo do canavial nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.3.12. Excedente Hídrico

A Figura 23 apresenta os valores de excedente hídrico (exc_125) total ao longo do ciclo dos canaviais nas diferentes localidades e safras analisadas. Quando avaliada a variação entre localidades, dado o fato que todas essas se encontram na mesma macrorregião, há pouca variação entre elas, com exceção de Monte Alto, cujos dados estão com menor representatividade e foram desconsiderados da modelagem. No entanto, quando se analisou as diferentes safras, observou-se que há uma maior mediana e variação dos valores nos anos de 2012 e 2018 e menor nos anos de 2013 e 2016. Quanto à distribuição geral, observou-se uma distribuição normal com ligeiro desvio para os valores mais baixos, com a mediana sendo da ordem de 480mm.

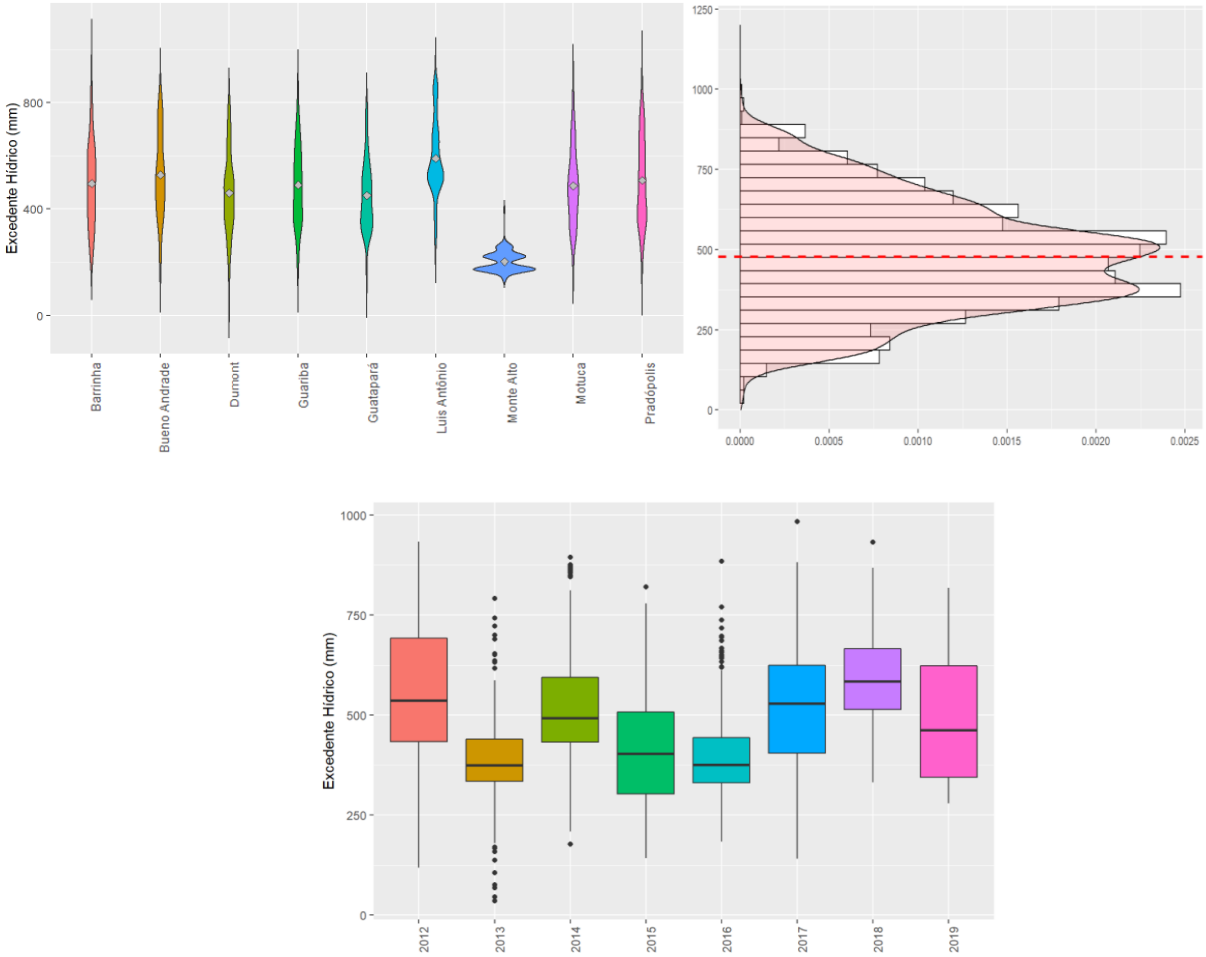


Figura 23. Excedente hídrico acumulada ao longo do ciclo do canavial nas localidades e safras em que o índice de infestação final da broca da cana-de-açúcar foi avaliado, no estado de São Paulo.

4.4. Determinação dos principais fatores determinantes do índice de infestação final utilizando-se diferentes técnicas de mineração de dados

O pacote *Caret*, cujas características foram apresentadas anteriormente, propicia identificar, por meio de diferentes técnicas de mineração de dados, as variáveis de maior importância nos modelos a serem gerados, elencando as de maior para menor importância, dentro de cada um deles.

4.4.1. *Random Forest* ou Floresta Aleatória

A Figura 24 apresenta a importância relativa das variáveis determinantes do II gerada pelo método *Random Forest*. Com a aplicação do algoritmo gerado por essa técnica de mineração de dados, obteve-se uma listagem dos fatores mais importantes para a determinação do II em ordem decrescente de importância.

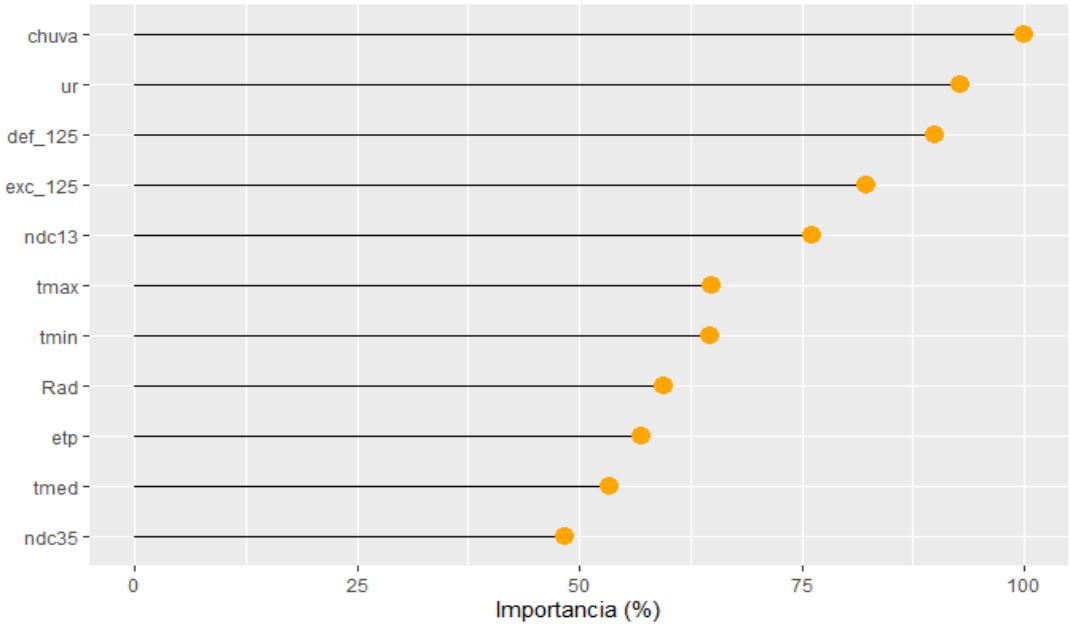


Figura 24. Classificação das variáveis meteorológicas que mais influenciam o índice de infestação final de broca da cana-de-açúcar (II), determinada pelo método *Random Forest*. Variáveis: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125).

Por meio da Figura 24 observa-se que todas as variáveis estudadas tiveram algum grau de importância no II, sendo a chuva a de maior peso relativo e o ndc35 a de menor, com grau de importância inferior a 50%. Na sequência e com grau de importância acima de 80%, encontram-se a ur, o def_125 e o exc_125. O ndc13, com grau de importância próximo dos 75%, também aparece como uma das principais variáveis a condicionar o II. As demais variáveis aparecem com grau de importância inferior a 70%. Esses resultados demonstram que a técnica *Random Forest* identificou como as principais variáveis que condicionam o II aquelas relacionadas às condições hídricas do canavial, com exceção apenas para o ndc13, que apareceu em quinto lugar na lista de fatores de importância para o II.

4.4.2. Boosting

A Figura 25 apresenta o grau de importância relativa das variáveis determinantes do II gerado pelo método do *Boosting*. Com a aplicação dessa técnica de mineração de dados foi possível observar que, assim como a técnica *Random Forest*, todas as variáveis meteorológicas analisadas apresentaram certo grau de importância no II, porém com quatro delas (etp, ndc13, ndc35 e Rad) ficando abaixo de 50%.

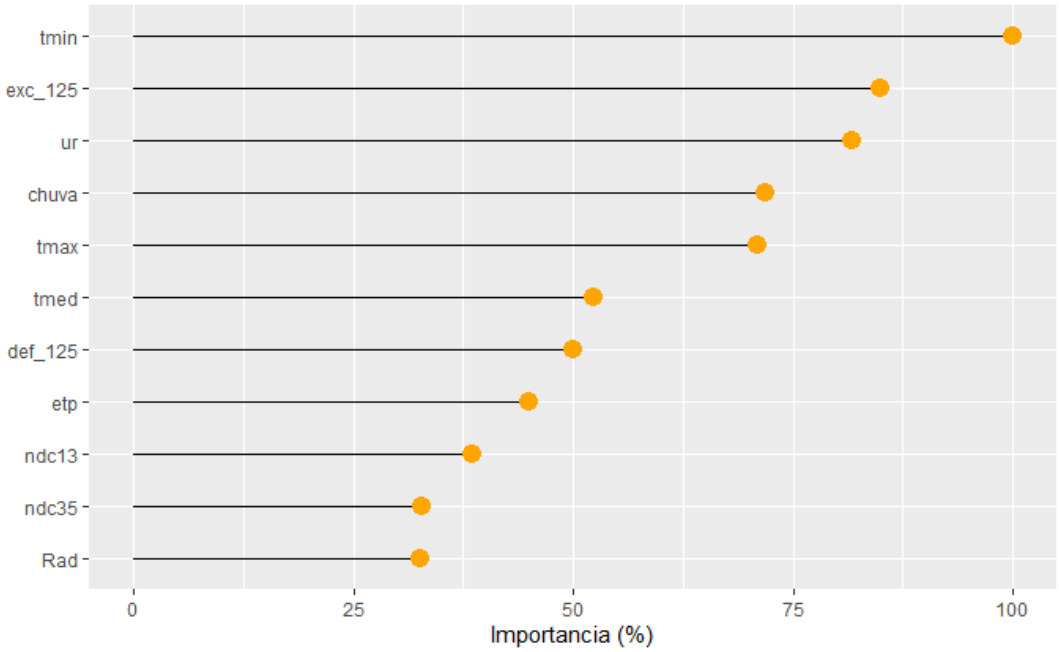


Figura 25. Classificação das variáveis determinantes do índice de infestação final de broca da cana-de-açúcar, determinada pelo método do *Boosting*. Variáveis: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125).

De acordo com a Figura 25 é possível se observar que a variável tmin foi a que teve o maior peso no II, aparecendo na sequência as seguintes variáveis: exc_125; ur; chuva; e tmax, todas com grau de importância superior a 70%, enquanto as variáveis tmed e def_125 apresentaram um grau de importância por volta de 50%. De acordo com esta técnica de mineração de dados é possível observar que houve um maior equilíbrio entre as variáveis térmicas (tmin, tmax) e hídricas (exc_125, ur e chuva) para o condicionamento do II.

