UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA AGRÍCOLA

AVALIAÇÃO DA INFLUÊNCIA DA TEMPERATURA E DA PRECIPITAÇÃO NA OCORRÊNCIA DA FERRUGEM ASIÁTICA DA SOJA POR MEIO DA TÉCNICA DE ÁRVORE DE DECISÃO

GUILHERME AUGUSTO SILVA MEGETO

CAMPINAS
JULHO 2012

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA AGRÍCOLA

AVALIAÇÃO DA INFLUÊNCIA DA TEMPERATURA E DA PRECIPITAÇÃO NA OCORRÊNCIA DA FERRUGEM ASIÁTICA DA SOJA POR MEIO DA TÉCNICA DE ÁRVORE DE DECISÃO

Dissertação de Mestrado submetida à banca examinadora para obtenção do título de Mestre em Engenharia Agrícola, na área de concentração de Planejamento e Desenvolvimento Rural Sustentável.

GUILHERME AUGUSTO SILVA MEGETO

Orientador: PROF. DR. STANLEY ROBSON DE MEDEIROS OLIVEIRA
Co-orientador: DR. CARLOS ALBERTO ALVES MEIRA

CAMPINAS
JULHO 2012

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

M472a

Megeto, Guilherme Augusto Silva, 1984-
Avaliação da influência da temperatura e da precipitação na ocorrência da ferrugem asiática da soja por meio da técnica de árvore de decisão / Guilherme Augusto Silva Megeto. --Campinas, SP: [s.n.], 2012.

Orientador: Stanley Robson de Medeiros Oliveira
Coorientador: Carlos Alberto Alves Meira.
Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mineração de dados (Computação). 2. Doenças das plantas - Epidemiologia. 3. Agricultura - Previsão. 4. Plantas - Doenças e pragas - Controle. 5. Sistemas de suporte de decisão. I. Oliveira, Stanley Robson de Medeiros. II. Meira, Carlos Alberto Alves. III. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. IV. Título.

Título em Inglês: Evaluation of the influence of temperature and precipitation in the occurrence of Asian soybean rust by using the technique of decision tree
Palavras-chave em Inglês: Data mining (computing), Plant diseases - Epidemiology, Agriculture - forecast, Plants - Diseases and pests - Control, Decision support systems
Área de concentração: Planejamento e Desenvolvimento Rural Sustentável
Titulação: Mestre em Engenharia Agrícola
Banca examinadora: Emerson Medeiros Del Ponte, Luiz Henrique Antunes Rodrigues
Data da defesa: 10-07-2012
Programa de Pós Graduação: Engenharia Agrícola

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Guilherme Augusto Silva Megeto**, aprovado pela Comissão Julgadora em 10 de julho de 2012, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Prof. Dr. Stanley Robson de Medeiros Oliveira – Presidente e Orientador
Feagri/Unicamp

Prof. Dr. Emerson Medeiros Del Ponte - Membro Titular
UFRGS

Prof. Dr. Luiz Henrique Antunes Rodrigues – Membro Titular
Feagri/Unicamp

Dedico
aos meus pais e avós,
pelo amor, carinho, educação,
princípios e valores.

AGRADECIMENTOS

Aos orientadores Prof. Dr. Stanley Robson de Medeiros Oliveira e Dr. Carlos Alberto Alves Meira pela orientação, oportunidade de crescimento profissional e pessoal, dedicação e incentivo à pesquisa multidisciplinar.

À CAPES pelo suporte financeiro.

À Embrapa Informática Agropecuária, especialmente ao Laboratório de Inteligência Computacional (LabIC), pela oportunidade de utilizar suas dependências físicas e a infra-estrutura computacional durante a realização do projeto.

Aos pesquisadores Prof. Dr. Emerson Medeiros Del Ponte, Prof. Willingthon Pavan e Adriano Otavian pelo fornecimento dos dados.

Ao colega Thiago Veloso dos Santos e ao grupo de epidemiologia de plantas da UFRGS (Universidade Federal do Rio Grande do Sul) pela vivência e informações sobre epidemiologia de plantas.

A todos os demais professores, pesquisadores e funcionários da Faculdade de Engenharia Agrícola, da Universidade Estadual de Campinas e da Embrapa Informática Agropecuária que me ajudaram, direta ou indiretamente.

Aos colegas do LabIC e à amiga Raquel Boschi, pelas dicas de resolução de problemas, conversas, almoços e cafés.

Aos amigos da República pela excelente convivência diária e aos familiares e amigos de Jundiaí pela compreensão e apoio.

Aos meus pais, Anaí Silvia e Luiz Antonio, e aos meus avós Ada e Sylvio (sempre presente), pelo apoio incondicional, carinho, amor, educação, princípios e valores para a vida.

Em especial, para Sheila, por tornar tudo melhor.

“Deveríamos aprender como se fôssemos viver para sempre,
e viver como se fôssemos morrer amanhã.”

RESUMO

A ferrugem asiática, causada pelo fungo *Phakopsora pachyrhizi*, atualmente é considerada uma das doenças mais importantes e agressivas da soja. A principal forma de controle é a aplicação calendarizada de fungicidas a qual desconsidera o risco de ocorrência da doença. Estudos epidemiológicos buscam compreender os fatores que influenciam na ocorrência e desenvolvimento das epidemias, especialmente aqueles relacionados ao ambiente tais como condições meteorológicas. Com o avanço da tecnologia da informação e do armazenamento de dados, técnicas de mineração de dados (*data mining*) apresentam-se promissoras para a descoberta de conhecimento em bases de dados epidemiológicos. Este trabalho tem como objetivo avaliar a influência da chuva e da temperatura na ocorrência da ferrugem asiática da soja utilizando árvores de decisão. Para tal, foram obtidos dados de ocorrências da doença em quatro safras, de 2007/2008 a 2010/2011, oriundos do banco de dados do Consórcio Antiferrugem, e dados meteorológicos, provenientes do sistema Agritempo. A análise exploratória dos dados permitiu obter subsídios para compor o conjunto de dados final e definir o escopo deste trabalho, buscando-se características intrínsecas à doença e sua interação com o ambiente, utilizando apenas variáveis de base meteorológica. As variáveis utilizadas foram relacionadas à precipitação e à temperatura, que deram origem a nove atributos avaliados para cada período temporal. Os atributos meteorológicos foram relacionados com o evento de ocorrência (Oc) e de não ocorrência (NaoOc) da doença (assumido como o trigésimo dia anterior ao evento da ocorrência). Os resultados englobam um modelo preditivo e um modelo interpretativo para classificar eventos de ocorrências e de não ocorrências da doença. O modelo preditivo utilizou 46 atributos e 12.591 registros, e teve uma taxa de acerto de 79,52%, avaliada por validação cruzada. O modelo interpretativo foi baseado no modelo preditivo, excluindo-se *outliers* e realizando podas no modelo, com o objetivo de reduzir o número de regras para interpretá-lo visualmente. As regras obtidas consideraram mais os atributos relacionados com a temperatura do que com a precipitação. Os valores detectados das temperaturas apresentaram coerência com os valores que favorecem ou não os processos de infecção da planta encontrados na literatura. O processo de mineração de dados e as vantagens preditivas e interpretativas de árvores de decisão mostraram-se capazes de

evidenciar o conhecimento acerca da influência meteorológica na ferrugem asiática da soja e predizer razoavelmente os eventos de ocorrência da doença no campo, podendo ser úteis em avaliações de risco visando o apoio à tomada de decisão quanto à aplicação de fungicidas.

PALAVRAS-CHAVE: Mineração de dados (Computação); Doenças das plantas – Epidemiologia; Agricultura – Previsão; Plantas - Doenças e pragas – Controle; Sistemas de suporte de decisão

ABSTRACT

The Asian soybean rust, caused by *Phakopsora pachyrhizi*, is now considered one of the most important and aggressive diseases of soybean. The main form of control is the scheduled application of fungicides which disregards the the risk of disease occurrence. Epidemiological studies seek to understand the factors that influence the occurrence and development of epidemics, especially those related to the environment such as weather conditions. With the development of information technology and data warehousing, data mining techniques appear to be promising for knowledge discovery in epidemiological databases. This study aims to evaluate the influence of rainfall and temperature on the occurrence of soybean rust by using decision trees models. To accomplish that, data of the occurence of the disease were collected from four seasons, 2007/2008 to 2010/2011, from the Consórcio Antiferrugem and weather data from the Agritempo system. Exploratory data analysis allowed for obtaining subsidies to generate the final data set and define the scope of this work, seeking intrinsic characteristics of the disease and its interaction with the environment, using only meteorological variables. The variables used were related to precipitation and temperature, resulting into nine attributes evaluated in different periods. Such attributes were related to the event of occurrence (Oc) and non occurrence (NaoOc) of the disease (assumed as the thirtieth day prior to the event of occurrence). The results include a predictive model and an interpretive model for classifying events of occurrences and non occurrences of the disease. The predictive model used 46 attributes and 12,591 records, and yielded an accuracy rate of 79.52%, being evaluated by cross-validation. The interpretive model was based on the predictive model, excluding outliers and performing pruning on the model, with the goal of reducing the number of rules to interpret it visually. The rules were obtained using more attributes related to temperature than to precipitation. The detected temperature values were consistent with the values that favor or not the processes of plant infection founded in the literature. The attributes related to the precipitation did not show good agreement with existing knowledge. The process of data mining and the predictive and interpretive advantages of decision trees shown to be able to demonstrate knowledge about the influence of weather on soybean rust and predict with acceptable accuracy rate their

occurrences in the field, and that could be useful in risk evaluations that aims to support decision making regarding the application of fungicides.

KEYWORDS: Data mining (computing), Plant diseases - Epidemiology, Agriculture - forecast, Plants - Diseases and pests - Control, Decision support systems

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

1 INTRODUÇÃO

O cultivo de soja é uma das atividades econômicas mais expressivas do agronegócio, tanto no contexto nacional quanto no internacional. Embora o Brasil possua apenas 5,8% de sua área agropecuária destinada à sojicultura, esta contribui com cerca de 27,1% e 39,0%, respectivamente, da produção e da exportação mundiais de soja em grãos (DALL'AGNOL et al., 2010).

Dentre os principais fatores que podem limitar a produção de soja estão as doenças e pragas. No Brasil, já foram identificadas 40 doenças causadas por bactérias, fungos, nematóides e vírus. A importância de cada doença varia de ano a ano e de região para região (EMBRAPA, 2010); nos últimos anos, no entanto, a ferrugem asiática se destacou por ser uma doença muito severa e que se disseminou rapidamente no Brasil.

A ferrugem asiática, causada pelo fungo *Phakopsora pachyrhizi*, é uma doença extremamente agressiva, que inicialmente apresenta pequenas lesões foliares de coloração castanha, que com o seu desenvolvimento pode causar o amarelecimento e desfolha precoce, comprometendo a formação e o enchimento das vagens e diminuindo a produção. No Brasil, o fungo foi identificado pela primeira vez em 2001, no estado do Paraná (YORINORI et al., 2005). Como exemplo do impacto econômico causado pela ferrugem asiática, o custo total estimado da ferrugem, que abrange tanto perdas na produção quanto custos de aplicações de defensivos, atingiu a marca de US$ 2 bilhões na safra 2003/2004 e prosseguiu aumentando até a safra 2007/2008 (DEL PONTE e MARTINS, 2008).

Pesquisas sobre cultivares resistentes à doença estão sendo desenvolvidas, porém, devido à grande variabilidade do patógeno, ainda não foram obtidos resultados definitivos. Atualmente, a tática de controle mais utilizada é o uso de fungicidas, aliada à adoção do "vazio sanitário" em escala regional, que é um período de 90 dias no qual deve haver ausência de plantas de soja entre a colheita e o plantio, a fim de evitar a sobrevivência do fungo (HENNING e GODOY, 2006).

Embora seja amplamente utilizado, os fungicidas nem sempre são utilizados de maneira racional. Segundo Godoy et al. (2009), as aplicações de fungicidas são realizadas de forma calendarizada, com início em um determinado estádio fenológico de florescimento e

aplicações subsequentes em intervalos de 14 a 21 dias. No entanto, este sistema não leva em consideração diversos fatores que influenciam a doença, podendo gerar aplicações desnecessárias ou mesmo atrasadas quando a doença ocorre precocemente.

O conhecimento sobre a epidemiologia da ferrugem asiática tem sido sumarizado em modelos matemáticos que buscam descrever de maneira quantitativa o desenvolvimento da doença, possibilitando assim prever seu comportamento no futuro (DEL PONTE et al., 2006a).