4.4.3. Support Vector Machine

A Figura 26 apresenta o grau de importância relativa das variáveis meteorológicas determinantes do II gerado pelo método *Support Vector Machine*. Com a aplicação desta técnica de mineração de dados foi possivel identificar que a ur e a etp foram as únicas variáveis com grau de importância relativa acima dos 50%. A variável ndc13 apresentou um grau de importância um pouco abaixo de 50%, enquanto as demais variáveis tiveram um grau de importância inferior a 25%, o que chama a atenção já que muitas dessas variáveis apresentaram grau de importância muito elevado quando empregadas as outras técnicas.

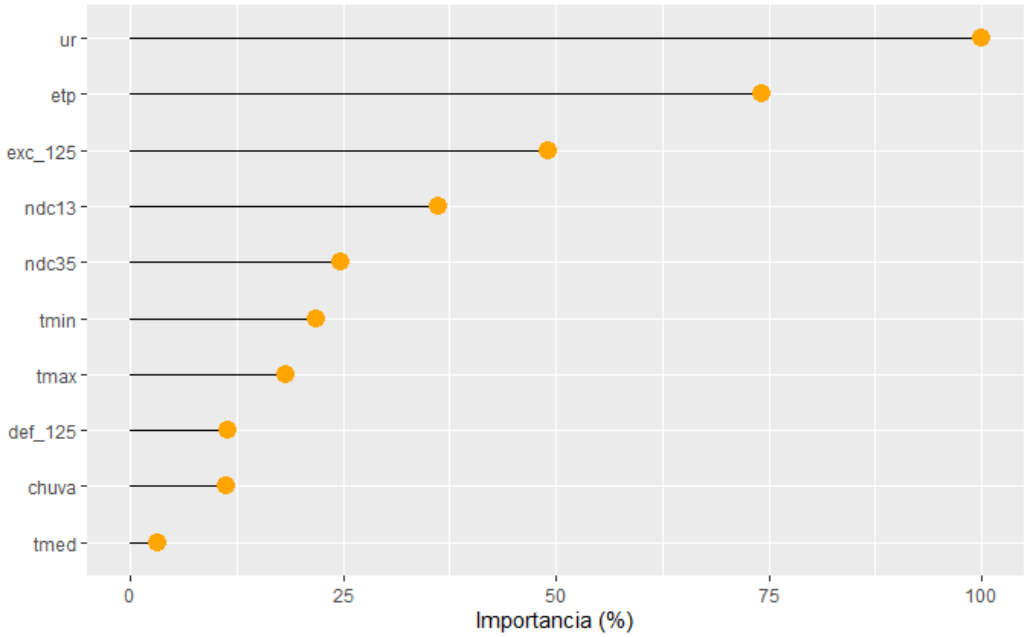


Figura 26. Classificação das variáveis determinantes do índice de infestação final de broca da cana-de-açúcar, determinada pelo método do *Support Vector Machine*. Variáveis: temperatura máxima (tmax); temperatura mínima (tmin); temperatura média (tmed); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; irradiância solar global (Rad); evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125).

Considerando-se apenas as cinco primeiras variáveis elencadas pelas três técnicas de mineração de dados, apresentadas no diagrama de Venn (Figura 27), observa-se que as únicas duas variáveis meteorológicas e de balanço hídrico que foram elencadas concomitantemente pelas três técnicas foram o exc_125 e a ur. As variáveis que apareceram elencadas em pelo menos duas técnicas foram a chuva e a ndc13. Assim, é possível observar que cada técnica de mineração de dados apresenta uma abordagem diferente na seleção das variáveis independentes, mas que, apesar disso, algumas delas acabam coincidindo, demonstrando que essas variáveis como o exc_125 e a ur realmente foram as que mais apresentaram impacto no II. A ur foi classificada como a primeira variável mais importante pela técnica do *Random Forest* (Figura 24), ao passo que pelas demais técnicas de mineração de dados essa variável foi a terceira mais importante (Figuras 25 e 26). Por outro lado, o exc_125 foi a segunda variável mais importante pela técnica do *Boosting* (Figura 25), a terceira pela técnica do *Support Vector Machine* (Figura 26) e a quarta pela técnica do *Random Forest* (Figura 24).

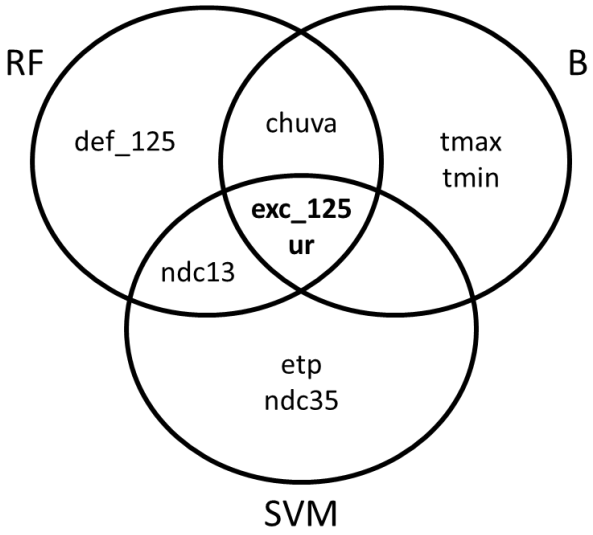


Figura 27. Diagrama de Venn considerando-se as cinco primeiras variáveis meteorológicas que influenciaram o índice de infestação final de broca da cana-de-açúcar de acordo com as três técnicas de mineração de dados: Random Forest (RF); Boosting (B); e Support Vector Machina (SVM). Variáveis: temperatura máxima (tmax); temperatura mínima (tmin); número de dias com Tmax ≤ 35°C (ndc35); número de dias com Tmin ≥ 13°C (ndc13); umidade relativa média (ur); chuva; evapotranspiração potencial (etp); deficiência hídrica (def_125); e excedente hídrico (exc_125).

4.5. Modelos para a estimação do índice de infestação final de broca por meio de técnicas de mineração de dados

Assim como para a ordenação dos fatores determinantes do índice de infestação final da broca da cana-de-açúcar, as técnicas de mineração de dados *Random Forest* (RF), *Boosting* (XGB) e *Support Vector Machine* (SVM) foram empregadas para a obtenção de modelos de estimação do índice de infestação final de broca.

Visando-se iniciar a modelagem por meio de mineração de dados, é necessário, primeiramente, se separar aleatoriamente a base de dados em duas, treinamento e testes, com diferente porcentagem de participação dos dados. Na primeira, referente ao treinamento dos algoritmos (70%), as características necessárias para representar a informação desejada são determinadas automaticamente, e outra para o teste dos modelos gerados (30%), que, neste estudo, foi feito por meio de validação cruzada (*cross validation*) (REFAEILZADEH et al., 2009).

A análise do desempenho dos modelos foi efetuada pela comparação entre os índices de infestação final de broca estimado pelos modelos e os valores observados a campo, por meio da análise de regressão. O desempenho dos modelos foi avaliado considerando-se os seguintes índices estatísticos: coeficiente de determinação (R^2) e coeficiente de correlação (r), que indicam a precisão das estimativas; coeficiente "D" de Willmott (1985), que representa a distância relativa dos valores estimados em relação à reta 1:1, ou seja a exatidão das estimativas; índice de confiança "C" de Camargo e Sentelhas (1997), que representa o grau de concordância do modelo, englobando em um único índice a precisão (r) e a acurácia (D); raiz quadrada do erro médio (*Root Mean Square Error*, RMSE), que indica a acurácia do modelo em termos do erro das estimativas; e erro absoluto médio (EAM), o qual indica a magnitude dos erros das estimativas. Os indicadores estatísticos decorrentes do processo de treinamento e teste dos modelos são apresentados na Tabela 5, enquanto a Figura 28 apresenta o desempenho em gráficos de dispersão. Observa-se que na fase de treinamento o modelo que apresentou o melhor resultado foi o *Boosting* com R^2 = 0,94 (r

= 0,969), D = 0,983, C = 0,953, EAM = 0,214% e RMSE = 0,357%, seguido pelo modelo *Random Forest* com R^2 = 0,749 (r = 0,865), D = 0,892, C = 0,772, EAM = 0,352% e RMSE = 0,536%. O modelo gerado pela técnica *Support Vector Machine* foi o pior, com valores muito baixos de R^2, D, C e erros elevados (Tabela 5; Figura 28). Já na fase de testes (Tabela 5; Figura 28), os modelos gerados pelas técnicas de *Random Forest* e *Boosting* apresentaram resultados similares, porém com o modelo *Random Forest* sendo ligeiramente melhor em relação aos erros, que foram menores (EMA = 0,539% e RMSE = 0,745%) do que os apresentados pelo modelo gerado a partir da técnica do *Boosting* (EMA = 0,548% e RMSE = 0,780%). O modelo gerado a partir da técnica de *Support Vector Machine* voltou a ter o pior desempenho estatístico também na fase de testes, o que já era esperado em razão do péssimo desempenho desse modelo na fase de desenvolvimento.

Com base nos resultados da fase de treinamento, foi possível observar que duas técnicas utilizadas (*Random Forest e Boosting*) apresentaram índices estatísticos de avaliação melhores e mais próximos, não havendo assim diferenças expressivas no desempenho de ambos, diferentemente do que foi observado com o modelo *Support Vector Machine*, que apresentou o pior desempenho. Os modelos *Random Forest e Boosting* apresentaram índices "C" entre muito bom e ótimo, segundo a classificação de Camargo e Sentelhas (1997), enquanto o modelo *Support Vector Machine* teve seu desempenho classificado como como péssimo. Analisando-se todos os índices estatísticos obtidos, pode-se considerar que os dois modelos gerados pelas técnicas RF e XGB tiveram desempenho satisfatório para o índice de infestação da broca da cana-de-açúcar, podendo ser empregados na estimação dessa variável para fins aplicados.

Em função do desempenho apresentado nas fases de treinamento e de testes, o modelo de estimação do índice de infestação de broca da cana-de-açúcar gerado a partir da técnica de mineração de dados *Random Forest* foi o selecionado para se conduzir as demais etapas deste estudo, ou seja, a caracterização da variabilidade temporal e espacial do risco de ocorrência da broca nas diferentes regiões do estado de São Paulo.

Tabela 5. Indicadores estatísticos para avaliação dos modelos empregados na estimativa do índice de infestação final de broca (II), empregando-se o conjunto de dados independentes: coeficiente de determinação (R^2); coeficiente de correlação (r); índice de Willmott (D); índice de confiança (C); erro absoluto médio (EAM); raiz quadrada do erro médio (RMSE).

Fase de Treinamento

Modelo	**R^2**	**r**	**D**	**C**	**EAM (%)**	**RMSE (%)**
Random Forest	0,749	0,865	0,892	0,772	0,352	0,536
Boosting	0,940	0,969	0,983	0,953	0,214	0,357
Support Vector Machine	0,249	0,499	0,610	0,305	0,778	1,259

Fase de Teste

Modelo	**R^2**	**r**	**D**	**C**	**EAM (%)**	**RMSE (%)**
Random Forest	0,396	0,629	0,735	0,462	0,539	0,745
Boosting	0,374	0,612	0,769	0,471	0,548	0,780
Support Vector Machine	0,195	0,442	0,563	0,249	0,826	1,275

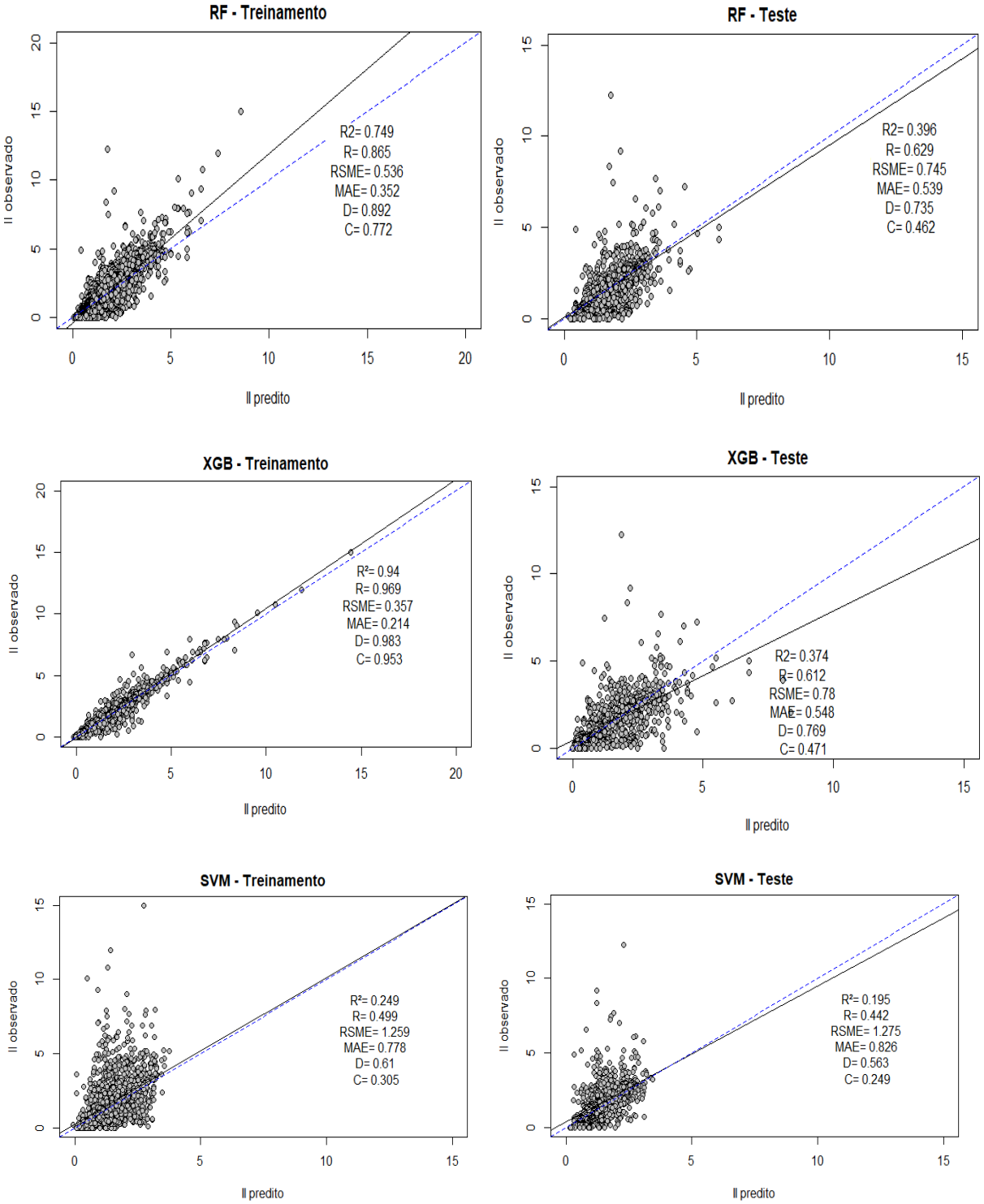


Figura 28. Relação entre o índice de infestação de broca da cana-de-açúcar estimado pelos modelos de mineração de dados (*Random Forest* – RF, *Boosting* – XGB e *Support Vector Machine* – SVM) com base em variáveis meteorológicas, nas fases de treinamento (esquerda) e teste (direita).

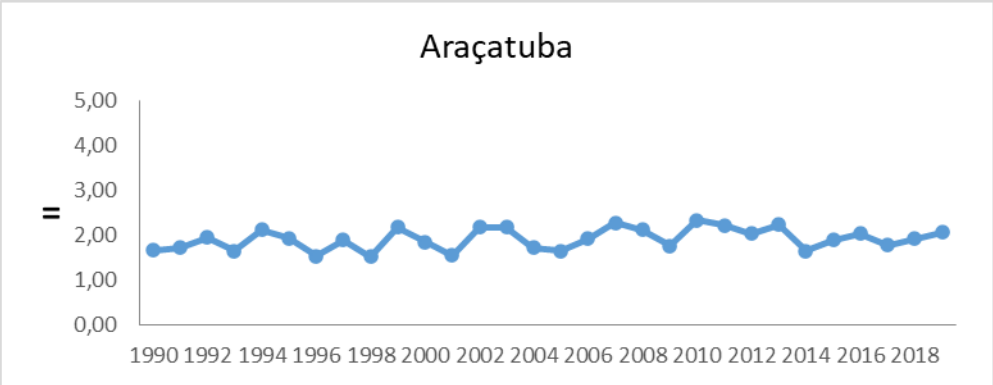
A abordagem de mineração de dados e uso de aprendizado de máquina para a estimação da ocorrência de pragas na cultura da cana-de-açúcar é recente, assim como para outros fins. Hammer et al. (2020) utilizou os modelos RF, XGB, e SVM para modelagem da produtividade da cultura da cana-de-açúcar no estado de São Paulo, tendo como

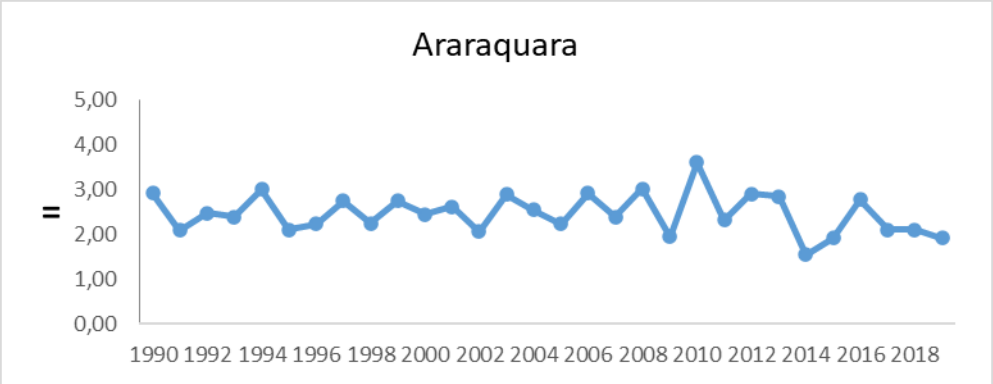
base dados meteorológicos e de produtividade de diferentes usinas. Os autores concluíram que os três modelos gerados por meio do emprego de técnicas de MD tiveram desempenho similar entre si, o que não aconteceu no estudo de Martins (2020), no qual o modelo XGB foi ligeiramente melhor do que o RF e o SVM para estimar os índices de qualidade da fibra de algodão. Assim como nos trabalhos acima citados, o desempenho dos modelos decorrentes das técnicas de MD não proporcionaram índices estatísticos tão elevados, o que se deve à grande complexidade das variáveis dos bancos de dados empregados. Além disso, os melhores valores de índice C encontrados por Hammer et al. (2020) em cada uma das técnicas empregadas foram: 0,56 para *Random Forest*; 0,56 para *Boosting;* e 0,57 para *Support Vector Machine*, enquanto Martins (2020) obteve valores de C da ordem de: 0,70 para *Random Forest*; 0,71 para *Boosting*; e 0,61 para *Support Vector Machine*, nos dois casos valores inferiores aos encontrados neste estudo na fase de treinamento.

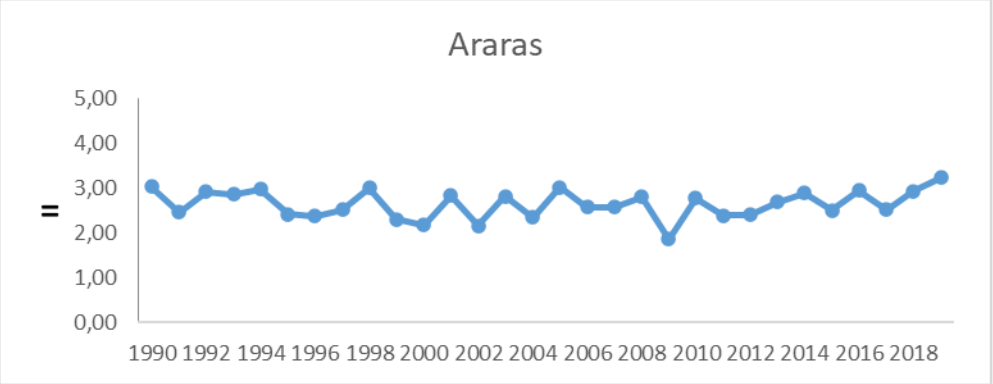
4.6. Determinação e mapeamento do risco de infestação da broca da cana-de-açúcar no estado de São Paulo

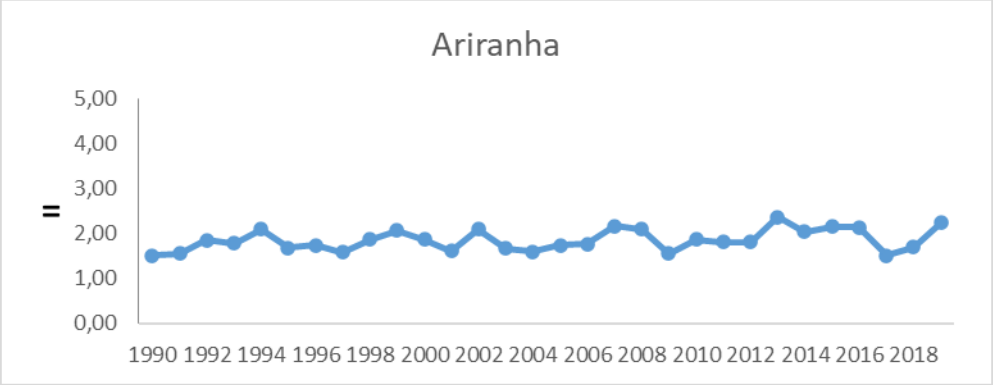
4.6.1 Variabilidade interanual do índice de infestação de broca da cana-de-açúcar

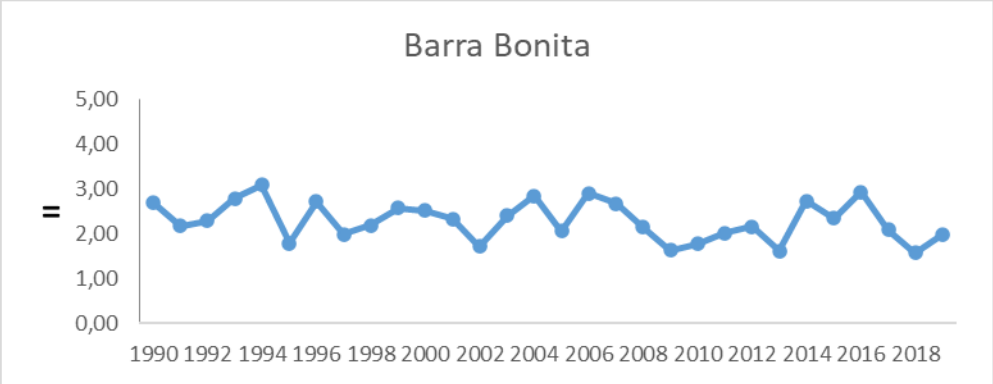
Como já mencionado anteriormente, para a determinação do risco de infestação da broca da cana-de-açúcar foram selecionados 25 locais no Estado de São Paulo, para os quais foi utilizado o modelo de estimação de II de melhor desempenho nas fases de treinamento e teste, no caso o *Random Forest*. Para cada um desses 25 locais foram obtidos os dados meteorológicos de um período de 30 anos (1990 a 2019), considerando-se todas as 11 variáveis necessárias para a estimação de II. Esses dados foram obtidos junto às mesmas bases já mencionadas anteriormente (INMET, CIIAGRO, SIGRH-DAEE-SP) e interpolados para as regiões de interesse. O modelo de infestação da broca-da-cana foi gerado anualmente resultando em 25 valores de II cujos resultados estão distribuídos anualmente na figura 29 abaixo.