As duas principais estratégias de modelagem da ferrugem asiática são: (i) utilizar dados provenientes de ambientes controlados; e (ii) utilizar dados de observações no campo. A partir desses dados podem ser aplicadas diversas técnicas de modelagem, sendo que Del Ponte et al. (2006a) fizeram uma revisão dos principais modelos desenvolvidos para a ferrugem asiática que utilizam ambas as estratégias mencionadas. Tais modelos podem utilizar abordagens tradicionais, como é o caso da regressão linear (DEL PONTE et al., 2006b), e também técnicas mais complexas, como redes neurais (BATCHELOR et al., 1997) e lógica fuzzy (KIM et al., 2005).

Porém, ainda existem lacunas no conhecimento acerca de alguns aspectos da ferrugem asiática, além de técnicas e dados ainda não explorados. O uso de estatísticas descritivas e técnicas tradicionais em conjuntos de dados podem, eventualmente, não evidenciar informações úteis ocultas nos dados (FAYYAD et al., 1996).

O Consórcio Antiferrugem (CAF) possui um banco de dados extenso, com informações de ocorrências da doença em plantações das principais áreas de cultivo, desde a safra de 2004/2005 (www.consorcioantiferrugem.net). Esses dados podem abrigar algum conhecimento novo e ser evidenciado a partir de aplicações de técnicas de mineração de dados.

Além das informações disponíveis no CAF, os dados meteorológicos nos períodos de ocorrência da ferrugem também podem contribuir para a descoberta de conhecimento, uma vez que as condições do ambiente afetam o desenvolvimento das plantas, dos patógenos e suas interações (COAKLEY, 1988).

A área de mineração de dados, muitas vezes considerada como sinônimo de descoberta de conhecimento em bases de dados, abrange técnicas conhecidas por serem

capazes de lidar com grandes quantidades de dados, extrair informações destes e transformá-los em conhecimento, ao passo que, caso fossem utilizadas outras técnicas, permaneceria oculto.

Dentre as técnicas de mineração de dados, destacam-se os algoritmos de árvores de decisão (QUINLAN, 1993), principalmente pelo potencial interpretativo do modelo simbólico gerado, uma vez que é possível visualizar as decisões do algoritmo em relação aos atributos, e então traduzir o modelo gerado como regras nos moldes de causa e efeito (se - então) (APTÉ e WEISS, 1997).

Tendo em vista o potencial de descoberta de conhecimento na base de dados do Consórcio Antiferrugem, a hipótese deste trabalho é que um processo de mineração de dados, utilizando algoritmos de árvores de decisão, aplicado em um conjunto de dados contendo registros de ocorrências da ferrugem asiática da soja e atributos meteorológicos, gerará modelos capazes de identificar os principais fatores que influenciam a ocorrência da doença no campo.

Este trabalho está organizado na forma de capítulos. No Capítulo 2 apresenta-se a revisão de literatura, que abrange os assuntos relativos à soja, à doença ferrugem asiática e aos conceitos de mineração de dados. O Capítulo 3 descreve o material e os métodos utilizados, desde a aquisição dos dados até a descrição do procedimento para a análise dos resultados. No Capítulo 4, os resultados são apresentados e discutidos com base na literatura. Finalmente, no Capítulo 5, as conclusões são expostas a partir da análise dos dados e dos modelos obtidos, assim como as sugestões de realização de trabalhos futuros.

1.1 Objetivos

O objetivo geral deste trabalho é:

•Executar e avaliar um processo de mineração de dados, utilizando o potencial interpretativo da técnica de indução de árvore de decisão para identificar a influência da temperatura e da precipitação nos eventos de ocorrência da ferrugem asiática da soja.

1.2 Objetivos específicos

•Definir o conjunto final de dados de ocorrência a serem utilizados no processo de modelagem bem como as variáveis meteorológicas com potencial preditivo baseando-se na análise exploratória.

•Desenvolver e interpretar os cenários gerados pelos modelos resultantes e identificar as condições meteorológicas que favoreçam ou não a ocorrência da ferrugem asiática da soja.

2 REVISÃO BIBLIOGRÁFICA

2.1 A ferrugem asiática da soja

2.1.1 Importância e epidemiologia da ferrugem asiática

A soja [*Glycine max* (L.) Merrill] é atacada por muitos patógenos (fungos, bactérias, vírus e nematóides), porém, dentre todas as doenças, a ferrugem asiática, causada pelo fungo *Phakopsora pachyrhizi* Syd. & P. Syd., tem sido considerada uma das doenças mais importantes (HENNING e GODOY, 2006).

Sua primeira ocorrência registrada foi em 1902 no Japão e então atingiu diversos países na primeira metade do século XX, passando por alguns da Ásia, assim como Austrália e Índia. Nos anos 90, foi registrada a sua presença na África, e em 2001 ocorreu pela pela primeira vez na América do Sul, sendo Brasil e Paraguai os países atingidos (YORINORI et al., 2005). O avanço temporal e espacial da ferrugem asiática pelo mundo pode ser acompanhado na Figura 2.1.

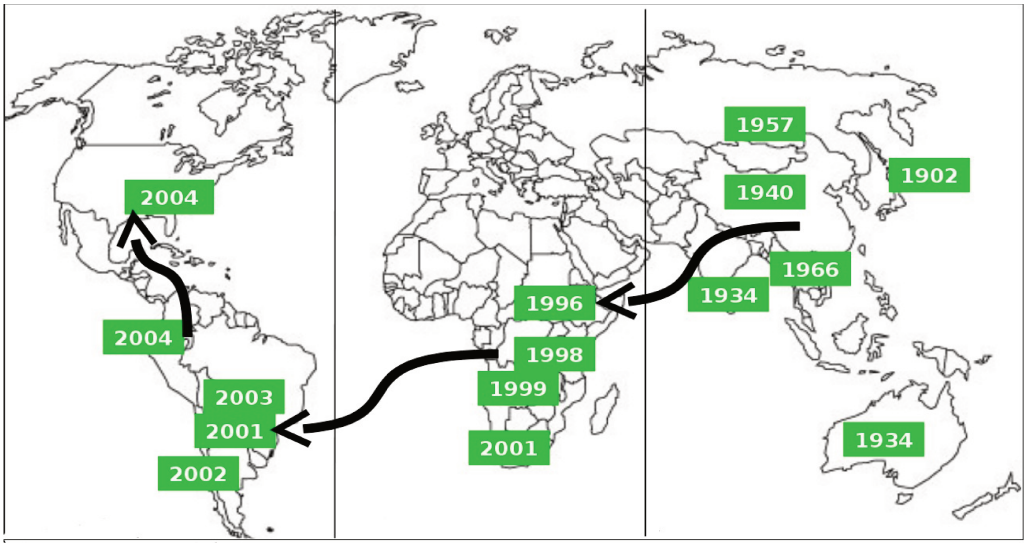


Figura 2.1: Dispersão espaço-temporal da ferrugem asiática no mundo. Fonte: www.consorcioantiferrugem.net

No Brasil, a doença se disseminou rapidamente para todas as regiões produtoras nos três anos posteriores à sua detecção. A partir da safra 2003/2004, os custos com a ferrugem atingiram a marca de US$ 2 bilhões, o que garantiu à ferrugem asiática o título de maior problema fitossanitário da cultura (YORINORI et al., 2005). Segundo Del Ponte e Martins (2008), na safra de 2007/2008 as perdas não passaram de 1% da produção nacional, porém, os custos com o controle, ou seja, aquisição e aplicação de fungicidas, representaram cerca de 80% do custo total. A Figura 2.2 ilustra o custo da ferrugem asiática no Brasil ao longo do tempo, a partir de dados consultados no *site* do Consórcio Antiferrugem (www.consorcioantiferrugem.net).

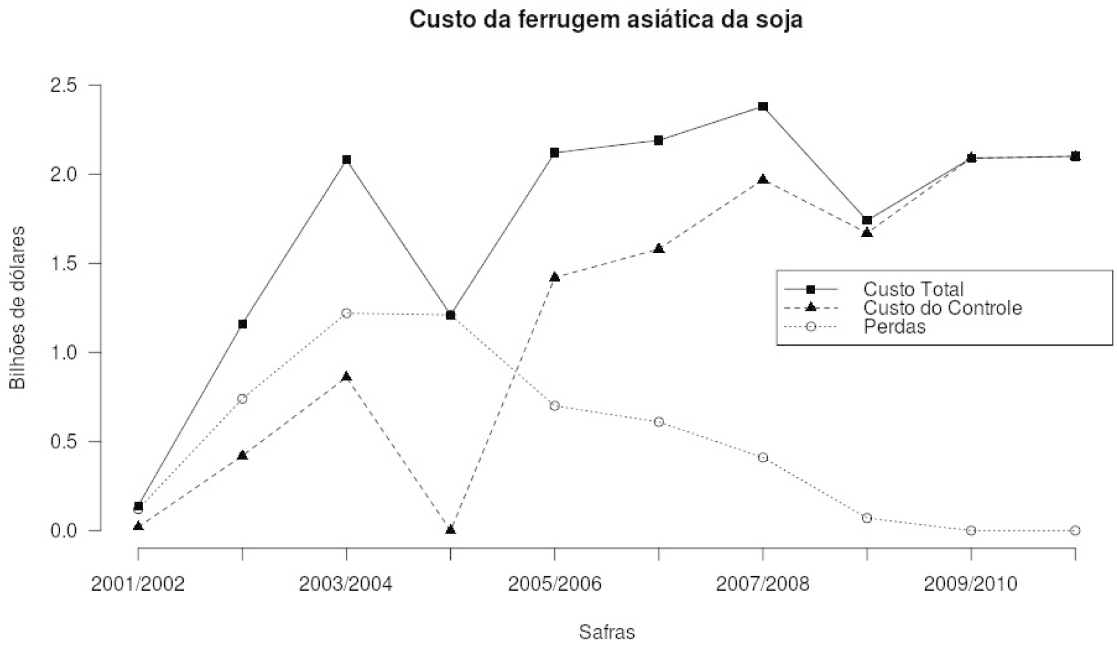


Figura 2.2: Custos da ferrugem asiática da soja no Brasil incluindo perdas em produtividade e gastos com aplicações de fungicidas. Fonte dos dados: www.consorcioantiferrugem.net

Com o agravamento do problema devido à falta de conhecimento sobre as estratégias de manejo da doença, em 2004 foi criado o Consórcio Antiferrugem (CAF), uma iniciativa público-privada que visa reunir, organizar e uniformizar o conhecimento gerado pela pesquisa e então divulgar as informações sobre a doença.

Desde então, o CAF tem cumprido seu papel e usa como meio principal de divulgação um *website* na Internet (http://www.consorcioantiferrugem.net). Nesse *site*, uma

rede de laboratórios credenciados cadastram as ocorrências de ferrugem em todas as regiões com o objetivo de monitorar e alertar quanto à presença do inóculo da doença nas regiões do Brasil. Cada ocorrência registrada é incluída na base de dados, que possui informações como a data da ocorrência, estádio da planta, tipo de cultivar etc., e também mantém relatórios, palestras e documentos disponíveis com todas as informações atuais sobre a ferrugem asiática.

Quanto ao manejo da doença, o desenvolvimento de cultivares resistentes ainda está em fase de estudos, o que faz com que o uso de fungicidas seja uma alternativa que viabilize o cultivo da soja na presença da doença. Porém, há a necessidade de se utilizar os defensivos químicos de maneira racional, tanto para diminuir os custos de produção, como por exemplo, gastos com produtos químicos e mão de obra para aplicação, quanto para a preservação do meio ambiente e da sociedade, evitando a contaminação do solo, da água e dos alimentos.

Atualmente, as aplicações de fungicidas que mantém um bom controle sobre a doença são realizadas de forma calendarizada, com início no estádio de florescimento e se repetindo em intervalos de 14 a 21 dias. Porém, esse método ignora fatores que influenciam as epidemias, o que pode resultar em aplicações desnecessárias (GODOY et al., 2009).

2.1.2 Sintomatologia, ciclo da doença e epidemiologia

Os sintomas iniciais podem aparecer em qualquer estádio de desenvolvimento da planta na forma de minúsculos pontos protuberantes (urédias) mais escuros que o tecido foliar sadio, na face inferior da folha, vistos mais facilmente contra um fundo claro. Progressivamente, as urédias adquirem cor castanho-clara e liberam os esporos que são facilmente espalhados pelo vento e chuva (EMBRAPA, 2010).

O aumento da intensidade da doença causa o amarelecimento e queda prematura das folhas, o que gera deficiência na fase de formação e enchimento dos grãos, causando perda na qualidade e no rendimento da produção. Em casos severos, pode causar o aborto e queda das vagens, resultando em perda total do rendimento (EMBRAPA, 2010).

O ciclo da ferrugem asiática da soja abrange diversas fases, onde, inicialmente, o fungo se dissemina através de esporos, que são facilmente levados pelo vento ou pelo impacto das gotas de chuva e se depositam em folhas das plantas de soja. Com a temperatura e molhamento foliar ideais, os esporos germinam e o fungo penetra na folha diretamente,

rompendo a epiderme. Logo são formadas as urédias (protuberâncias) que passam a produzir os uredósporos, que são disseminados pelo vento, e assim o ciclo é completado. O ciclo de vida simplificado da doença pode ser ilustrado como na Figura 2.3.

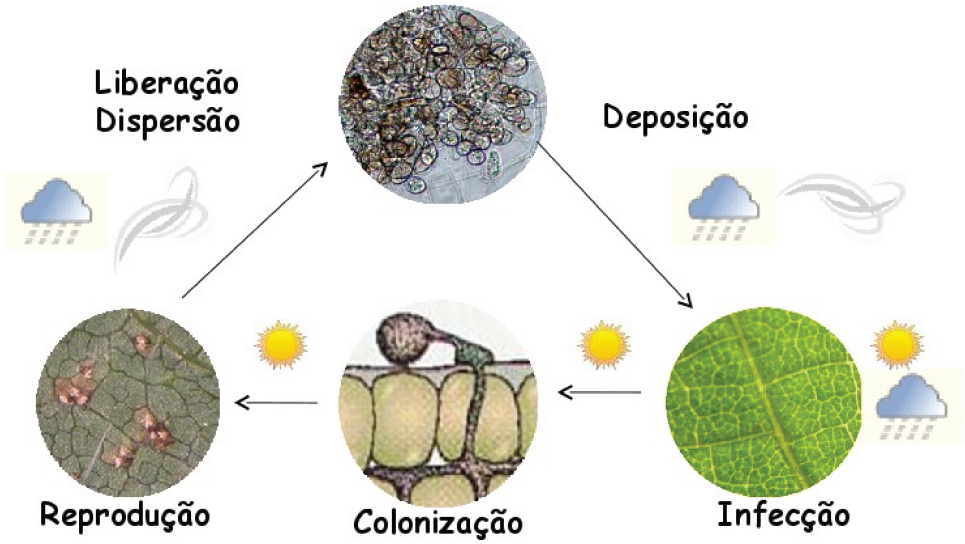


Figura 2.3: Ciclo da ferrugem asiática da soja. Fonte: www.consorcioantiferrugem.net

Em cada fase do ciclo da doença, diferentes fatores ambientais exercem influência sobre o comportamento do patógeno e sua interação com a planta. Nesse contexto, na década de 70, com o aumento da produção de soja nos EUA, começaram a ser realizados diversos estudos a fim de se conhecer melhor o patógeno *Phakopsora pachyrhizi* e as condições ambientais para o desenvolvimento da ferrugem asiática da soja. Nesses estudos e em outros mais recentes, os resultados abordaram a influência ambiental na infecção, colonização e desenvolvimento da doença, além de fatores aerobiológicos como liberação, dispersão e deposição de esporos (DEL PONTE e ESKER, 2008).

Dentre os resultados, destacam-se os obtidos para a infecção, que é uma sequência de eventos onde é necessário água livre para ocorrer a germinação de uredósporos, formação de tubo germinativo e depois um apressório. Em experimentos realizados em casas de vegetação e estufas, em condições úmidas, após a deposição do esporo na superfície da folha, a germinação pode iniciar em uma hora e encerrar em seis a oito horas (MELCHING et al.,

1989; ALVES et al., 2006). Se a água não for um fator limitante, a germinação é influenciada pela temperatura (DEL PONTE e ESKER, 2008).

Marchetti et al. (1976) observaram que o fungo foi capaz de germinar entre 7 e 28ºC em água-agar, e o intervalo de temperatura ótimo foi entre 15 e 25ºC. Em experimentos realizados por Alves et al. (2006) utilizando isolados do Brasil, houve germinação dos esporos em temperaturas variando de 8 a 30ºC e o intervalo ótimo foi de 15 a 25ºC.

Bonde et al., (2007) buscaram verificar se havia diferenças no efeito da temperatura na germinação e crescimento do tubo germinativo entre isolados de *P. Pachyrhizi* de diferentes regiões, como Taiwan, Zimbábue, Havaí e Brasil. Nesse estudo, não foram encontradas diferenças significativas entre as respostas da germinação, crescimento do tubo germinativo e início da infecção. O intervalo de temperatura ótimo para todos esses processos foi de 17 a 28ºC.

A infecção se estabelece com o mínimo de 6 horas de molhamento foliar a uma temperatura de 24ºC (KITANI e INOUE, 1960 apud DEL PONTE e ESKER, 2008). Em 18ºC ou 26,5ºC , é necessário um período maior de molhamento foliar (MARCHETTI et al., 1976; MELCHING et al., 1989) e temperaturas acima de 28ºC são prejudiciais para a infecção (MARCHETTI et al., 1976). Em temperaturas abaixo de 15ºC ou acima de 30º, não houve infecção (CALDWELL et al., 2005 apud DEL PONTE e ESKER, 2008). Alves et al. (2006) relataram que o fungo não conseguiu infectar a planta em temperaturas acima de 30ºC, e na temperatura de 10ºC os sintomas demoraram 20 dias para aparecer.

O efeito da temperatura no período latente foi estudado por Alves et al. (2006), que resultou na Equação 1 ($R^2 = 0,99$), em que Y é o período latente em dias e T é a temperatura (mantida constante no experimento) em graus Celsius (ºC).

$$Y = 0,11\,T^2 - 5,20\,T + 69,53 \quad (1)$$

As principais informações de estudos de campo foram obtidas em experimentos em Taiwan nos anos 80 (TSCHANZ et al., 1984 apud DEL PONTE et al., 2006a), nos quais os resultados indicam que temperaturas noturnas abaixo de 14ºC reduzem ou previnem a infecção. Outra constatação foi que as primeiras chuvas e a quantidade de precipitação em épocas de seca apresentaram ser fatores importantes para o desenvolvimento da epidemia.

2.1.3 Modelagem e previsão da ferrugem asiática da soja

Um modelo pode ser definido como uma simplificação da realidade (MADDEN et al., 2007) ou uma representação simplificada de um sistema (BERGAMIN FILHO, 2006).

Os modelos geralmente são usados para descrição, entendimento, predição ou comparação de fenômenos. Os modelos abstratos são baseados em símbolos e regras que representam uma forma ou função de uma realidade em particular, como por exemplo, funções matemáticas, gráficos e desenhos esquemáticos. Esses modelos podem ser qualitativos ou quantitativos. Os modelos qualitativos buscam mostrar os principais componentes do fenômeno em estudo, mas sem utilizar equações, matemática ou estatísticas. Já os modelos quantitativos utilizam funções matemáticas, símbolos e regras para representar os aspectos de interesse (MADDEN et al., 2007).

No caso de modelagem de doenças de plantas, especificamente no caso de epidemias, há um grande interesse em determinar os elementos e condições que iniciam as epidemias, assim como os fatores que influenciam a taxa de aumento e a direção delas. Assim, são desenvolvidos experimentos, medições, fórmulas matemáticas com o objetivo de predizer a intensidade, a direção e o tempo de duração de uma epidemia em um determinado local (AGRIOS, 2004).

Alguns objetivos da previsão de doenças de plantas são: a redução de custos de produção, ao aplicar os agrotóxicos no momento correto; e a segurança, ao reduzir o número de aplicações de agrotóxicos, que não apenas reduz os riscos de uma possível intoxicação da plantação, mas também como o de pessoas e do meio ambiente (HARDWICK, 2006).

Muitos estudos e pesquisas são realizados para obter informações meteorológicas, ambientais e climáticas, buscando relacioná-las com o comportamento das doenças. Geralmente são utilizadas informações de precipitação, como frequência e intensidade, e em sistemas mais complexos, podem ser utilizadas temperaturas máxima, mínima e média, umidade relativa, molhamento foliar, velocidade do vento, entre outras (HARDWICK, 2006).

Estudos a respeito da influência das condições meteorológicas nas doenças de plantas indicam que a doença é mais afetada pelas condições microclimáticas no dossel das plantas do que pelas condições macroclimáticas. Entretanto, condições macroclimáticas produzem o microclima, e existe um limite na extensão com que o microclima pode facilitar o

desenvolvimento da doença sob condições macroclimáticas desfavoráveis. Além disso, é possível usar regras para determinar relacionamentos entre o macro e o microclima (COAKLEY, 1988).

Ao obter medidas suficientes para diversos fatores epidemiológicos, e em diferentes situações, é possível desenvolver um modelo matemático, geralmente uma ou mais equações, que descrevam a epidemia (AGRIOS, 2004).

Além de modelos matemáticos clássicos, nos últimos anos a área de tecnologia da informação também apresentou contribuições significativas na área de epidemiologia de doenças de plantas. Segundo Newton et al. (2006), uma das definições de tecnologia da informação na epidemiologia de doenças de plantas é: "o uso de tecnologia baseada em computação para coletar, processar e disseminar a informação e o conhecimento para aplicações para o estudo de epidemiologia de doenças de plantas".

As simulações de epidemias são feitas com o auxílio de computação e podem agregar informações de dinâmica populacional e epidemiológica. Geralmente, as aplicações de simulações são utilizadas para mapear e estimar o risco de epidemias nas principais regiões produtoras, e levam em consideração, principalmente, a presença da doença e variáveis ambientais (CANTERI et al., 2007).

Outros exemplos de aplicações do uso de tecnologia da informação na epidemiologia de doenças de plantas são sistemas de alertas, sistemas especialistas e sistemas de suporte à decisão, que, basicamente, auxiliam a tomada de decisão do produtor em relação ao momento de aplicação da tática de controle, geralmente o uso de fungicidas.

Para esses sistemas, as técnicas de modelagem também podem ser diferentes das convencionais, como por exemplo, o uso de correlação de variáveis e janelas temporais (COAKLEY et al., 1988), redes neurais para modelar o desenvolvimento da ferrugem asiática da soja (BATCHELOR et al., 1997) e árvores de decisão para a análise e o alerta da ferrugem do cafeeiro (MEIRA et al., 2008; 2009).

Diversos modelos matemáticos têm sido desenvolvidos por meio de diferentes técnicas de modelagem, variáveis e situações, a fim de se ter um melhor entendimento sobre os fatores que podem influenciar o avanço da ferrugem asiática da soja no tempo e no espaço, e, eventualmente, prever o comportamento futuro da doença.

Segundo Del Ponte et al. (2006a), em uma revisão de doze modelos existentes para a ferrugem asiática da soja, os autores agruparam os modelos em dois tipos quanto a abordagem adotada na sua construção: empíricos ou de simulação (mecanístico). Os modelos empíricos geralmente buscam relações estatísticas entre as variáveis utilizando dados experimentais. Já os modelos de simulação utilizam conceitos epidemiológicos para melhorar o conhecimento da estrutura e do comportamento do sistema biológico em questão.

Os modelos de simulação foram divididos em duas categorias: epidemiológicos ou aerobiológicos. Os epidemiológicos buscam simular os processos biológicos do ciclo da doença, enquanto que os aerobiológicos, geralmente mais complexos que os epidemiológicos, levam em consideração os fatores que afetam o transporte e a disseminação do inóculo da ferrugem asiática em longas distâncias (DEL PONTE et al., 2006a).

O primeiro modelo epidemiológico de simulação desenvolvido para a ferrugem asiática da soja foi o SOYRUST (YANG et al., 1991), calibrado e testado com dados de epidemia observados em Taiwan. O modelo simula de maneira dinâmica o aumento da severidade diariamente em duas variedades de soja utilizando fatores como taxa de infecção e período latente, que refletem as variações ambientais. Já mais recentemente, Pivonia e Yang (2006) ajustaram um modelo analítico para avaliar o potencial de estabelecimento da doença a partir da estimativa do aumento do número de unidades infecciosas. Para isso, esse modelo utiliza variáveis como eficiência da infecção, período latente, período infeccioso, número de esporos produzidos por lesão, entre outras, com parâmetros ajustados com base na literatura.

Os modelos aerobiológicos simulam o movimento de partículas, incluindo micro-organismos, na atmosfera, de uma região para outra. Dentre os modelos existentes, destacam-se os baseados no modelo atmosférico de transporte HYSPLIT (*Hybrid Single-Particle Lagrangian Integrated Trajectory*) (NOAA, 2011), e o SRAPS (*Soybean Rust Aerobiology Prediction System*) (ISARD et al., 2005).