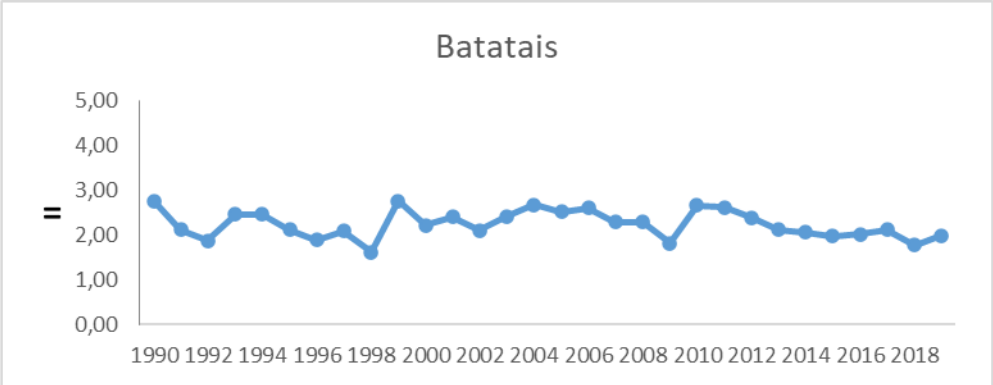

Araçatuba
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

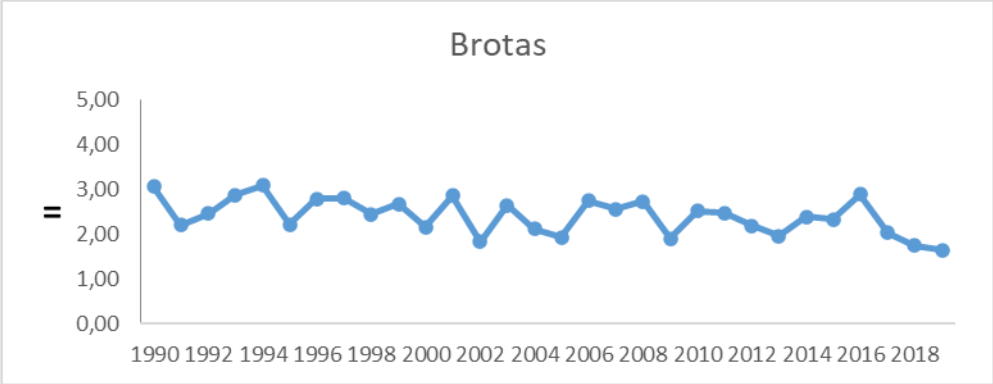

Araraquara
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

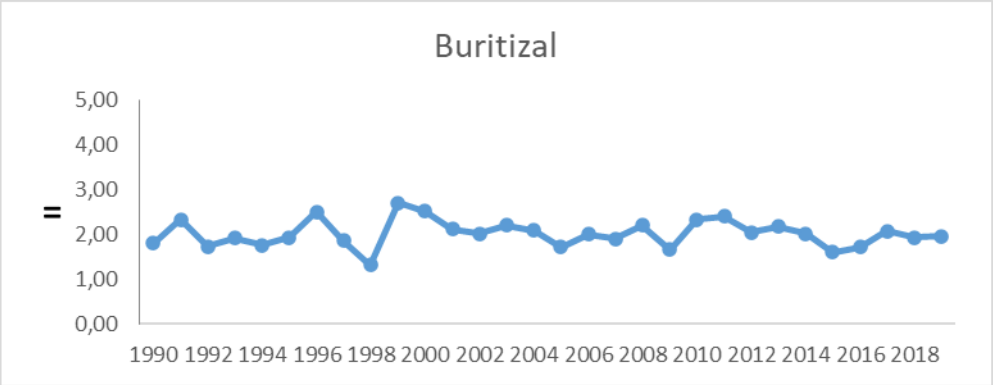

Araras
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

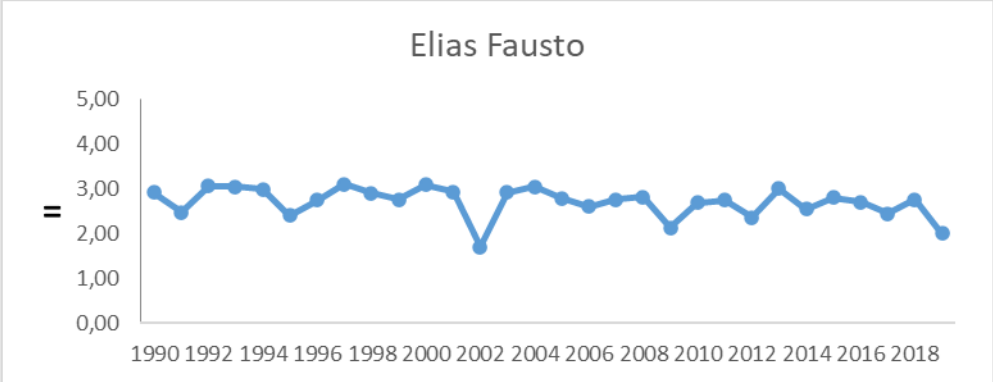

Ariranha
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

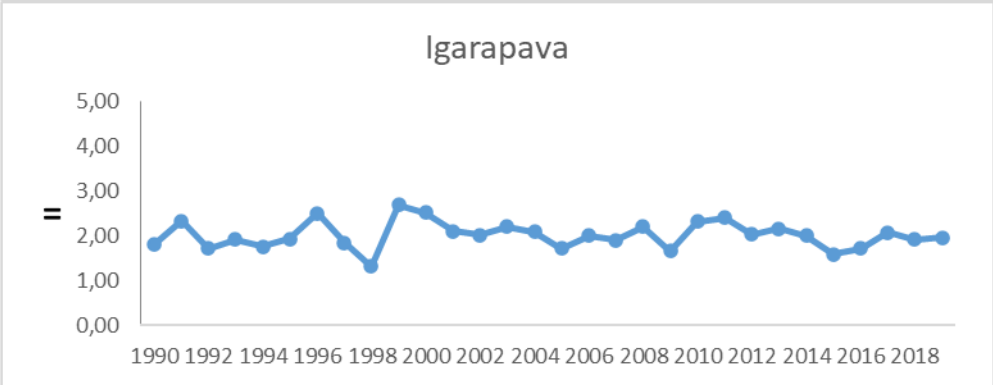

Barra Bonita
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

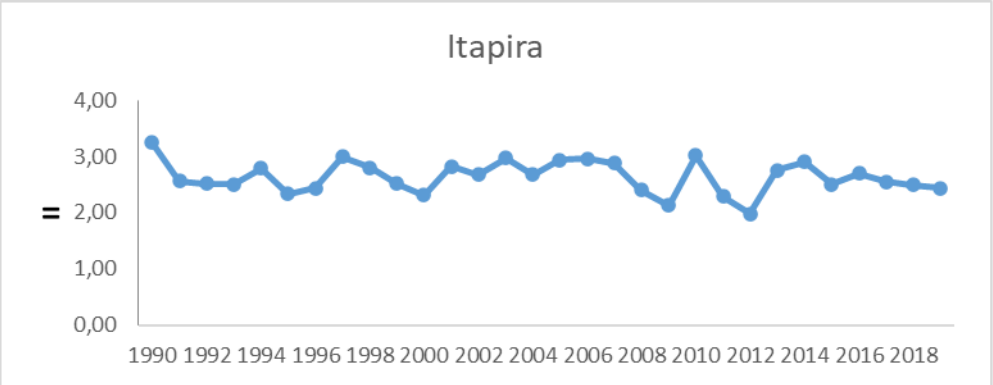

Batatais
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

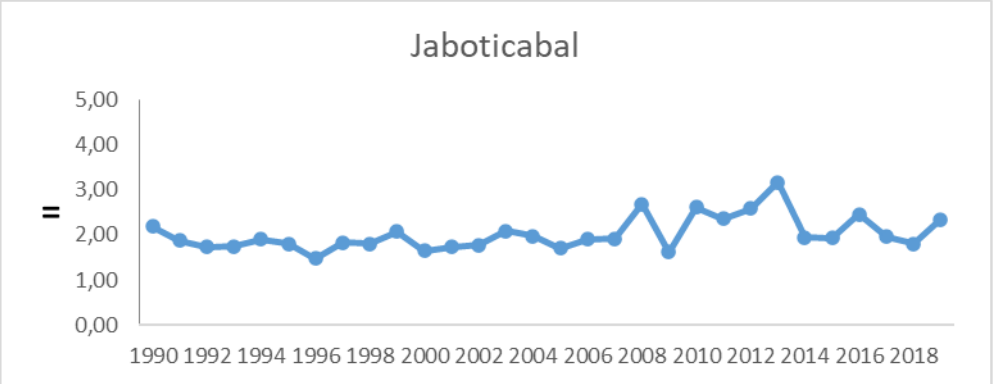

Brotas
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

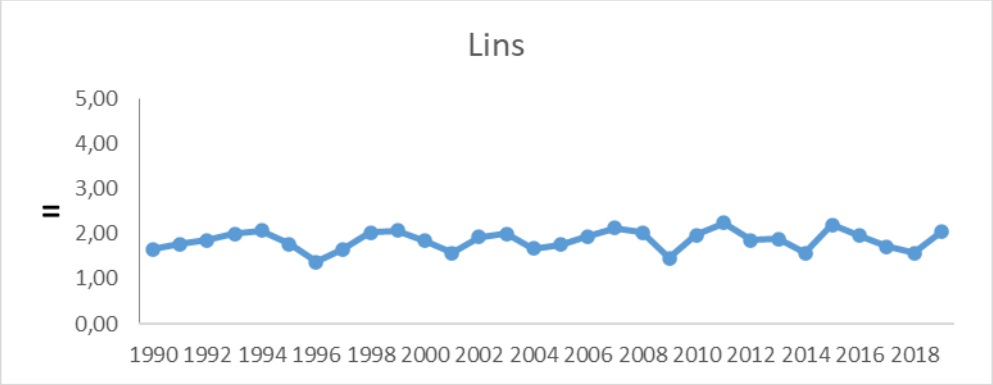

Buritizal
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

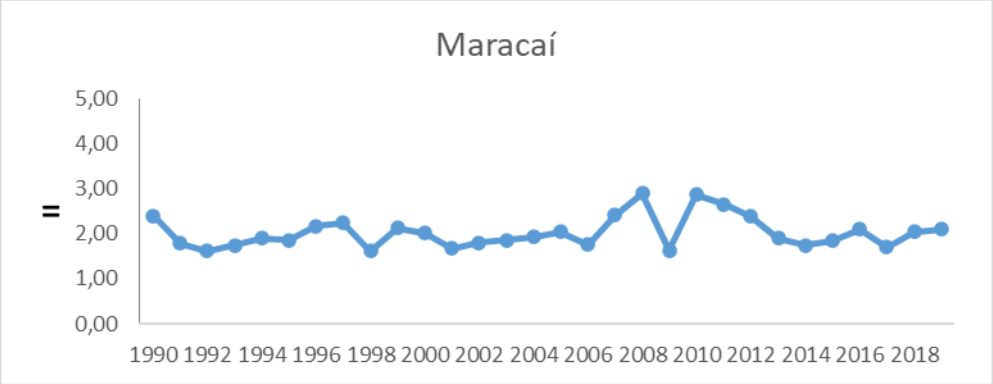

Elias Fausto
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

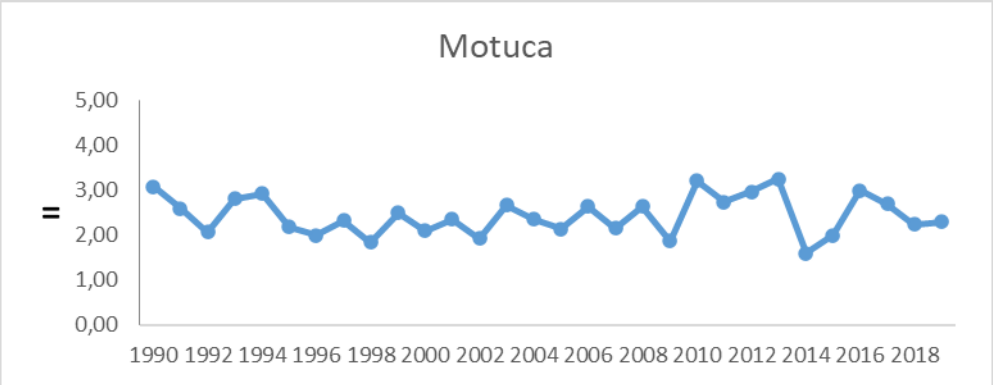

Igarapava
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

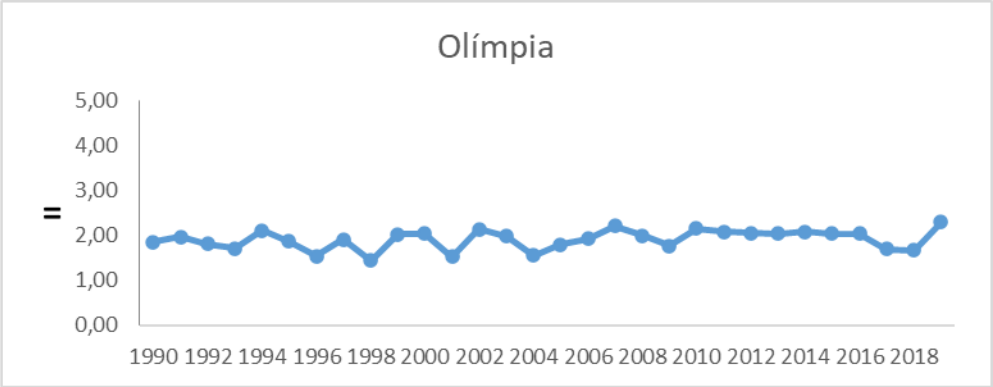

Itapira
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

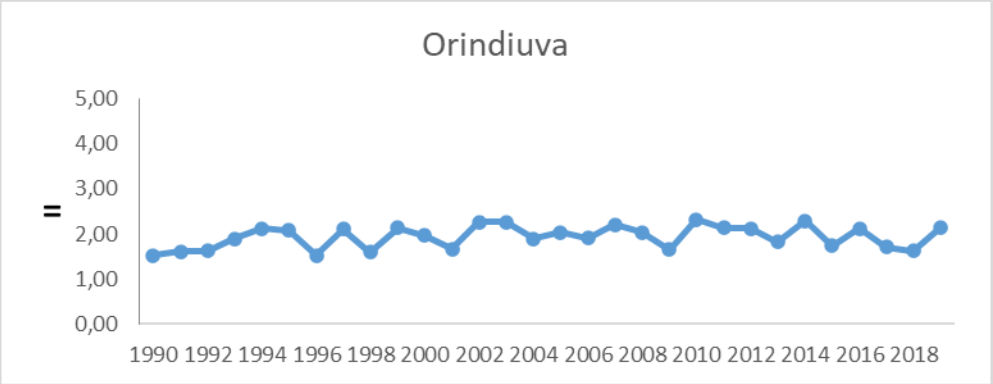

Jaboticabal
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

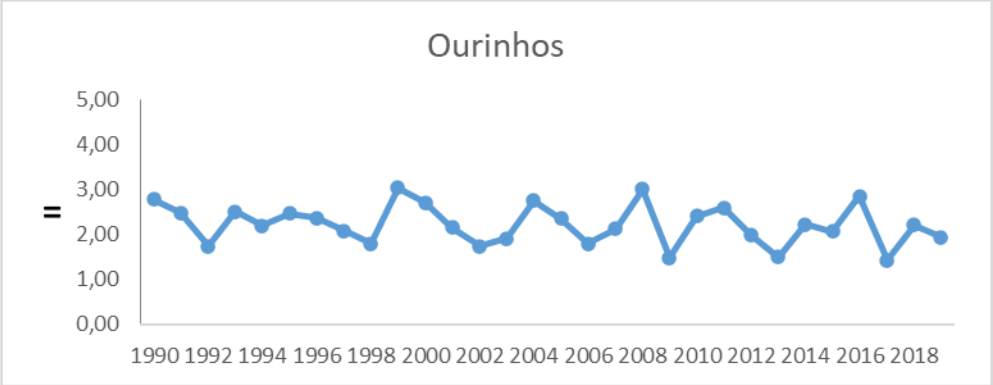

Lins
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