O HYSPLIT é um modelo que simula trajetórias, dispersão, concentração e deposição de partículas originadas de uma certa localização geográfica e período do ano. Um exemplo de aplicação foi o modelo de Pan et al. (2006), que agregou informações como força da fonte de inóculo, produção de esporos, sobrevivência e deposição ao modelo HYSPLIT, além de ter utilizado dados de previsões climáticas a fim de predizer a disseminação de esporos no futuro.

Isard et al. (2005) também utilizaram informações de diversos estágios do processo aerobiológico, como por exemplo, a produção de esporos, mortalidade de esporos por exposição à radiação solar, deposição de esporos, entre outros, para desenvolver o modelo SRAPS.

Já os modelos classificados por Del Ponte et al. (2006a) como empíricos foram utilizados de três formas: (i) verificar e predizer períodos críticos ou favoráveis à infecção; (ii) descrever e predizer o desenvolvimento da doença; e (iii) predizer a severidade máxima. Por exemplo, Reis et al. (2004) propuseram um modelo, de base climática, a fim de predizer a probabilidade diária de infecção baseando-se em dados de período de molhamento foliar e temperatura durante esse período fornecidos pela literatura (MELCHING et al., 1989). Foi desenvolvida uma função que relaciona a intensidade da infecção (lesão.cm^{-1}) com as variáveis de período de molhamento foliar e temperatura nesse período, a qual gerou uma "tabela de períodos críticos". Em experimentos, este modelo foi integrado a estações agrometeorológicas automáticas instaladas na lavoura e, a partir dos dados obtidos, foi possível obter a probabilidade de infecção diária e emitir um alerta.

Outro modelo similar foi desenvolvido por Canteri et al. (2004) utilizando dados previamente analisados por Marchetti et al. (1976). Neste modelo, foi realizada uma regressão não-linear entre a intensidade da infecção em função do período de molhamento foliar e temperatura nesse período. Esta função foi utilizada para avaliação de risco de epidemia no estado do Paraná, a partir de dados meteorológicos de 6 anos, em Canteri et al. (2005).

Já modelos que predizem o desenvolvimento da ferrugem utilizaram redes neurais e lógica *fuzzy* como técnicas de modelagem. Batchelor et al. (1997) desenvolveram um sistema de redes neurais a fim de predizer a severidade diariamente. O modelo utilizou sete variáveis de entrada, dentre elas, o dia de plantio, dia da observação da ocorrência da doença, número de dias acumulados com umidade relativa do ar acima de 90%, graus-dia acumulados para o desenvolvimento da ferrugem asiática e graus-dia acumulados para o desenvolvimento da soja.

Kim et al. (2005) desenvolveram regras *fuzzy* a partir de conhecimento prévio da epidemiologia da ferrugem asiática da soja com o objetivo de se estimar a taxa de infecção aparente a partir de um conjunto de dados onde os valores (quantitativos) foram transformados

("fuzzyficação") em linguagem natural como "muito baixo" ou "muito alto" para dar entrada no sistema.

Finalmente, Del Ponte et al. (2006b) utilizaram dados de epidemias de ferrugem observadas nas safras 2002/2003 a 2004/2005 de 34 campos experimentais, de 21 locais diferentes do Brasil. A modelagem se baseou na técnica de regressão linear entre dados de severidade máxima da ferrugem asiática da soja e dados de variáveis meteorológicas de 30 dias após a data de detecção da doença. Embora este modelo tenha considerado dados de temperatura, essa variável não resultou em incremento no desempenho do modelo para predizer a severidade final da doença.

2.2 Informações meteorológicas para a modelagem

Existem diversos métodos para estabelecer as relações empíricas entre variáveis climáticas ou meteorológicas com dados de doenças de plantas. Em situações de experimentos em ambientes controlados, as condições ambientais são simuladas por equipamentos a fim de se ter um controle maior sobre as variáveis e sobre os efeitos que essas podem proporcionar na interação entre a planta e o patógeno.

Para estudar a ferrugem asiática da soja, diversos experimentos desse tipo foram realizados buscando-se relacionar a temperatura, molhamento foliar e outras variáveis ambientais com as diversas etapas do ciclo da doença, visando ao entendimento acerca dos fatores chave em cada processo (MELCHING et al., 1989; ALVES et al., 2006; MARCHETTI et al., 1976; BONDE et al., 2007).

Porém, as situações em ambientes fechados podem não representar adequadamente o comportamento das epidemias de doenças de plantas, uma vez que existem muitas variáveis ambientais atuando ao mesmo tempo, o que pode gerar respostas diferentes das obtidas pelos experimentos em ambientes fechados (DEL PONTE et al., 2008).

Existe uma quantidade menor de informações quantitativas sobre os efeitos de variáveis ambientais sob condições de campo do que quando comparada sob condições de câmara de crescimento ou estufas (DEL PONTE et al., 2008). Nesse caso, são necessários métodos e equipamentos para medir as variáveis meteorológicas no microambiente da cultura, o que é feito normalmente com estações meteorológicas de superfície, tanto automáticas

quanto convencionais. Uma estação meteorológica de superfície automática é composta de uma unidade de memória central ("data logger"), ligada a vários sensores dos parâmetros meteorológicos (pressão atmosférica, temperatura e umidade relativa do ar, precipitação, radiação solar, direção e velocidade do vento, etc), que, a cada hora, integra automaticamente os valores observados minuto a minuto (INMET, 2012).

Uma estação meteorológica convencional é composta de vários sensores isolados que registram continuamente os parâmetros meteorológicos (pressão atmosférica, temperatura e umidade relativa do ar, precipitação, radiação solar, direção e velocidade do vento etc.), que são lidos e anotados por um observador a cada intervalo, e posteriormente enviados a um centro coletor por um meio de comunicação qualquer (INMET, 2012).

Com o avanço da tecnologia, as estações meteorológicas automáticas estão ficando mais acessíveis para o produtor, porém, o que pode parecer uma vantagem na hora de coletar os dados mais próximos da plantação, pode ser arriscado quando não se tem uma formação técnica para instalar e calibrar esses equipamentos.

Segundo Gleason et al. (2008), os principais erros cometidos estão na instalação da estação automática no campo, uma vez que os sensores e instrumentos de medidas da estação devem estar rigorosamente posicionados e calibrados. Devem, também, passar por constantes averiguações ao longo do tempo para verificar se não há problemas.

Para evitar problemas relacionados à manutenção de equipamentos por pessoas não treinadas, foram criadas as redes de estações meteorológicas, mantidas por investimentos públicos ou de empresas privadas que, além de manter os equipamentos e coletar os dados, também podem fornecer produtos, como sistemas de alerta, para o produtor (GLEASON et al., 2008).

Os principais órgãos operacionais de meteorologia do Brasil que mantêm uma rede de observação em nível nacional são: Instituto Nacional de Meteorologia (INMET), do Ministério da Agricultura, Pecuária e Abastecimento; Departamento de Controle do Espaço Aéreo (DECEA) do Comando da Aeronáutica e a Diretoria de Hidrografia e Navegação (DHN) do Comando da Marinha, ambos do Ministério da Defesa, além do Instituto Nacional de Pesquisas Espaciais do Ministério da Ciência e Tecnologia (INPE) (INMET, 2012).

Outra fonte pública de dados meteorológicos é o Agritempo (Sistema de Monitoramento Agrometeorológico), um sistema que agrega informações de todas as outras fontes públicas de dados sejam estaduais ou estatais, e que permite aos usuários o acesso, via Internet, às informações meteorológicas e agrometeorológicas de diversos municípios e estados brasileiros (EVANGELISTA et al., 2003). O sistema permite a atualização de cadastro de estações e dados climáticos diários (temperaturas máxima e mínima, e precipitação), criação de boletins agrometeorológicos e visualização de mapas. Os dados são recebidos de várias instituições, em diferentes formatos, e passam por um processo de migração, incluindo a validação, antes de serem inseridos no banco de dados (www.agritempo.gov.br).

As estações agrometeorológicas de superfície cadastradas no Agritempo estão representadas na Figura 2.4.

Figura 2.4: Estações agrometeorológicas de superfície cadastradas no Agritempo. Fonte: www.agritempo.gov.br

Mesmo com uma aparente alta densidade de estações, há regiões com uma grande falta de estações meteorológicas. Uma alternativa para cobrir a falta de dados meteorológicos é a utilização de modelos climáticos ou meteorológicos, simulando os dados faltantes com base nos dados existentes. Existem diferentes métodos de se completar a informação faltante, e uma delas é a simulação por meio das chamadas estações virtuais (ROMANI et al., 2007).

As estações virtuais foram definidas pela simulação dos dados de precipitação e temperaturas máxima e mínima para uma determinada latitude e longitude. Romani et al. (2007) utilizaram um hidroestimador desenvolvido pelo CPTEC (Centro de Previsão e Estudos Climáticos) para simular os dados de precipitação. Já os dados de temperaturas máxima e mínima foram simuladas a partir de estações agrometeorológicas reais, numa vizinhança de até 100km, usando uma média ponderada com o inverso do quadrado da distância (ROMANI et al., 2003b).

Os principais objetivos da criação das estações virtuais foram de (i) simular dados para períodos acumulados (3, 5, 7, 10 dias e mensais), uma vez que os dados diários não apresentaram uma confiabilidade de 100%; e (ii) estimar dados faltantes de estações de superfície (ROMANI et al., 2007).

Atualmente, de acordo com Otavian (2011), o Agritempo substituiu os dados simulados desse hidroestimador pelos dados estimados provenientes do radar de precipitação a bordo do satélite TRMM - *Tropical Rainfall Measuring Mission* (http://trmm.gsfc.nasa.gov/).

O satélite TRMM foi desenvolvido em conjunto pela Agência Espacial Norte Americana (NASA) e pela Agência de Exploração Aeroespacial do Japão (JAXA) com o objetivo de realizar medidas de precipitação nas áreas tropicais e subtropicais (KUMMEROW et al., 2000). Seu lançamento foi realizado em 27 de novembro de 1997 no Japão, com uma expectativa de duração da missão de 3 a 3,5 anos, porém, continua ativo até o presente momento (TRMM, 2012).

Os dados pluviométricos provenientes do satélite TRMM foram estimados a partir de uma grade regular de 0,25º x 0,25º (aproximadamente 625 km^2) para todo o Brasil. Assim, de acordo com Otavian (2011), após o processamento de dados pelo Agritempo, cada ponto dessa grade se tornou referência para a criação de uma estação virtual, ou seja, cada ponto foi

caracterizado pela estimativa de precipitação pelo radar do TRMM e pela simulação de temperaturas máxima e mínima, como descrita por Romani et al. (2003b).

Em geral, os dados de superfície são considerados os "verdadeiros", mas são espaçados de forma irregular e representam apenas características pontuais e arredores. Pela baixa densidade de medições, a interpolação é utilizada, mas pode gerar resultados que deixam a desejar na qualidade, especialmente para medidas de precipitação. O uso do satélite TRMM é uma opção, mesmo que ainda algumas regiões apresentem dados superestimados, e outras, subestimados (ROZANTE et al., 2010). Outros estudos que apresentaram resultados satisfatórios para as estimativas do TRMM, como Nicholson (2005) para uma porção do oeste da África, e no Brasil com Collischonn et al. (2006) sobre a bacia do São Francisco até Três Marias e Collischonn et al. (2007) para a região da bacia do Alto Paraguai.

2.3 Mineração de dados

2.3.1 Conceitos de mineração de dados

Com o aumento do volume de dados nas mais diversas áreas de conhecimento humano, técnicas que auxiliem as pessoas a obterem informações úteis (conhecimento) a partir desses dados são cada vez mais necessárias. As técnicas de análise de dados tradicionais são baseadas em análises manuais, que podem ser demoradas, caras e subjetivas. Assim, novas técnicas e métodos foram e continuam sendo desenvolvidos na área conhecida como Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases – KDD*) a fim de transformar informações contidas em grandes quantidades de dados em formas mais simples e fáceis para interpretação humana (FAYYAD et al., 1996).

A definição mais aceita atualmente de KDD é: "Descoberta de Conhecimento em Bases de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em repositórios de dados" (FAYYAD et al., 1996).

Nesta definição, Fayyad et al. (1996) também explicam que o termo "dados" são um conjunto de fatos (por exemplo, registros de uma base de dados); "padrões" podem ser expressões descrevendo subconjuntos dos dados, ou podem ser modelos aplicáveis a

subconjuntos dos dados; “processo” indica que KDD compreende várias fases (por exemplo, preparação dos dados ou busca por padrões).

Os padrões descobertos devem ser “válidos” em novos dados, com algum grau de certeza, além de ser desejável que sejam “novos” e “potencialmente úteis”, de preferência para o usuário, podendo conduzir a algum benefício; por fim, os padrões devem ser “compreensíveis”, se não imediatamente, então, após algum pós-processamento (FAYYAD et al., 1996)

O termo “mineração de dados” (*data mining*, em inglês) é muitas vezes usado como sinônimo KDD. Porém, também pode ser considerada como uma das etapas do processo KDD (HAN et al., 2011), na qual seriam aplicados algoritmos e, após um período aceitável de tempo, padrões (modelos) seriam produzidos sobre os dados (FAYYAD et al., 1996).

A Figura 2.5 apresenta uma visão geral sobre um processo KDD (FAYYAD et al., 1996).

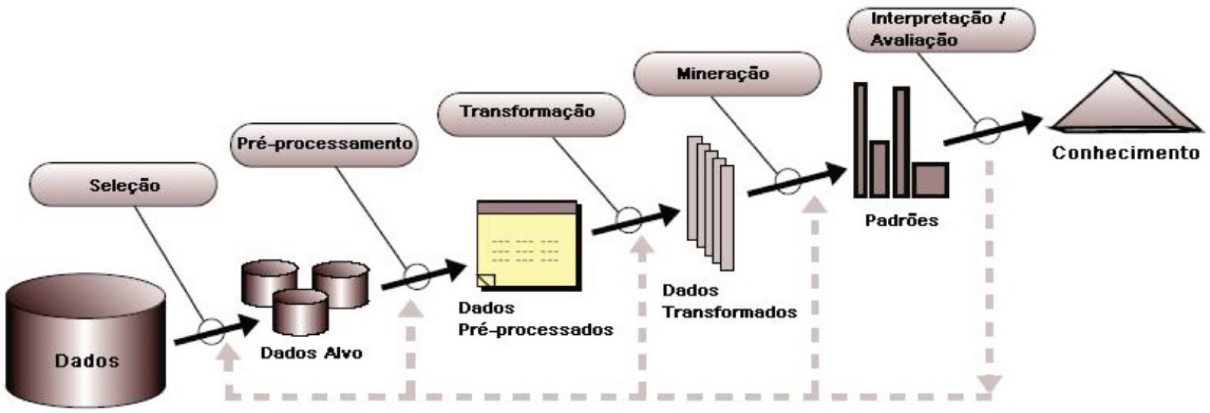


Figura 2.5: Processo KDD (Adaptado de Fayyad et al. (1996))

Primeiramente, é preciso compreender o domínio de aplicação, e identificar os objetivos pelo ponto de vista do usuário. Em seguida, sobre os dados selecionados é feito um pré-processamento, como por exemplo, eliminação de ruído e tratamento de dados faltantes, e então podem ser aplicadas transformações nos dados (por exemplo, conversão de dados e derivação de novos atributos), a fim de se produzir um conjunto de dados pronto para aplicar técnicas de mineração de dados (modelagem). Por fim, é feita a avaliação e interpretação dos resultados e o conhecimento descoberto pode ser aplicado e distribuído conforme o planejamento.

KDD se baseia em técnicas conhecidas de aprendizado de máquina, de reconhecimento de padrões e de estatística para encontrar padrões nos dados. As técnicas de visualização de dados estimulam naturalmente a percepção e a inteligência humana, aumentando a capacidade de entendimento e associação de novos padrões (REZENDE et al., 2002).

2.3.2 Tarefas e técnicas de mineração de dados

Na prática, os dois objetivos principais da mineração de dados são a predição e a descrição. A predição envolve o uso de variáveis com valores conhecidos para predizer um valor desconhecido ou futuro de outra variável (atributo meta). A descrição caracteriza propriedades gerais encontradas nos dados, com foco em padrões interpretáveis pelo ser humano. Esses objetivos podem ser alcançados por meio de vários tipos de tarefas. A escolha de uma ou mais tarefas depende do problema em questão. As tarefas tradicionais de mineração de dados estão representadas na Figura 2.6 e são brevemente descritas a seguir.

Classificação: é o processo de encontrar um conjunto de modelos (funções) que descrevam e distingam classes ou conceitos pré-definidos, com propósito de utilizá-los para predizer classes de objetos que ainda não foram classificados. O modelo construído baseia-se na análise prévia de um conjunto de dados de amostragem ou dados de treinamento, contendo objetos classificados corretamente.

Regressão: consiste em descobrir uma função que mapeie um item de dados para uma variável de predição de valor numérico contínuo. Conceitualmente é similar à classificação, porém, em vez de classes discretas, são números contínuos.

Associação: é a descoberta de regras ou padrões (como um conjunto de atributos ou sequência de eventos) que ocorrem frequentemente juntos em um conjunto de dados.

Agrupamento (*clustering*): consiste em agrupar os dados em classes ou *clusters*, tal que os elementos do mesmo grupo tenham alta similaridade entre si e sejam diferentes dos elementos das outras classes. Ao contrário da classificação e regressão, não há um atributo meta ou classes pré-definidas para compor um conjunto treinamento.

Figura 2.6: Tarefas de mineração de dados (REZENDE et al., 2002)

Sumarização: envolve meios de encontrar uma descrição compacta para um subconjunto de dados, como, por exemplo, média, desvio padrão ou técnicas de visualização.

Cada tarefa de mineração de dados possui técnicas diferentes associadas. Dentre as mais populares estão: árvores de decisão, regras de classificação, redes neurais, regressão linear ou não linear, análise de cesta de mercado ou k-vizinhos mais próximos.

Dentre as técnicas de mineração de dados, não é possível indicar a melhor delas, pois cada uma possui vantagens e desvantagens. A escolha de uma técnica requer uma análise mais detalhada do problema, principalmente levando em consideração o formato dos dados e como o conhecimento descoberto pode ser representado. Melhor ainda é poder aplicar mais de uma técnica para resolver o mesmo problema e no final apresentar os melhores resultados.

2.3.3 Classificação com árvores de decisão

Classificação e regressão são técnicas utilizadas para problemas de predição. Se o objetivo da predição é um valor discreto (alfanumérico), usa-se a classificação. Se o objetivo for numérico e contínuo, usa-se a regressão. Dentre os métodos de classificação, a indução de árvore de decisão contribui para a compreensão dos dados. A razão disto é que esta técnica

apresenta o conhecimento gerado de forma simbólica (estruturas compreensíveis), ou seja, interpretável por humanos. Neste sentido, os especialistas humanos podem verificar o conhecimento extraído e relacioná-lo à sua própria experiência (MONARD e BARANAUSKAS, 2002b).

Um modelo de árvore de decisão tem uma estrutura similar a de um fluxograma no formato de árvore (invertida), onde o primeiro nó (mais acima) é chamado de nó raiz. Nos nós internos, ou nós de decisão, são realizados testes sobre os valores dos atributos (variáveis independentes), gerando ramos, e cada um destes representa uma saída do teste. Os nós terminais, também conhecidos como nós folhas, contêm as classes (atributo meta) onde um exemplo é classificado (HAN et al., 2011).

Por ser uma tarefa de classificação, o objetivo é construir um modelo, geralmente para prever uma classe discreta e categórica (alfanumérica), tais como, “sim” ou “não”, ou também como “alto”, “médio” ou “baixo”. Para isso, o classificador precisa de um conjunto de treinamento, que é um conjunto de exemplos (também chamados de registros, instâncias, amostras ou objetos) com os atributos, ou variáveis independentes, classificados corretamente pelo atributo meta (classe), ou variável dependente, de onde o modelo pode “aprender” como classificar outros registros, no mesmo formato, chamado de conjunto de teste.

As árvores de decisão apresentam um potencial de interpretação único proporcionado pelos modelos simbólicos. Os resultados podem ser diretamente inspecionados a fim de se compreender as decisões do método e, consequentemente, os padrões existentes nos dados (APTÉ e WEISS, 1997). Um exemplo típico de uma árvore de decisão está representado na Figura 2.7, na qual tem-se um modelo com o objetivo de recomendar a saída para uma viagem ou não, baseando-se em dados relativos às condições do tempo. O atributo meta possui duas classes: “vá” ou “não_vá”.

O conhecimento representado em árvores de decisão pode ser extraído na forma de regras de classificação “se-então”. Uma regra é criada para cada caminho entre a raiz e um nó folha. Essas regras podem ser mais fáceis de compreender, especialmente se a árvore de decisão for muito grande (HAN et al., 2011).

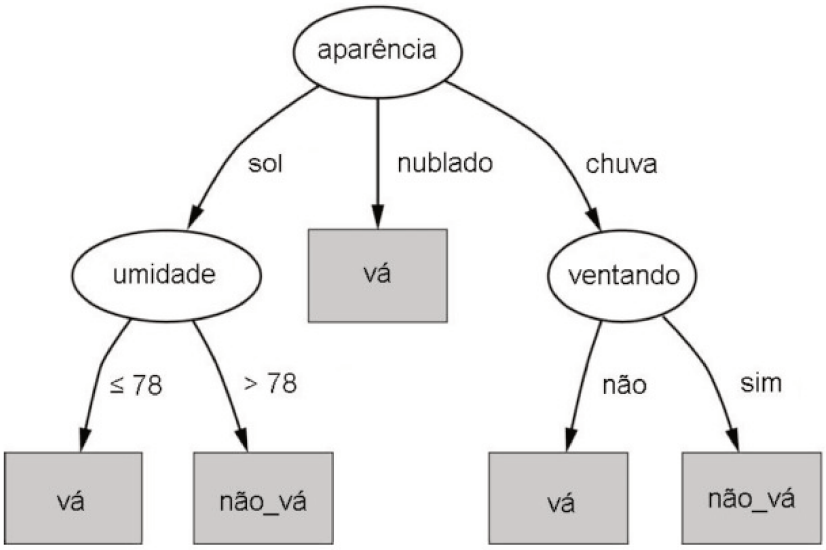


Figura 2.7: Exemplo de árvore de decisão (MONARD e BARANAUSKAS, 2002b)

No caso da Figura 2.7, as seguintes regras que podem ser extraídas:

SE aparência = sol E umidade ≤ 78 ENTÃO classe = vá

SE aparência = sol E umidade > 78 ENTÃO classe = não_vá

SE aparência = nublado ENTÃO classe = vá

SE aparência = chuva E ventando = não ENTÃO classe = vá

SE aparência = chuva E ventando = sim ENTÃO classe = não_vá

2.3.4 Indução de árvores de decisão

A indução é uma forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos (MONARD e BARANAUSKAS, 2002a). O propósito básico da indução de uma árvore de decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema. No último caso, a intenção é compreender quais variáveis e interações entre elas conduzem o fenômeno estudado.

O algoritmo básico de indução de árvores de decisão constrói a árvore recursivamente, de cima para baixo, segundo a abordagem conhecida como “dividir-e-conquistar” (HAN et al., 2011; WITTEN et al., 2011). O conjunto de treinamento é, repetidamente, dividido em subconjuntos a partir de testes sobre os atributos preditivos, onde

se busca os subconjuntos mais homogêneos, ou seja, que seja possível atribuir um único valor do atributo meta para cada subconjunto.

Para escolher quais atributos serão selecionados em cada repetição, os métodos mais comuns (HAN et al., 2011) são os baseados na teoria da informação, como o ganho de informação e a razão de ganho de informação (QUINLAN, 1993), e o índice Gini (BREIMAN et al., 1984).

O ganho de informação, por exemplo, é uma medida baseada na teoria da informação, que busca selecionar o melhor atributo considerando os atributos com menor "impureza". O quanto antes os subconjuntos se tornarem mais homogêneos em relação à classe, menos ramos (divisões) o modelo terá, simplificando a árvore gerada.

Outro conceito importante é sobre as suposições que o algoritmo deve fazer em relação aos dados faltantes. Na indução de árvores de decisão, se um teste é realizado sobre um atributo com valor faltante, há o problema de qual ramo aquela instância deve seguir. Uma das soluções existentes é a de fracionar as instâncias utilizando pesos numéricos entre 0 e 1 que indiquem a proporção do número de instâncias que seguiram o mesmo ramo e obtiveram a mesma classificação. Assim, para classificar esta instância, depois de induzida, as decisões em cada nó folha devem ser recombinadas utilizando os pesos de cada nó folha (WITTEN et al., 2011).

Durante a construção de uma árvore de decisão, muitos ramos podem representar os ruídos indesejados ou *outliers* (HAN et al., 2011), assim como a árvore pode ser muito específica para o conjunto de treinamento, causando o efeito conhecido como *overfitting*, ou, super-ajuste (MONARD e BARANAUSKAS, 2002b).