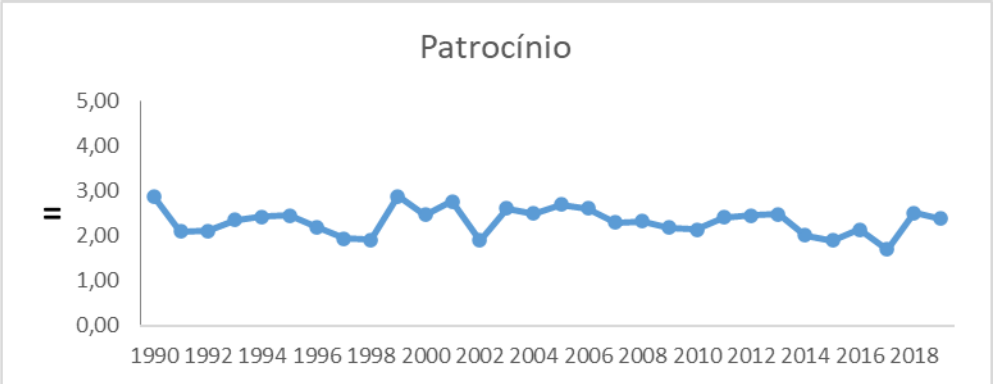

Maracaí
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

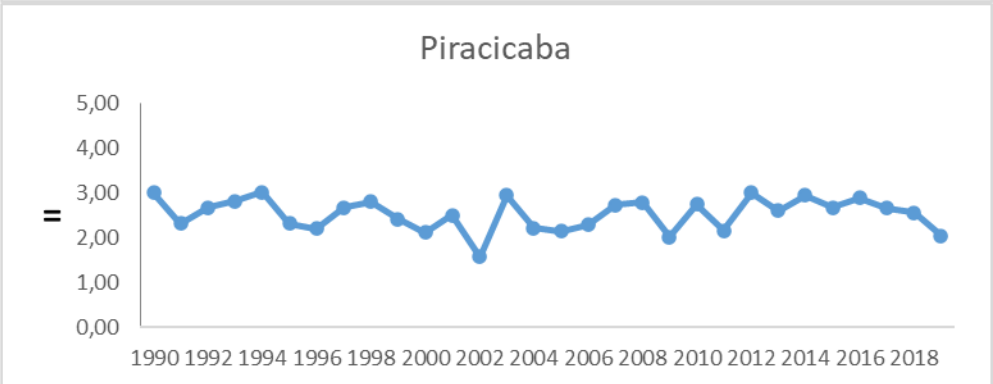

Motuca
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

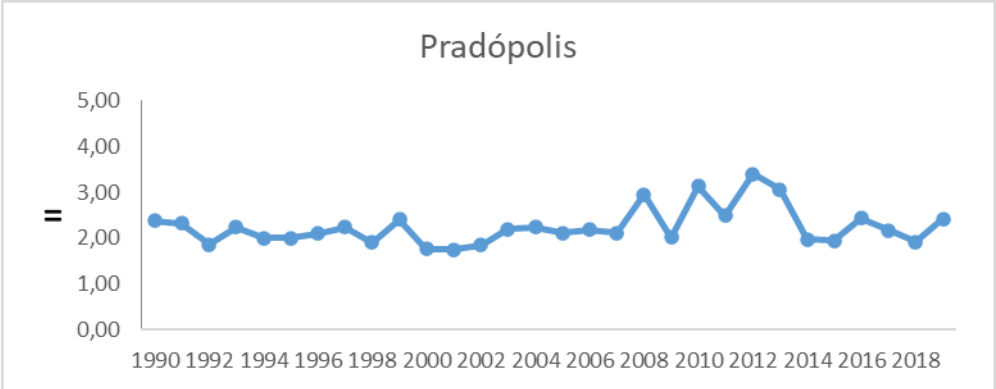

Olímpia
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

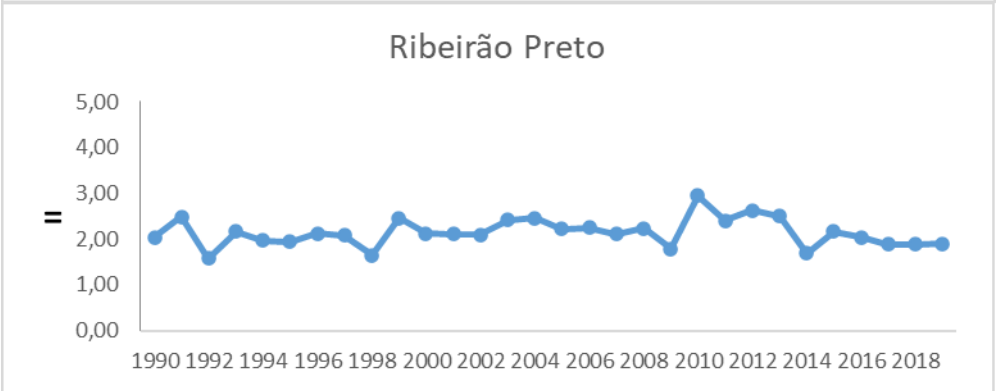

Orindiuva
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

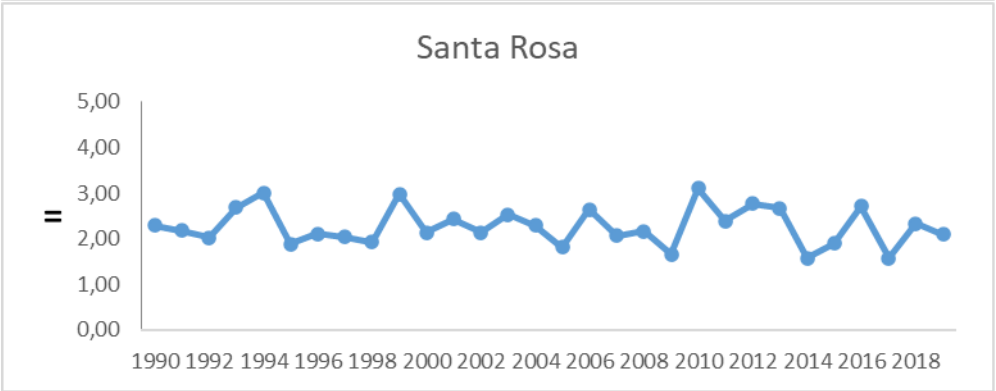

Ourinhos
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

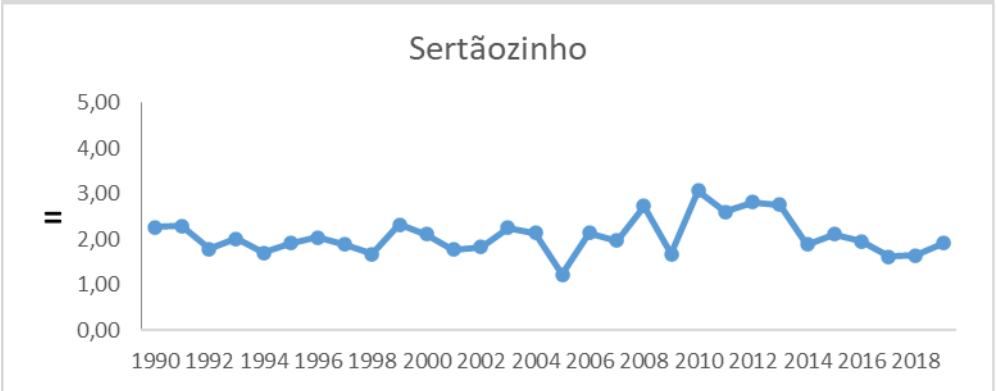

Patrocínio
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018


Piracicaba
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018


Pradópolis
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018


Ribeirão Preto
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018


Santa Rosa
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018


Sertãozinho
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

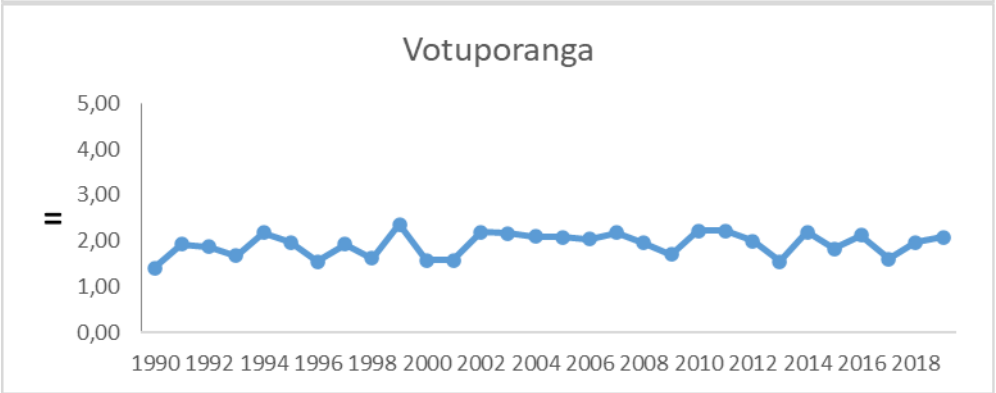

Votuporanga
5,00
4,00
3,00
2,00
1,00
0,00
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018

Figura 29. Variabilidade interanual do índice de infestação final da broca da cana-de-açúcar (II) ao longo do período entre 1990 a 2019 em diferentes localidades do estado de São Paulo, com base nas estimativas do modelo gerado a partir da técnica de mineração de dados *Random Forest*.

Observa-se nas localidades apresentadas na Figura 29 uma grande variação do II ao longo dos anos, a qual é, basicamente, influenciada pelas condições meteorológicas que vigoraram em cada ano da série histórica analisada. Cada localidade apresenta condições meteorológicas distintas em relação às 11 variáveis empregadas no modelo, as quais capturaram bem as diferenças observadas entre elas. É importante ressaltar que o modelo tem as suas variáveis de maior peso e importância, sendo que no caso do *Random Forest* as cinco principais variáveis são: chuva; umidade relativa; déficit hídrico; excedente hídrico; e número de dias com temperatura mínima ≥ 13 ^{o}C.

Analisando-se os resultados das estimativas das 25 localidades que serviram como base para a determinação de risco da broca da cana-de-açúcar, os cinco locais com maiores índices médios de infestação final de broca da cana-de-açúcar (II) ao longo da série de dados de 30 anos pelo modelo *Randon Forest* foram os seguintes locais: Elias Fausto; Itapira; Araras; Piracicaba; e Araraquara. Os cinco locais com menores índices médios de infestação final de broca (II) foram: Olímpia; Araçatuba; Maracaí; Ariranha; e Lins, sendo que os demais locais apresentaram II intermediário.

Os resultados obtidos neste estudo corroboram aqueles encontrados por Carbognin (2016) que, baseados em modelos matemáticos em função dos graus-dia e equações de diferenças, demonstraram haver maior ocorrência de adultos de broca no período das chuvas, principalmente, de janeiro a abril, não ocorrendo correlação entre os as condições meteorológicas e a flutuação populacional da praga em Pradópolis e Valparaíso, ao contrário do que ocorreu em Quirinópolis, onde a umidade foi limitante para o crescimento populacional desta espécie. Assim, os autores verificaram que houve a correlação entre a infestação da broca e as variáveis meteorológicas apenas em Quirinópolis, GO, onde houve relação positiva ($p<0,05$) entre a dinâmica de adultos de *D. saccharalis* e a chuva nas safras de 2013/2014 e 2014/2015. Esses autores também obtiveram correlação positiva ($p<0,01$) entre a umidade e o número de mariposas capturadas. Em um trabalho seguinte, Carbognin (2019) analisou outros três locais e verificou que as variáveis chuva, temperatura média e umidade afetaram a ocorrência de broca-da-cana. Esses autores observaram que em Barra Bonita, Pradópolis e Valparaíso, no estado de São Paulo, há uma maior ocorrência de adultos dessa praga dos canaviais quando a temperatura, a chuva e a umidade são mais elevadas, o que corrobora os resultados obtidos no presente estudo, que identificou maiores riscos de ocorrência da broca-da-cana quando as condições térmicas e hídricas são mais elevadas.

4.6.2 Variação espacial do risco de infestação de broca da cana-de-açúcar no estado de São Paulo

Para a determinação da variação espacial do risco de infestação da broca-da-cana foram selecionados 25 locais no Estado de São Paulo, já mencionados anteriormente, para os quais foi utilizado o modelo de *Random Forest* para a estimação de II de melhor desempenho nas fases de treinamento e teste. O modelo de infestação da broca-da-cana foi gerado anualmente resultando em 25 valores de II. Com base nesses dados foi estabelecida a frequência relativa de anos com infestações acima do limiar considerado pelas empresas sucro-energéticas como definidor da necessidade de controle químico, ou seja, de 2%, para cada uma das 25 localidades.

A essa frequência (FII>2%) se denominou de risco de infestação da broca-da-cana, o qual foi classificado da seguinte forma: Baixo (≤ 30%); Moderado (30,1 a 60%); Alto (60,1 a 90%); e Muito Alto (> 90%). Os dados de FII>2% desses 25 locais foram correlacionados, por meio de uma regressão linear múltipla, com as coordenadas geográficas (latitude – Lat e longitude – Long) e com a altitude (Alt), de modo a gerar um modelo linear para a espacialização dessa variável. Finalmente, o desempenho do mapa temático, foi avaliado considerando-se os dados independentes de 10 localidades bem distribuídas no estado de São Paulo, também já mencionadas.

O mapa obtido com o procedimento acima (Figura 30) apresenta o risco de infestação de broca-da-cana para o Estado de São Paulo. Com esse mapa é possível se observar uma variação regional do risco de infestação, com risco muito alto na faixa sudeste do estado, englobando as mesorregiões de Itapetininga, Campinas, Mococa e Vale do Paraíba. Já nas porções centro-norte, que engloba as mesorregiões de Avaré, Araraquara e Ribeirão Preto, e no extremo noroeste do estado (Ouroeste e Santa Albertina), o risco é alto. Na região central e oeste do estado, onde se encontram as regiões de Assis, Marília, Lins, Catanduva, São José do Rio Preto, Araçatuba e Andradina, o risco é médio, passando para baixo somente em algumas áreas do Pontal do Paranapanema (Teodoro Sampaio e Rancharia). Esse mapa apresentou um desempenho bastante satisfatório quando comparado com dados independentes, ou seja, r = 0,80, D = 0,91, C = 0,73, EM = 1,48, EAM = 6,78, RSME = 6,64 e EAMP = 11,2%, similar ao obtidos por Sentelhas et al. (2016) para o mapa de risco da ferrugem alaranjada da cana.

Esse direcionamento de risco da broca-da-cana para o estado de São Paulo é algo inovador e inédito no setor sucroenergético abrindo diversas possibilidades de manejo e planejamento tanto ao nível estadual, como apresentado no mapa da Figura 30, quanto ao nível local, permitindo identificar o nível de risco e a variabilidade interanual de infestação, como uma função das condições meteorológicas (Figura 29). Tal modelagem, apesar de ser exploratória e realizada com banco de dados pré-existente e sem a formatação ideal para estudos dessa natureza, pode funcionar como ferramenta para identificação de locais com potencial risco de alta infestação da broca da cana-de-açúcar em diferentes localidades do Estado de São Paulo e, assim, auxiliar no manejo desta praga tão importante para os canaviais.

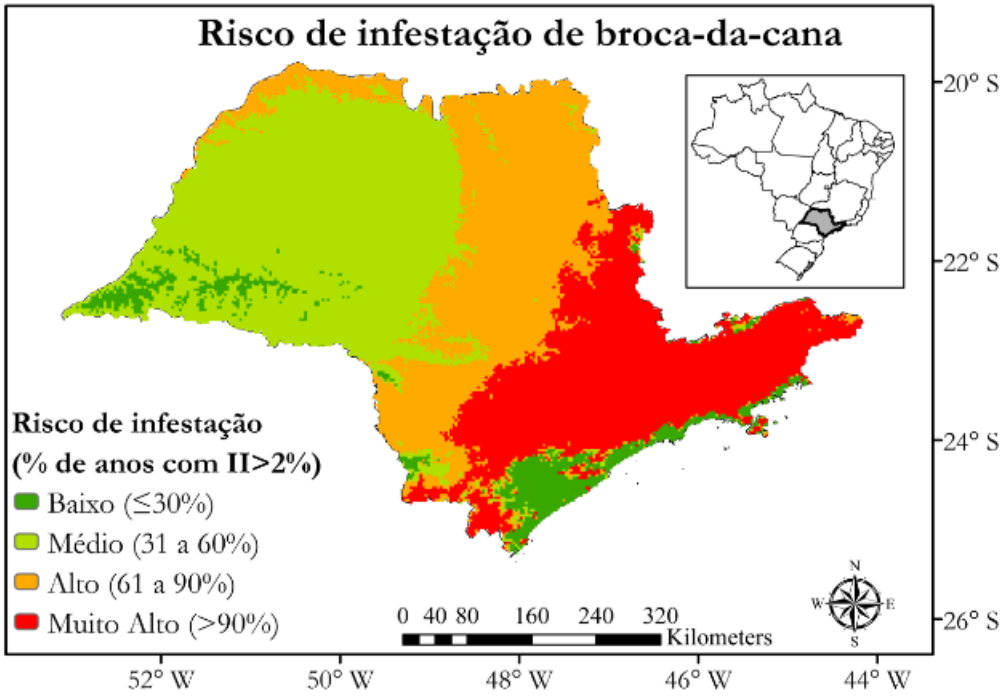


Figura 30. Risco de infestação da broca da cana-de-açúcar no estado de São Paulo, considerando-se a porcentagem de anos em que o índice de infestação (II) foi superior a 2%, de acordo com o modelo de estimação baseado na técnica de mineração de dados *Random Forest*.

5.CONCLUSÕES

Os resultados obtidos no presente estudo permitiram concluir que:

a) Há uma grande dependência do índice de infestação final de broca (II) com as variáveis meteorológicas;

b) As técnicas de mineração de dados permitiram identificar as variáveis meteorológicas que mais influenciam a infestação de broca nos canaviais, porém com uma grande heterogeneidade entre elas quanto às variáveis de maior importância;

c) Considerando-se as variáveis elencadas pelas três técnicas de mineração de dados, as variáveis meteorológicas e de balanço hídrico que mais influenciam a infestação de broca são o excedente hídrico e a umidade relativa do ar, já que ambas foram selecionadas concomitantemente por todos os modelos. Na sequência, as variáveis que apareceram elencadas por pelo menos duas técnicas foram à chuva e o número de dias com a temperatura Tmin ≥ 13°C;

d) Os modelos de melhor desempenho para a estimação da infestação de broca-da-cana foram os provenientes das técnicas de *Boosting* e de *Random Forest;*

e) O risco de ocorrência da broca-da-cana no estado de São Paulo varia entre as regiões, sendo alto a muito alto nas regiões canavieiras do sul, centro-norte, leste e extremo noroeste do estado, enquanto que na região centro-oeste o risco oscila entre médio e baixo;

f) A abordagem empregada neste estudo, apesar de suas limitações associadas ao banco de dados empregado, possibilita a identificação dos anos e das áreas de maiores riscos para a infestação da broca-da-cana, contribuindo, assim, para o manejo racional dessa praga.

REFERÊNCIAS

AMARAL S.F. & H.V. ARRUDA. **Grau de suscetibilidade de 15 variedades de cana-de-açucar à broca *Diatraea saccharalis*** (F., 1794) (Lepidoptera-Pyralidae). Arq. Inst. Biol. 31:79-83, 1964.

AMARASEKARE, P.; COUTINHO, R. M. The intrinsic growth rate as a predictor of population viability under climate warming. **Journal of Animal Ecology**, United Kingdom, v. 82, p. 1240–1253, 2013.

AMARASEKARE, P.; SIFUENTES, R. Elucidating the temperature response of survivorship in insects. **Functional Ecology**, West Sussex, v. 26, p. 959-968, 2012.

AGRIANUAL 2009: anuário da agricultura brasileiro. São Paulo: IFNP, 2009. 497p.

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).

ARNOLD, C.Y. The determination and significance of the base temperature in a linear heat unit system. **American Society for Horticultural Science**, St. Joseph, v. 74, p. 430-445, 1959.

ARRIGONI, E. de B. Broca da cana-de-açúcar: importância econômica e situação atual. In: ARRIGONI, E. de B.; DINARDO-MIRANDA, L.L.; 72 ROSSETO, R. Pragas da cana-de-açúcar: Importância econômica e enfoques atuais. Piracicaba: Stab, 2002, p. 1-4.