Para evitar esses problemas, utilizam-se técnicas de poda da árvore para reduzir o número de nós internos, gerando árvores menores e menos complexas, consequentemente, mais fáceis de serem compreendidas.

Existem as técnicas de pré-poda e de pós-poda. As técnicas de pré-poda buscam parar a construção da árvore mais cedo, decidindo não dividir mais os nós e transformando-os em folhas, mesmo que não classifique corretamente todas as instâncias. Isso pode ser feito utilizando diversos critérios de parada, dentre eles o que impede que a árvore atinja determinado nível de profundidade e o que impede que novos ramos se formem ao atingir um

determinado número mínimo de exemplos no conjunto que está sendo avaliado, criando-se um nó folha (WITTEN et al., 2011).

As técnicas de pós-poda atuam após a construção completa de uma árvore e realizam cortes em alguns nós e substituem alguns ramos por nós folha com a classe que apresentou maior frequência nos ramos excluídos.

2.3.5 Avaliação de modelos de classificação

A avaliação dos modelos de classificação consiste, basicamente, em testar o desempenho do modelo. Para isso, normalmente, se divide o conjunto de dados em duas partes disjuntas, sendo uma o conjunto de treinamento e outra o conjunto de teste, e cada uma destas deve ser representativa do conjunto de dados inicial.

Dentre os métodos de divisão dos dados mais utilizados estão o *holdout* e o de validação cruzada (*cross-validation*). O método *holdout* particiona aleatoriamente o conjunto de dados inicial em, tipicamente, dois terços para o conjunto de treinamento e um terço para o conjunto de teste. Já o método de validação cruzada separa aleatoriamente o conjunto de dados em k partições disjuntas *(folds),* sendo que uma delas para o conjunto de teste, e as outras, juntas, são utilizadas como conjunto de treinamento. Esse procedimento é realizado k vezes, cada vez utilizando um conjunto de teste diferente, e no final, a estimativa do erro é uma média de todos os erros obtidos.

Vale ressaltar que a separação dos dados, geralmente, é realizada de maneira estratificada, ou seja, mantendo a proporção das classes de exemplos existentes no conjunto original nos conjuntos de treinamento e de teste. Estudos apontam que a divisão do conjunto de dados em dez partes é uma das que realizam melhor a estimativa do erro, prática que ficou conhecida como *10-fold cross-validation* (WITTEN et al., 2011).

Um dos resultados de um classificador, após realizar o treinamento e o teste do modelo, é a matriz de confusão, que permite identificar e medir os acertos e erros do modelo. A Tabela 2.1 ilustra a matriz de confusão para um problema de duas classes, denominadas C_+ (classe positiva) e C_- (classe negativa). Nesse caso, existem quatro possíveis resultados em relação à classificação de exemplos (MONARD e BARANAUSKAS, 2002a). São eles:

•Verdadeiros positivos (VP), quando os exemplos pertencem à classe C_+ e foram preditos como pertencentes à essa mesma classe.

•Falsos negativos (FN), quando os exemplos pertencem à classe C_+ e foram preditos como pertencentes à classe C_-.

•Verdadeiros negativos (VN), quando os exemplos pertencem à classe C_- e foram preditos como pertencentes à essa mesma classe.

•Falsos positivos (FP), quando os exemplos pertencem à classe C_- e foram preditos como pertencentes à classe C_+.

Tabela 2.1: Matriz de confusão para classificação com duas classes.

		Predita		
		C_+	C_-	Total
Verdadeira	C_+	VP	FN	P
	C_-	FP	VN	N
Total		P'	N'	P + N

Na coluna *Total* da Tabela 2.1, *P* é o valor total de casos positivos e *N* é o total de casos negativos existentes no conjunto de treinamento. Já na linha *Total*, *P′* é o total de casos que o modelo classificou como casos positivos e *N′* é o total de casos classificados como negativos.

A partir da matriz de confusão, é possível extrair as métricas de avaliação de desempenho. A taxa de acerto, ou acurácia (*accuracy*), é a porcentagem de exemplos que foram classificados corretamente pelo classificador e pode ser expressa como na Equação 2.

$$taxa\ de\ acerto = \frac{VP + VN}{P + N} \tag{2}$$

Uma medida similar é a **taxa de erro** (*error rate*), que é 1 – **taxa de acerto**, também expressa pela Equação 3.

$$taxa\ de\ erro = \frac{FP + FN}{P + N} \tag{3}$$

Outras medidas podem ser derivadas da matriz de confusão, tais como **sensitividade**, ou precisão da classe C_+ (*sensitivity*) e a **especificidade**, ou precisão da classe C_- (*specificity*), calculadas a partir das Equações 4 e 5 respectivamente.

$$sensitividade = \frac{VP}{P} \tag{4}$$

$$especificidade = \frac{VN}{N} \tag{5}$$

Os bons modelos apresentam altos valores na diagonal principal da matriz (VP e VN) e baixos valores na outra diagonal (FP e FN).

A estatística *Kappa* também pode ser utilizada para medir o desempenho do classificador (COHEN, 1960). O kappa é uma medida de concordância entre as classes preditas e observadas, que deduz o número esperado de acertos (utilizando uma classificação ao acaso) do número real de acertos do classificador (WITTEN et al., 2011).

O valor máximo desta medida de concordância é 1, onde este valor representa total concordância e os valores próximos a 0, indicam nenhuma concordância, ou a concordância foi exatamente a esperada pelo acaso.

A Equação 6 é utilizada para calcular o Kappa, a partir da matriz de confusão da Tabela 2.1 é:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

onde p_o é equivalente à proporção de acertos nas classes (Equação 7) e p_e é a probabilidade de concordância esperada considerando eventos independentes (Equação 8).

$$p_o = \frac{VP + VN}{P + N} \tag{7}$$

$$p_e = \frac{(P\,'\,P) + (N\,'\,N)}{(P + N)^2} \tag{8}$$

Uma possível interpretação do desempenho dos modelos a partir da estatística Kappa foi introduzida por Landis e Koch (1977) e é expressa na Tabela 2.2:

Tabela 2.2: Desempenho de modelos de classificação a partir do Kappa.

Estatística kappa	Qualidade
< 0,00	Péssima
0,00 – 0,20	Ruim
0,21 – 0,40	Razoável
0,41 – 0,60	Boa
0,61 – 0,80	Muito Boa
0,81 – 1,00	Excelente

2.4 Modelo do processo de descoberta de conhecimento em bases de dados

Com o objetivo de padronizar o processo de descoberta de conhecimento, em 1996 foi criado o modelo de processo CRISP-DM (*CRoss-Industry Standart Process for Data Mining*), que divide o ciclo de vida de um projeto de mineração de dados em seis fases, a saber: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição (CHAPMAN et al., 2000).

As fases do modelo do processo estão ilustradas na Figura 2.8. O círculo externo traduz o aspecto cíclico de um projeto de mineração de dados, uma vez que após encontrar uma solução, o projeto não é necessariamente finalizado, e a partir de novos conhecimentos adquiridos podem ocorrer novos questionamentos que levam a ações mais específicas. As setas internas ilustram as relações entre as fases, que indicam que a sequência entre elas não é rígida, sendo comum haver a necessidade de voltar ou avançar entre as fases.

A seguir, encontra-se uma breve descrição de cada fase do processo:

•**Compreensão do domínio:** compreender os objetivos e requisitos do projeto e transformar esse conhecimento em um problema de mineração de dados. Definir um plano preliminar para atingir esses objetivos.

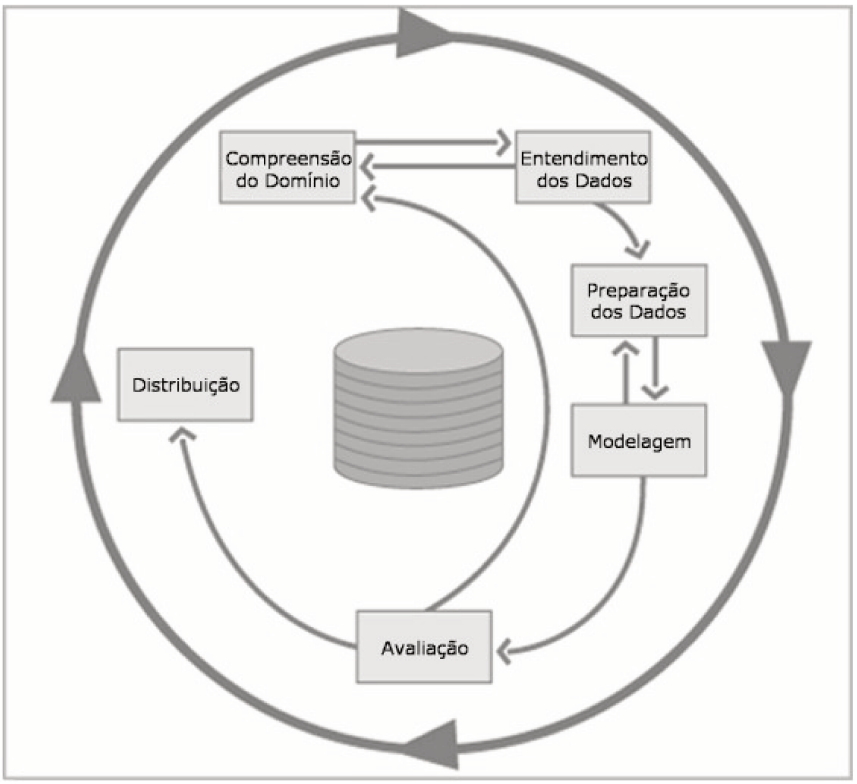


Figura 2.8: Fases do modelo de processo CRISP-DM (Adaptado de Chapman et al. (2000))

•**Entendimento dos dados:** inicia-se com uma coleção de dados inicial e prossegue com atividades de exploração de dados, para se familiarizar, identificar problemas de qualidade, fazer as primeiras hipóteses e identificar possíveis subconjuntos que possam abrigar informações ocultas sobre esses dados.

•**Preparação dos dados:** a fase de preparação dos dados contempla todas as atividades necessárias para a construção do conjunto de dados final, no qual serão aplicadas as técnicas de modelagem. As atividades incluem, por exemplo, limpeza de dados, seleção e transformação de atributos, entre outras.

•**Modelagem:** nessa fase são escolhidas e aplicadas as técnicas de mineração de dados, e seus parâmetros são calibrados. Diversas técnicas podem ser aplicadas ao mesmo problema, embora cada técnica necessite de formatos específicos e necessite voltar para a fase de preparação de dados.

•**Avaliação:** nesse estágio, tem-se o modelo (ou modelos) com boa qualidade. Os resultados são comparados e interpretados conforme a área de aplicação. É importante reavaliar todas as etapas do processo para se ter a certeza de que o modelo atende às necessidades e aos objetivos do projeto.

•**Distribuição:** a criação de modelos geralmente não finaliza um projeto. O conhecimento obtido deve ser documentado, organizado e apresentado para os usuários, para que estes possam saber quais ações devem ser realizadas para aproveitar os modelos criados.

3 MATERIAL E MÉTODOS

3.1 Considerações iniciais

Neste trabalho foi utilizada uma base de dados na ordem de milhares de ocorrências da ferrugem asiática da soja e os respectivos dados meteorológicos para cada caso de ocorrência. Além da coleta dos dados, foram realizadas tarefas de exploração e limpeza dos dados com o objetivo de minimizar os ruídos e aumentar o potencial de descoberta de conhecimento nos dados.

A influência de variáveis meteorológicas nos eventos de ocorrência da doença foi feita considerando variáveis de temperatura e precipitação em períodos ou janelas temporais anteriores aos eventos de ocorrência e não ocorrência da doença.

Para distinguir os fatores que favorecem a ocorrência da ferrugem dos fatores que não favorecem, desenvolveu-se um procedimento para simular um evento de não ocorrência da doença. Para cada registro de ocorrência foi criado um evento de não ocorrência, com a premissa de que em um momento anterior, no caso 30 dias antes da ocorrência, a doença não havia sido detectada .

A técnica de modelagem escolhida foi a indução de árvores de decisão para a classificação. Tal técnica foi aplicada utilizando o classificador J48, versão adaptada do algoritmo C4.5 de Quinlan (1993), embutido no software livre Weka (HALL et al., 2009), ambiente que contém um conjunto de algoritmos de aprendizado de máquina para solucionar problemas de mineração de dados.

Para dar suporte aos procedimentos realizados neste trabalho, optou-se por seguir um modelo de processo descrito por Chapman et al. (2000) para projetos de descoberta de conhecimento em bases de dados, conhecido como CRISP-DM (*CRoss-Industry Standart Process for Data Mining*). A aplicação do modelo CRISP-DM neste trabalho, com as atividades desenvolvidas nas respectivas fases, pode ser acompanhada nas seções seguintes.