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.

BERTELS, A. Estudos da influência da umidade sobre a dinâmica de populações de lepidópteros-pragas do milho. **Pesquisa agropecuária brasileira**, Brasília, v. 5, p.567-579, 1970.

BOETHEL, D.J. & R.D. Eikenbary. **Interactions of plant resistance and parasitoids and predators os insects. Chichester**, Ellis Horwood, 224p, 1986.

BOTELHO, P. S. M.; MENDES, A. C.; MACEDO, N.; SILVEIRA-NETO, S. Influences of climatic factors on the population fluctuations of the sugarcane moth borer, *Diatraea saccharalis* (Fabr., 1794) (Lep. – Crambidae). In: **Proceedings, 16th Congress of the International Society of Sugarcane Technologists**, São Paulo- SP, p. 643-655, 1977.

BOTELHO, P.S.M.; MENDES, A.C. MACEDO, N.; SILVEIRA NETO, S. Testes comparativos de armadilhas para coleta de Diatraea saccharalis (Fabr., 1794) (Lep., Pyralidae). **Brasil Açucareiro**, Rio de Janeiro, v. 88, p. 38-72, 1976. Cálculo da dispersão da broca da cana-de-açúcar, Diatraea saccharalis (Fabr., 1974). **Brasil Açucareiro**, Rio de Janeiro, v. 5, p. 34-38, 1978.

BOTELHO, P. S. M.; MACEDO, N. *Cotesia flavipes* para o controle de *Diatraea saccharalis*. In: **Controle Biológico no Brasil – Parasitóides e Predadores**, 1. ed. São Paulo: Manole, 2002. cap. 25, p. 409-425.

BOTELHO, P. S. M.; MAGRINI, E. A.; SILVEIRA NETO, S.; MACEDO, N. Flutuação populacional de machos de *Diatraea saccharalis* (Fabricius.) através de armadilhas de feromônios. **Anais da Sociedade Entomológica do Brasil**, Londrina, v.22, n.2, p.293-297, 1993.

BOMMIREDYY, P. L.; PARAJULEE, M. N.; PORTER, D. O. Influence of constant temperatures on life history of immature Lygus elisus (Hemiptera: Miridae). **Environmental Entomology**, Cary, v. 33, p. 1549–1553, 2004.

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.

BUG Agentes Biológicos. Disponível em: <http:bugagentesbiologicos.com.br>. Acesso em: 05 jun. 2019.

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

BRAGA, B. et al. Introdução à engenharia ambiental. São Paulo: Prentice Hall, 2003, 143 p.

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.

BRUNINI, O. Ambientes climáticos e exploração agrícola da cana-de-açúcar. In: DINARDO MIRANDA, L. L.; VASCONCELOS, A. C. M. de; ANDRADE LANDELL, M. G. de. **Cana-de-açúcar**. Campinas: Instituto Agronômico, 2008. p. 205-218.

BRUCE, P.; BRUCE, A. **Estatística Prática para Cientistas de Dados: 50 conceitos essenciais**. 1. ed. Rio de Janeiro: Alta Books, 2019.

BRUNINI, O.; LISBÃO, R. S.; BERNARDI, J. B.; FORNASIER, J. B.; PEDRO JUNIOR, M. J. Temperatura-base para alface cultivar "White-Boston", em um Sistema de unidades térmicas. **Bragantia**, Campinas, v. 35, n. 19, p. 213-219, 1976.

CARBOGNIN, E. R. **Modelos de previsão de ocorrência de adultos de *Diatraea saccharalis* (Fabricius, 1794) (Lepidoptera: Crambidae) em cana-de-açúcar (*Saccharum* spp.)**. 2016. 53 f. Dissertação (Mestrado em Entomologia Agrícola) - Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista "Júliode Mesquita Filho", Jaboticabal.

CARBOGNIN, E. R. Modelagem de previsão **de *Diatraea saccharalis* (Fabricius, 1794) (Lepidoptera: Crambidae) em cana-de-açúcar (*Saccharum* spp.)**. 2019. 110 f. Doutorado (Entomologia Agrícola) - Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista "Júliode Mesquita Filho", Jaboticabal.

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.

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

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.

CAMPOS, M.B.S; MACEDO, N. Cana-de-açúcar - ampliando campo de ataque. Cultivar: Grandes Culturas, v. 6, n. 68, p. 23-26, 2004.

CAMARGO, A. P.; SENTELHAS, P. C. Avaliação do desempenho de diferentes modelos 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.

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.

CEBRASA. Centro de Soluções Ambientais. **Controle biológico**. 2008. Disponível em:< http://www.cebrasa.agr.br/controle_biologico.html>. Acesso em: 29 jan. 2019.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Acompanhamento da safra brasileira**. **Cana-de-açúcar**. Brasília, 2020. 17 p. Acesso em: 15 dez. 2020.

DE CONTI, B. F.; BUENO, V. H. P.; SAMPAIO, M. V.; SIDNEY, L. A. Reproduction and fertility life table of three aphid species (Macrosiphini) at different temperatures. **Revista Brasileira de Entomologia**, São Paulo, v. 54, p. 654–660, 2010.

DERNEIKA, O. & F.M. LARA. **Resistência de cana-de-açúcar a *Diatraea* saccharalis (Fabr., 1794) (Lepidoptera-Pyralidae): comportamento de variedades em três cortes e em quatro locais do Estado de São Paulo**. An. Soc. Entomol. Brasil 20:359-368, 1991.

DINARDO-MIRANDA, L. L. Pragas. In: **Cana-de-açúcar** (Ed.) DINARDO-MIRANDA, L. L.; VASCONCELOS, A. C. M.; LANDELL, M. G. A de. Campinas: Instituto Agronômico, 2008, cap. 17.

DINARDO-MIRANDA, L.L.; FRACASSO, J.V.; ANJOS, I.A. dos et al. Influência da infestação de Diatraea saccharalis (Fabr.) sobre parâmetros tecnológicos da cana-de-açúcar. Bragantia, v. 71, n. 3, p.342-345, 2012.

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.

DOORENBOS, J.; KASSAM, A. H. **Yield response to water.** Rome: FAO, 1979. 179 p. (FAO. Irrigation and Drainage Paper, 33).

FABRE, F.; PIERRE, J. S.; DEDRYVER, C. A.; PLANTEGENEST, M. Barley yellow dwarf disease risk assessment based on Bayesian modelling of aphid population dynamics. **Ecological Modelling**, Amsterdam, v. 193, p. 457-466, 2006.

FERRARO, D.O.; RIVERO, D.E.; GHERSA, C.M. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. **Field Crops Research,** Amsterdam, v. 112, p. 149-157, 2009.

FIGUEIREDO, M. L. C.; CRUZ, I.; SILVA, R. B. Aspectos Bioecológicos de *Diatraea saccharalis* Fabr. (Lepidoptera: Crambidae) em Plantas de Milho (*Zea mays* L.). In: CONGRESSO NACIONAL DE MILHO E SORGO, 28., 2010, Goiânia. **Resumos...** Goiânia: ABMS, 2010. p. 230-235.

FUSER, I. Os perigos da monocultura. **Revista Atualidades Vestibular**, 2008, 241p.

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.

GALLO D. **Infestação da broca nas principais variedades de cana**. Rev. Agric. 29:149-155, 1954.

GALLO, D. et al. **Entomologia agrícola**. Piracicaba: FEALQ, 2002. 920p.

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.

GASCHO, G. J.; SHIH, S. F. Sugarcane. In: TEARE, I.D.; PEET, M.M. (Ed.). **Crop-water relations.** New York: Wiley-Interscience, 1983. p. 445-479.

GEREMIAS, L. D. **Exigências térmicas e danos de *Diatraea saccharalis* (Fabricius, 1794) (Lepidoptera: Crambidae) em milho e base para seu controle com *Trichogramma galloi* Zucchi, 1988 (Hymenoptera: Trichogrammatidae).** 2013. Tese (Doutorado em Ciências) – Escola Superior de Agricultura Luiz de Queiroz-ESALQ, Universidade Estadual de São Paulo – USP, Piracicaba-SP.

GILLESPIEL, T.J.; SENTELHAS, P.C. Agrometeorology and plant disease management – a happy marriage**. Sci. Agric.,** Piracicaba, v.65, special issue, p.71-75, December 2008.

GUTIERREZ, A. P. **Applied population ecology: a supply-demand approach.** 1996. 300 p.

HAMMER, R. G. **Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados**. 2016. (Dissertação de Mestrado). Escola Superior de Agricultura "Luiz de Queiroz"-Universidade de São Paulo, Piracicaba, 2016.

HAMMER, R. G.; SENTELHAS, P. C.; MARIANO, J. C. Q. Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. **Sugar Tech**, v.22, p.216-225, 2020.

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

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

HADDAD, J. R.; PARRA, J. R. P. **Métodos para estimar os limites térmicos e a faixa ótima de desenvolvimento das diferentes fases do ciclo evolutivo dos insetos**. Piracicaba: Fundação Luiz de Queiroz, 1984. 12 p.

HADDAD, J. R.; PARRA, J. R. P.; MORAES, R. C. B. **Métodos para estimar os limites térmicos inferior e superior de desenvolvimento de insetos.** Piracicaba: Fundação Luiz de Queiroz, 1999. 15 p.

HANNON, B.; RUTH, M. **Dynamic modeling**, 2001. New York, Springer, 2. ed, 409 p.

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

HIGLEY, L. G.; PEDIGO, L. P.; OSTLIE, K. R. DEGDAY: a program for calculating degree-days, and assumptions behind the degree-day approach. **Environmental Entomology**, Annapolis, v. 15, n. 5, p. 999-1016, 1986.

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

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

JAROSIK, V.; KRATOCHVIL, L.; HONÉK, A.; DIXON, A. F. G. A general rule for the dependence of developmental rate on temperature in ectothermic animals. Proceedings of the Royal Society of London, London, v. 271, p. 219-221, 2004.

JONES, J.W. Integrating models with expert systems and data bases for decision making. Climate and Agriculture: Systems Approaches to Decision Making. **American Meteorological Society**, Massachusetts, p. 194-211, 1989.

KAISER, H.F. The varimax criterion for analytic rotation in factor analysis. Psychometrika, New York, v. 23, p. 187-200, 1958.

KING, N. J.; MUNGOMERY, R. W.; HUGUES, C. G. **Manual of cane growing**. New York, Elsevier, 1965.

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.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistical Software**, v.28, n.5, p.1-25, 2008.

LARA, F.M. **Princípios de resistência de plantas a insetos**. São Paulo, Ed. Ícone, 336, 1991.

LANDELL, M. G. A.; CAMPANA, M. P.; FIGUEIREDO, P.; SILVA, M. A. **Programa de melhoramento genético: estratégias para o desenvolvimento de novas variedades de cana-de-açúcar**. In: LANDELL M. G. A.; VASCONCELOS, A. C. M.; DINARDOMIRANDA, L. L. (Coords.) WORKSHOP, 2, Programa Cana IAC. Instituto Agronômico de Campinas. p. 7-10. 2003a.

LEITE, G. M. V.; ANDRADE, L. A. B.; GARCIA, J. C.; ANJOS, I. A. Efeitos de fontes e doses de silicato de cálcio no rendimento agrícola e na qualidade tecnológica da cana-de-açúcar, cultivar SP80-1816. **Ciência e Agrotecnologia**, Lavras, v. 32, n. 4, p. 1.120-1.125, 2008.

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.

LIMA, E. A. B. F.; FERREIRA, C. P.; GODOY, W. A. Ecological modeling and pest population management: a possible and necessary connection in a changing world. **Neotropical Entomology**, Londrina, v. 38, n. 6, p. 699-707, 2009.

LIMEIRA, E. H. **Modelagem matemática aplicada ao controle da praga da cana-de-açúcar para a produção de etanol: estratégias ótimas de controle.**
Dissertação (Mestrado em Energia) - Universidade Federal do ABC, Santo André, 2010.

LYRA NETO, A. M. C.; LIRA, M. de. A.; FRANÇA, J. G. E.; MACIEL, G. A.; FREITAS, E. V. Flutuação populacional de *Diatraea* spp. (lepidoptera: Pyralidae) em sorgo com armadilha luminosa. **Anais da Sociedade Entomológica do Brasil**, Londrina, v.19, n.2, p.355-360, 1990.

LYRA, W. S.; SILVA, E.C.; ARAÚJO, M.C.U.; FRAGOSO, W.D. Classificação periódica: Um exemplo didático para ensinar análise de componentes principais. **Química Nova**, v.33, p.1594-1597, 2010.

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).

MACEDO, N., A.C. MENDES & P.S.M. BOTELHO. **Susceptibility of the main commercial sugarcane varieties to the moth borer, *Diatraea saccharalis* (Fabr., 1974) in the Southern-Central Brazil**, p.693-701. In Proc. Congr. Int. Soc. Sug. Cane Tech, 16, São Paulo, 1977.

MACEDO, N.; BOTELHO, P. S. M. Controle integrado da broca da cana-de-açúcar *Diatraea saccharalis* (Fabricius, 1794) (Lepidoptera, Pyralidae). **Brasil Açucareiro**, Rio de Janeiro, v.162, n. 2, p.2-11, 1988.

MACEDO, N.; LAVORENTI, N. Novo método de amostragem de intensidade de infestação da broca da cana-de-açúcar (Diatraea saccharalis). STAB, v.22, n.3, p.32-41, 2004.

MACHADO, L. A.; HABIB, M. **Perspectivas e impactos da cultura de cana-de-açúcar no Brasil.** 2009. Artigo em Hypertexto. Disponível em: http://www.infobibos.com/Artigos/2009_2/Cana/index.htm>. Acesso em: 29 jan. 2019.

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.

MARTINS, I. T. A. **Qualidade da fibra de diferentes cultivares brasileiras de algodão e sua relação com as condições meteorológicas.** 2020. (Dissertação de Mestrado). Escola Superior de Agricultura "Luiz de Queiroz"- Universidade de São Paulo, Piracicaba, 2020.

MELO, L. A. S.; SILVEIRA-NETO, S.; VILLA NOVA, N. A.; REIS, P. R. Influência de elementos climáticos sobre a população de cigarrinhas-das-pastagens. **Pesquisa Agropecuária Brasileira**, Brasília, v. 19, n. 1, p. 9-19, 1984.

MELO, A. B. P de.; PARRA, J. R. P. Exigências térmicas e estimativa do número de gerações anuais de broca da cana-de-açúcar em quatro localidades canavieiras de São Paulo. **Pesquisa Agropecuária Brasileira**, Brasília, v. 23, n. 7, p. 691-696, 1988b.

MELÉM JÚNIOR, N. J.; FONSECA, I. C. B.; BRITO, O. R.; DECAËNS, T.; CARNEIRO, M. M.; MATOS, M. F. A.; GUEDES, M. C.; QUEIROZ, J. A. L.; BARROSO, K. O. Análise de componentes principais para avaliação de resultados analíticos da fertilidade de solos do Amapá. **Semina: Ciências Agrárias**, v.29, p.499- 506, 2008.

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.

MILLER, P.; LANIER, W.; BRANDT, S. **Using growing degree days to predict plant stages**. Bozeman: Montana State University, Extension, 2018. p. 1-7. Disponível em: http://www.msuextension.org/publications/agandnaturalresources/mt200103ag.pdf. Acesso em: 15 nov. 2020.

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.

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.

MOZAMBANI, A.E. et al. História e morfologia da cana-deaçúcar. In: SEGATO, S.V. et al. Atualização em produção de cana-de-açúcar. Piracicaba: **Cadernos Planalsucar**. 2006.p.11-18.

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

MCKINION, J. M. Getting started: basics of modeling strategies. In: GOODENOUGH, J. L.; MCKINION, J. M. (ed.). **Basics of Insect Modeling**, 1992, cap. 1.

NAVA, D. E.; HADDAD, M. L.; PARRA, J. R. P. Exigências térmicas, estimativa do número de gerações de Stenoma catenifer e comprovação do modelo em campo. **Pesquisa Agropecuária Brasileira, Brasília**, v. 40, p. 961-967, 2005.

NAKANO et al. **Entomologia agrícola**. Piracicaba: FEALQ, 2002, 920p.

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.

ODUM, E. P. **Ecologia.** Rio de Janeiro: Guanabara Koogan, 1988, p.184, 185 e 244.

OLIVEIRA, C. M.; AUAD, A. M.; MENDES, S. M.; FRIZZAS, M. R. Crop losses and the economic impact of insect pests on Brazilian agriculture. **Crop Protection**, Amsterdam, v. 56, p. 50-54, 2014.

PAPA, G. **Broca da cana-de-açúcar**. 2007. UNESP/ Ilha Solteira. Disponível em:<http://www.ilhasolteira.com.br/colunas/index.php?acao=verartigo&idartigo=119401242>. Acesso em: 29 jan. 2019.

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.

PINTO, A. de S. **Controle de pragas da cana-de-açúcar**. Sertãozinho: Biocontrol, 2006, 64p. (Boletim Técnico Biocontrol, n.1).

PINTO, A. de S.; BOTELHO, P. S. M.; OLIVEIRA, H. N. **Guia ilustrado de pragas e insetos benéficos da cana-de-açúcar**. Piracicaba: CP 2, 2009, 160 p.

PRECETTI, A. A. C. M.; TERAN, F. O.; SANCHEZ, A. G. Alterações nas características tecnológicas de algumas variedades de cana-de-açúcar, devidas ao dano da broca *Diatraea saccharalis*. **Boletim Técnico Copersucar**, v. 41, p. 3-8, 1988.

RAVINDRAN, A., PHILLIPS, D. T., SOLBERG, J. J. **Operations Research Principles and Practice**, 2. ed, 1987, 656 p.

R Core Team (2021) R: **A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.r-project.org/>. Acesso em: 20 jan. 2020.

RÉAUMUR, M. Observations du thermometre, faites à Paris pendant l'année - Comparées avec celles qui ont été faites sous la ligne; à l' Ile de France, à Alger et en quelques-unes de nos îles de l' Amérique. **Académie Royale des Sciences**, Paris, p.545-576, 1735.

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

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

ROSSATO JÚNIOR, J. A. S. ***Diatraea saccharalis* (Fabr.) (Lepidoptera:Crambidae) e *Mahanarva fimbriolata* (Stal) (Hemiptera: Cercopidae) em cana-de-açúcar: impacto na qualidade da matéria-prima, açúcar e etanol.** 2012. Tese (Doutorado em Entomologia Agrícola) - Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Jaboticabal-SP.

ROSSATO JÚNIOR, J. A. S., COSTA, G. H. G., MADALENO, L. L., MUTTON, M. J. R., HIGLEY, L. G., FERNANDES, O. A. Characterization and impact of the sugarcane borer on sugarcane yield and quality. **Agronomy Journal**, Madison, v. 105, p. 643–648, 2013.

SANDOVAL, ; SENÔ, K.C.A. Comportamento e controle da Diatraea saccharalis na cultura da cana-de-açúcar. Revista Nucleus, v. 7, n. 1, p. 1-16, 2010.

SANGUINO, A. Situação atual da pesquisa em doenças da cana-de-açúcar. **Summa Phytopathologica**. v.24, n.1, p.90-91, 1988.

SANTOS, A. S. Doenças causadas por fungos e bactérias em cana-de-açúcar. In: REUNIÃO ITINERANTE DE FITOSSANIDADE DO INSTITUTO BIOLÓGICO – RIFIB, 9, Catanduva, SP. **Anais...** p.10-17, 2003.

SANTOS, R. F.; CARLESSO, R. Déficit hídrico e os processos morfológico e fisiológico das plantas. **Revista Brasileira de Engenharia Agrícola e Ambiental**. Campina Grande, v. 2, n. 3. p. 287-294, 1998.

SAYAD, S. **Support vector machines**. Toronto: University of Toronto, 2010. Disponível em: . Acesso em: 10 abr. 2016.

SEGATO, S. V.; MATTIUZ, C. F. M.; MOZAMBANI, A. E. Aspectos fenológicos 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: ESALQ, 2006. cap. 2, p. 19-36.

SENTELHAS, P.C.; SANTOS, D.L.; MONTEIRO, L.A.; SOARES-COLLETTI, A.R.; PALLONE FILHO, W.J.; DONZELLI, J.L.; ARRIGONI, E.B. Agro-climatic favorability zones for sugarcane orange rust as a tool for cultivar choice and disease management. **Crop Protection**. Amsterdan, v. 84, p. 88-97, 2016.

SILVA, L. M.; SILVA, M. C.; SILVA, S. M. F. A.; ALVES, R. C.; SIQUEIRA, H. A. A.; MARQUES, E. J. Toxin gene contentes and activity of Bacillus thuringiensis strains against two sugarcane borer species, Diatraea saccharalis (F.) and Diatraea flavipennella (Box). **Neotropical entomology**, Londrina, v. 47, p. 292-301, 2018.

SILVEIRA-NETO, S. NAKANO, O.; BARBIN, D.; VILLA NOVA, N. A. **Manual de Ecologia dos insetos**. São Paulo: Ed. Ceres, 1976, 419 p.

SUTHERST, R. W.; MAYWALD, G. F. A computerised system for matching climates in ecology. **Agriculture, Ecosystems and Environment**, v. 13, p. 281-299, 1985.

SCHOWALTER, T. D. **Insect ecology: an ecosystem approach**. San Diego: Elsevier, 2006, 572 p.

SCOMPARIN, A. L. X. **Desenvolvimento de diferentes populações de *Cydia pomonella* (Lepidoptera: Tortricidae) em temperaturas variáveis e consequências na modelagem fenológica.** 2009. Tese (Doutorado em Entomologia Agrícola) – Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Jaboticabal-SP.

SMERAGE, G. H. Representation of development in insect population models. In: GOODENOUGH, J. L.; MCKINION, J. M. (Ed.). **Basics of Insect Modeling**, 1992, cap. 8.

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.** 2008. 99 p. Dissertação (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2018.

TAIZ, L.; ZEIGER, E. **Fisiologia vegetal**. 4. ed. Porto Alegre: Artmed, 2009. 819 p.

TERAN, F. O. Dinâmica populacional de adultos de *Diatraea saccharalis* (Fabricius,1794) em canaviais do estado de São Paulo. **Anais da Sociedade Entomológica do Brasil**, Londrina, v. 8 n. 1, 1979.

TERAN, F. O.; PRECETTI, A. A. C. M.; DERNEIKA, O. Broca da cana-de-açúcar *Diatraea saccharalis.* In: Reuniáo Técnica Agronômica: Pragas da Cana-de-Açúcar, 1., 1983, Piracicaba, **Anais...** Piracicaba: Copersucar, 1983, p. 4-15.

THORNTHWAITE, C.W.; MATHER, J.R. **The water balance**. Publications in Climatology, New Jersey, Drexel Institute of Tecnology, 104p., 1955.

UNICA – **União da Indústria de Cana-de-açúcar**. 2012. Disponível em: <www.unica.com.br>. Acesso em: 29 jan. 2019.

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.

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.

WILLMOTT, C. J. On the validation of models. **Physical Geography**, 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.

YAMADA, E. S. M**. Zoneamento agroclimático da Jatropha curcas L. como subsídio ao desenvolvimento da cultura no Brasil visando à produção de biodiesel**. 2011. 135 f. Dissertação (Mestrado em Física do Ambiente Agrícola) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba. Disponível em: Acesso em: 30 jan. 2019.

YAMADA, E. S. M.; SENTELHAS, P. C. Agro-climatic zoning of *Jatropha curcas* as subside for crop planning and implementation in Brazil. **International Journal of Biometeorology**, v. 58, n. 9, p. 1995-2010, 2014.

YUKAWA, J.; ICHINOSE, M.; KIM, W.; UECHI, N.; GYOUTOKU, N.; FUJII, T. Lower development threshold temperatures and thermal constants for four species of Asphondylia (Diptera: Cecidomyiidae) in Japan and their larval developmental delay caused by heat stress**. Applied Entomology and Zoology**, Japan, v. 51, n.1, p. 71- 80, 2016.