3.2 Compreensão do domínio

Fase que resultou na elaboração do Capítulo 2, e que foi desenvolvida por meio de revisão na literatura e reuniões com especialistas das áreas de fitopatologia, soja, banco de dados e mineração de dados, com o objetivo de se conhecer melhor as questões, propriedades e restrições que cercam o assunto da soja, da ferrugem asiática e dos dados.

3.3 Entendimento dos dados

3.3.1 Coleção inicial dos dados e descrição

Dados de ocorrências da ferrugem asiática da soja

Por meio de consultas ao banco de dados do CAF foi possível selecionar quais atributos poderiam descrever um evento de ocorrência da ferrugem. Os registros foram organizados em um único arquivo no formato de tabela tipo ".csv" (*comma separated values* – valores separados por vírgulas), no qual cada linha representa um registro de ocorrência e cada coluna representa um atributo relacionado à ocorrência. Os atributos selecionados estão listados na Tabela 3.1.

O conjunto de dados brutos proveniente do CAF, entre os anos 2005 e 2011, continha 12.554 registros de ocorrências da doença caracterizados por 12 atributos (Tabela 3.1), considerando todos os registros de ocorrência, desde a criação do banco de dados até o momento da coleta de dados para este trabalho.

Para compreender melhor o atributo "estadio", a descrição dos estádios de desenvolvimento da soja está baseada em Fehr e Caviness (1977), como mostra a Tabela 3.2.

Dados meteorológicos

Para extrair os dados do banco de dados do Agritempo, foi utilizada uma lista com os estados do Brasil que tiveram pelo menos um registro de ocorrência da ferrugem asiática da soja. Para estes estados, todas as estações agrometeorológicas disponíveis foram separadas em diretórios específicos, um para estações reais e outro para estações virtuais.

Tabela 3.1: Descrição dos dados disponíveis para cada relato de ocorrência de ferrugem asiática nos municípios do Brasil e disponibilizados pelo Consórcio Antiferrugem.

Atributo	Descrição	Exemplo
dataOcorrencia	Data: DD/MM/AAAA.	21/01/2010
safra	Ano (AAAA) ou Ano/Ano (AAAA/AAAA).	2009/2010
mesPlantio	Mês de plantio: MM.	11
quinzenaPlantio	Quinzena do mês de plantio: Numérico (1 ou 2).	1
IBGEmunicipio	Numérico: 7 dígitos de identificação para o município, definidos pelo IBGE (Instituto Brasileiro de Geografia e Estatística).	4314100
municipio	Nome por extenso do município.	Passo Fundo
estado	Sigla com dois caracteres do estado (unidade federativa do Brasil).	RS
latitude	Latitude global do município.	-28.263
longitude	Longitude global do município.	-52.407
cultivar	Tipo de cultivar na qual foi identificada a ocorrência.	BMX Apollo
estadio	Estádio de desenvolvimento da planta no momento do registro da ocorrência (Tabela 3.2).	R5
tipoDeArea	Tipo de área da plantação (“COMERCIAL” ou “OUTRAS”).	COMERCIAL

Tabela 3.2: Codificação e descrição dos estádios fenológicos de desenvolvimento da soja.

Estádio	Descrição
VC	Da emergência a cotilédones abertos.
V1	Primeiro nó; folhas unifolioladas abertas.
V2	Segundo nó; primeiro trifólio aberto.
V3	Terceiro nó; segundo trifólio aberto.
Vn	Enésimo (último) nó com trifólio aberto
R1	Início da floração: até 50% das plantas com flor.
R2	Floração plena: maioria dos racemos com flores abertas.
R3	Final da floração: flores e vagens com até 1
R4	Maioria das vagens no terço superior com 2-4cm.
R5	Formação e enchimento dos grãos.
R6	Vagens com granação de 100% e folhas verdes.
R7	Amarelecimento de folhas e vagens.
R8	Desfolha.
R9	Ponto de maturação de colheita.

Para fins de organização e documentação, os arquivos com as estações reais e virtuais foram nomeados, respectivamente, como segue:

•Nome do arquivo: “2901205.CPTEC.0006.csv”, onde “2901205” representa o código do município e “CPTEC.0006” representa a identificação da estação.

•“2900207-9.000_-39.500.csv”, onde “2900207” representa o código do município e “-9.000_-39.500” representa a identificação da estação, no caso, a latitude e longitude, respectivamente.

O conteúdo dos arquivos de dados meteorológicos está descrito na Tabela 3.3.

Tabela 3.3: Descrição dos dados meteorológicos brutos obtidos a partir do Agritempo.

Atributo	**Descrição**
data	Data: DD/MM/AAAA.
tmin	Temperatura mínima do dia, em graus Celsius (ºC).
tmax	Temperatura máxima do dia, em graus Celsius (ºC).
precipitacao	Precipitação acumulada do dia, em milímetros (mm).

Como os dados do Consórcio Antiferrugem estavam separados por município, e muitas vezes esses dados apresentavam mais de uma estação meteorológica para o mesmo município, foi desenvolvido um *script* na linguagem de programação *Python* que, basicamente, identificou as estações do mesmo município e calculou a média aritmética entre os valores das variáveis (“tmin”, “tmax” e “precipitacao”) para todos os dias disponíveis.

Os novos arquivos foram nomeados com o código do município (p. ex., “1702208.txt”) tanto para as estações meteorológicas reais quanto para as virtuais, ressaltando-se que foram separadas em diretórios diferentes.

3.3.2 Exploração dos dados

Na fase de exploração dos dados buscou-se uma visão geral dos dados, incluindo qualidade, distribuição, semelhanças e diferenças entre os principais subconjuntos dos dados. Foi utilizado principalmente o *software* R para a visualização de tabelas e gráficos.

Para os dados do Consórcio Antiferrugem, buscou-se compreender a distribuição no tempo e no espaço das ocorrências da doença, com o objetivo principal de selecionar possíveis subconjuntos que pudessem abrigar algum conhecimento oculto.

O conjunto de dados original, no formato tabela (“.csv”), era composto por 12 atributos (descritos na Tabela 2.8) e 12.554 registros de ocorrências da ferrugem asiática da soja. As primeiras ações tiveram como objetivo analisar a distribuição dos registros em relação a cada atributo e, posteriormente, combinados.

Em relação ao atributo “tipoDeArea”, contendo as categorias “COMERCIAL” e “OUTRAS”, este apresentou uma distribuição discrepante entre as duas categorias, uma vez que apenas 617 registros (menos de 5% do total) possuíam o valor “OUTRAS”. Assim, a fim de minimizar o ruído nos dados, foram utilizados apenas os registros da categoria “COMERCIAL”.

Os atributos “mesPlantio”, “quinzenaPlantio” e “cultivar” também foram desconsiderados, pois não apresentavam potencial de abrigar algum conhecimento oculto, devido, principalmente, às distribuições irregulares e dados faltantes.

De acordo com as análises em relação aos outros atributos do Consórcio Antiferrugem, foi decidido não utilizá-los no conjunto de dados final para a fase de modelagem por diversos motivos. Considerando a Tabela 3.1, os atributos “dataOcorrencia”, “safra”, “IBGEmunicipio”, “municipio”, “estado”, “latitude” e “longitude” foram desconsiderados para que os modelos não estivessem restritos a uma localização ou a um período específico.

Assim, o único atributo restante do conjunto de dados do CAF seria o “estadio”, referente ao estádio de desenvolvimento da planta. Porém, devido à construção do atributo meta, explicada na seção seguinte, o atributo “estadio” ficaria descaracterizado nos registros de não ocorrência pelo fato de não haver uma forma bem definida de deslocar o estádio no tempo. Por exemplo, para uma planta no estádio R5, não se saberia dizer em qual estádio ela estaria 30 dias antes. Assim, a opção foi desconsiderar este atributo no conjunto de dados final.

Os dados meteorológicos foram obtidos por meio de consultas ao banco de dados do Agritempo. Foram utilizados os dados de temperaturas mínima e máxima e precipitação, provenientes de duas fontes: (i) estações meteorológicas de superfície, e (ii) estações

meteorológicas virtuais, que são pontos fictícios em latitude e longitude específicas para os quais são simulados ou estimados os valores das variáveis meteorológicas.

Ao avaliar os dados de estações de superfície disponíveis para os municípios do conjunto de dados do CAF, foi verificado que, de um total de 852 municípios com registros de ocorrências, existiam apenas 194 com dados meteorológicos, o que, praticamente, impossibilitou o uso de estações agrometeorológicas reais.

A alternativa encontrada foi utilizar os dados de estações virtuais. A utilização e validação desses dados já foram verificadas por Romani et al. (2003b; 2007), e especificamente, os dados pluviométricos provenientes do radar do satélite TRMM, foram validados por outros estudos (KUMMEROW et al., 2000; COLLISCHONN et al., 2007; ROZANTE et al., 2010).

Vale ressaltar que os dados das estações virtuais utilizando o TRMM tiveram início em setembro de 2007, que coincidiu com o início dos dados disponíveis para a safra 2007/2008. Assim, os dados de registros de ocorrência da doença anteriores a setembro de 2007 foram descartados, reduzindo o conjunto de dados de 12.554 para 7.810 registros.

Considerando 4 safras, de 2007/2008 a 2010/2011, no tipo de área comercial, foi obtido um total de 582 municípios. Para validar o uso de dados de estações virtuais, além da literatura consultada, foi realizada uma comparação entre os dados de estações reais e virtuais por meio do coeficiente de correlação de Pearson. Como esse estudo só poderia ser realizado entre municípios que continham dados de estações reais e virtuais, os resultados obtidos foram para um máximo de 134 municípios.

Para cada município foi realizado o teste de correlação entre dados acumulados de temperaturas mínima, máxima e precipitação de 5 dias, quantidade mínima de dias a ser utilizado no conjunto de dados final. O método de comparação de dados acumulados por período de comparação de dados já foi utilizado por Romani et al. (2003b; 2007).

Alguns municípios apresentaram muitos dados faltantes e não foi possível realizar o estudo para todos os 134. Para cada variável (temperatura mínima, máxima e precipitação) foi avaliado o número máximo de municípios com dados disponíveis.

Para relacionar os dados meteorológicos às ocorrências de ferrugem asiática da soja, os dados foram preparados segundo a descrição da seção a seguir.

3.4 Preparação dos dados

Após as fases de compreensão do domínio e entendimento dos dados, foram obtidos subsídios para prosseguir com a fase de preparação dos dados. Os objetivos da preparação de dados foram:

•Desenvolver um método para distinguir fatores favoráveis e desfavoráveis à ocorrência da ferrugem asiática da soja.

•Desenvolver um método para utilizar os dados meteorológicos de estações virtuais e relacioná-los com as ocorrências da doença.

•Integrar os dados em um formato específico, necessário para aplicar a técnica de indução de árvores de decisão para classificação utilizando o classificador J48, embutido no software Weka.

Para distinguir os fatores que favorecem a ocorrência daqueles que desfavorecem a ocorrência da ferrugem, o ideal seria ter registros de ambos os casos. Porém, o banco de dados do Consórcio Antiferrugem traz apenas as informações sobre eventos de ocorrências.

Para resolver esta questão, foi desenvolvido um procedimento que cria um registro de não ocorrência da doença com base em um registro de ocorrência, para o mesmo local. Ou seja, para cada registro de ocorrência, um novo registro equivalente foi criado, mas referente a um evento de não ocorrência da ferrugem.

Para definir o período que separaria o evento de ocorrência e o evento de não ocorrência da doença, foram levados em consideração as seguintes premissas:

•O inóculo do fungo estava presente na plantação no evento de não ocorrência.

•A infecção da planta é influenciada pelas condições ambientais.

•Existe um período entre a infecção da planta e a detecção da doença.

A presença do inóculo, a princípio, não foi considerado um fator limitante para eventos de ocorrência ou de não ocorrência, uma vez que para alguns modelos de previsão de risco considera-se que o fungo esteja presente na plantação e busca-se indicar o risco da doença se estabelecer ou atingir níveis preocupantes (DEL PONTE et al., 2006a).

Assim, fez-se a suposição que as condições ambientais precedentes à data da possível infecção dos registros de não ocorrência seriam desfavoráveis à infecção, enquanto que as

condições ambientais precedentes à data da possível infecção dos registros de ocorrências seriam favoráveis à infecção.

Para definir o período entre o evento de ocorrência e não ocorrência da doença, foi levado em consideração que houve um período anterior, no qual a planta estava infectada mas não apresentava sintomas. Estudos apontam que o período entre a infecção e o aparecimento de sintomas pode levar de 5 a 9 dias (MARCHETTI et al., 1975, 1976; GOELLNER et al.,2010; ZAMBENEDETTI et al., 2007).

Por serem dados de levantamento de ocorrências da doença, não há informações precisas sobre a severidade da doença ou há quanto tempo exatamente ela está presente na propriedade. Nessas condições, especialistas sugeriram que o tempo adequado buscando-se garantir que não haveria planta infectada no campo a partir da data que foi detectada é de 30 dias anteriores à data de registro de ocorrência.

Além de supor um período para garantir um evento de não ocorrência, também foi necessário estimar a possível data de infecção, a fim de relacionar a infecção com os atributos meteorológicos. Assim, para estimar a possível data da infecção, que pode variar de acordo com a temperatura, foi utilizado o conceito de período latente definido por Alves et al. (2006), que utiliza a Equação 1.

Após estabelecer como seriam os eventos de ocorrência e não ocorrência da doença, fez-se necessário definir quais seriam os atributos ambientais utilizados para caracterizar esses eventos. Práticas comuns na agricultura são combinar, acumular ou fazer a média de medidas meteorológicas em períodos específicos que precedem o evento que se está estudando, a fim de relacioná-los com o evento. Assim, foram desenvolvidos atributos relacionados à temperatura e precipitação em períodos anteriores à ocorrência e não ocorrência da doença. Maiores detalhes sobre a criação dos atributos meteorológicos preditivos podem ser encontrados na Seção 3.4.2.

3.4.1 Especificação do atributo meta

O atributo meta, também chamado de classe ou variável dependente, é um atributo especial para problemas de classificação, e deve ser categórico (não numérico). Ele foi criado para evidenciar eventos de ocorrência e não ocorrência da ferrugem asiática da soja. Para isso,

o atributo meta (classe) foi denominado “classeOcNaoOc”, podendo conter duas categorias: (i) “Oc”, para eventos de ocorrência e (ii) “NaoOc”, para eventos de não ocorrência.

Para criar os registros de não ocorrência e o atributo meta a partir do conjunto de dados do Consórcio Antiferrugem, foi desenvolvido um *script* em *Python*, denominado “01-criaNaoOcorrenciaFinal.py”, que, para cada registro de ocorrência, adicionava um registro de não ocorrência em uma data anterior mas no mesmo local; procedimento chamado de deslocamento de classe, ilustrado na Figura 3.1.

O registro criado de não ocorrência foi composto pelos mesmos valores dos atributos (os da Tabela 3.1) do registro de ocorrência, exceto os atributos temporais, como “dataOcorrencia” e “estadio”. A “dataOcorrencia” foi deslocada para trás em 30 dias e ao estádio de desenvolvimento da planta foi atribuído valor faltante (o caractere “?” representou o valor faltante), uma vez que o estádio não pode ser estimado devido à falta de um procedimento consolidado para estimar períodos de duração de cada estádio (FEHR e CAVINESS, 1977).

A partir desse procedimento, foi criado um novo arquivo contendo os registros de ocorrências e não ocorrências (atributo meta “classeOcNaoOc”) com os respectivos atributos da Tabela 3.1. Um exemplo do conjunto de dados após a criação dos registros de não ocorrência e do atributo meta "classeOcNaoOc" pode ser encontrado na Tabela 3.4.

Após a definição do atributo meta e de como os registros estariam organizados, seguiu-se com a atividade de especificar os atributos meteorológicos.

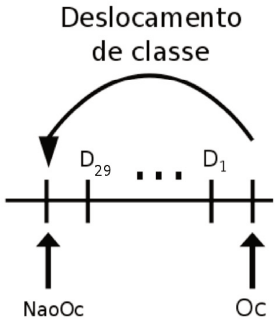


Figura 3.1: Deslocamento de classe na criação de registros de não ocorrência. Ao atributo meta “classeOcNaoOc” foi atribuído o valor “Oc” em casos de registros de ocorrência e o valor “NaoOc” para os registros de não ocorrência. O deslocamento de classe foi considerado como um período de 30 dias. D_1 é o dia anterior ao registro de ocorrência e D_{29} é o 29º dia antes do registro de ocorrência.

Tabela 3.4: Exemplo de instâncias do conjunto de dados após a criação dos registros de não ocorrência e do atributo meta "classeOcNaoOc".

dataOcorrencia	safra	mes Plantio	quinzena Plantio	IBGE municipio	municipio	estado	latitude	longitude	cultivar	estadio	tipoDeArea	classe OcNaoOc
10/12/2009	2009/2010	10	1	5100607	Alto Taquari	MT	-17.826	-53.28	TMG 123 RR	R3	COMERCIAL	Oc
10/11/2009	2009/2010	10	1	5100607	Alto Taquari	MT	-17.826	-53.28	TMG 123 RR	?	COMERCIAL	NaoOc
10/12/2009	2009/2010	10	1	5104609	Itiquira	MT	-17.209	-54.15	Msoy 7908 RR	R4	COMERCIAL	Oc
10/11/2009	2009/2010	10	1	5104609	Itiquira	MT	-17.209	-54.15	Msoy 7908 RR	?	COMERCIAL	NaoOc

3.4.2 Especificação dos atributos preditivos meteorológicos

Os atributos preditivos (variáveis independentes) são aqueles que pretendem ser utilizados para explicar o atributo meta (varável dependente). Como o objetivo geral deste trabalho é identificar os fatores, principalmente meteorológicos, que influenciam na ocorrência da ferrugem asiática da soja no campo, vale ressaltar que o mais interessante seria analisar características que precederam a infecção da planta. Assim, as variáveis meteorológicas deveriam ser analisadas em períodos precedentes à suposta data de infecção planta.

A partir disso, foram escolhidos os seguintes períodos: período latente estimado, 5, 10, 15 e 20 dias antes da possível infecção, como ilustrado na Figura 3.2. Também foi tomado o cuidado para que o período analisado não viesse a ultrapassar o suposto evento de não ocorrência. Assim, o período máximo considerado foi de aproximadamente 30 dias, somando-se o período latente (cerca de 8 a 10 dias) com o período de 20 dias antes da possível data de infecção.

O período latente foi calculado baseado na Equação 1, porém, foi realizada uma adaptação para uma forma dinâmica, ilustrada na Figura 3.3. A cada dia considerado, a nova temperatura era considerada no cálculo, gerando um novo *PL*. A partir do momento que o número de dias considerados se igualasse ao número de dias do período latente, esse seria o período latente considerado.

Após estabelecer os períodos que seriam utilizados, também deveriam ser definidas as características ambientais que seriam utilizadas para a modelagem. Essas características deveriam estar relacionadas com as propriedades da soja e da ferrugem asiática.

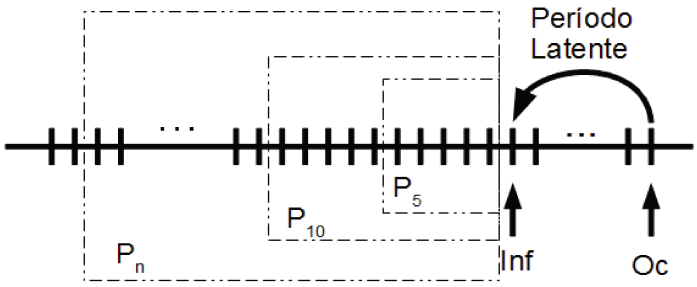


Figura 3.2: Períodos utilizados para compor os atributos preditivos meteorológicos. *Oc* – Dia da ocorrência; *Período Latente* – calculado a partir do algoritmo da Figura 3.3 ; *Inf* – data da possível infecção; P_5 – período de 5 dias que antecedem a data da infecção; P_{10} – período de 10 dias que antecedem a data da infecção; P_n, generaliza o procedimento aplicado para P_5 e P_{10} para os outros períodos de 15 e 20 dias.

A partir dos estudos da revisão bibliográfica, foram definidas que as variáveis meteorológicas seriam as temperaturas mínima, média e máxima, a precipitação média e acumulada, bem como o número de dias chuvosos (precipitação acima de 1 mm) e a maior chuva (em mm) de cada período. Também foram sugeridas variáveis como número de dias com temperaturas críticas para a não ocorrência da ferrugem, representadas por temperaturas abaixo de 15ºC e acima de 30ºC.

Enquanto n<= PL{

$$0{,}11\left(\sum_{i=1}^{n}\frac{T_i}{n}\right)^2-5{,}20\left(\sum_{i=1}^{n}\frac{T_i}{n}\right)+69{,}53=PL$$

n=n+1

}

Figura 3.3*:* Algoritmo adaptado de Alves et al. (2006). T_i é a temperatura média no i-ésimo dia anterior ao dia da ocorrência, *n* é o número de dias anteriores à ocorrência e *PL* é o número de dias do período latente.

Os atributos meteorológicos que foram utilizados para o conjunto de dados final na fase de modelagem estão listados na Tabela 3.5.

Tabela 3.5: Descrição dos atributos derivados das temperaturas mínima, máxima e da precipitação avaliados em cada um dos cinco períodos, sendo período latente, 5, 10, 15 e 20 dias antes da possível data de infecção.

Atributo	Descrição
tmin_media	Média da temperatura mínima no período.
dias_tmin_menor15	Número de dias nos quais a temperatura mínima foi abaixo de 15ºC.
tmax_media	Média da temperatura máxima no período.
dias_tmax_maior30	Número de dias nos quais a temperatura máxima ultrapassou 30ºC.
tmed_media	Média da temperatura média no período.
prcpt_media	Média da precipitação (em mm) no período.
prcpt_acum_mm	Precipitação (em mm) acumulada no período.
dias_prcpt_maior1mm	Número de dias nos quais a precipitação ultrapassou 1mm.
prcpt_maior	A maior precipitação (em mm) no período.

Cada atributo meteorológico foi avaliado nos períodos definidos (ilustrados na Figura 3.2). Para identificar o período ao qual o atributo está se referindo, foi acrescentado o número de dias do período ao título do atributo. Considerando-se os nove atributos da Tabela 3.5 para cada um dos cinco períodos em estudo, tem-se um total de 45 atributos meteorológicos para cada registro, ou seja, para cada caso de ocorrência e de não ocorrência. A Tabela 3.6 lista todos os atributos meteorológicos criados durante a fase de preparação dos dados.

Após a definição dos períodos e dos atributos que seriam criados e avaliados no conjunto de dados final, foi desenvolvido um *script* em *Python*, denominado "02-integraLatenteFinal.py", responsável pela criação dos atributos meteorológicos, assim como o cálculo dos valores nos respectivos períodos e a integração de todos esses dados em um único arquivo (".csv") que foi utilizado para a fase de modelagem.

O conjunto de dados final foi composto pelos atributos meteorológicos de cada período, listados na Tabela 3.6 e o atributo meta "classeOcNaoOc", classificando o registro em "Oc" para casos de ocorrência e "NaoOc", para casos de não ocorrência.

Embora o uso de dados meteorológicos provenientes de estações virtuais fornecesse dados para um número maior de municípios, nem todos os municípios do banco de dados do CAF apresentavam dados meteorológicos, uma vez que os pontos (resolução) de obtenção de dados do satélite TRMM, eventualmente, podem não abranger algum município. Assim,

fizeram parte do conjunto de dados final apenas os registros com dados meteorológicos de estações virtuais.

Outras atividades como remoção de *outliers* e dados inconsistentes que, a princípio, são classificadas como relevantes na fase de preparação de dados, estão descritas na seção seguinte (modelagem), pois, logo que realizadas essas atividades, a técnica de modelagem foi aplicada diversas vezes com parâmetros diferentes, a fim de avaliar o desempenho dos modelos gerados.

Tabela 3.6: Atributos meteorológicos agregados a cada registro, seja de ocorrência ou não ocorrência, do conjunto de dados final.

Atributos meteorológicos		
tmin_media_Latente	prcpt_acum_mm_5dias	dias_tmax_maior30_15dias
dias_tmin_menor15_Latente	dias_prcpt_maior1mm_5dias	tmed_media_15dias
tmax_media_Latente	prcpt_maior_5dias	prcpt_media_15dias
dias_tmax_maior30_Latente	tmin_media_10dias	prcpt_acum_mm_15dias
tmed_media_Latente	dias_tmin_menor15_10dias	dias_prcpt_maior1mm_15dias
prcpt_media_Latente	tmax_media_10dias	prcpt_maior_15dias
prcpt_acum_mm_Latente	dias_tmax_maior30_10dias	tmin_media_20dias
dias_prcpt_maior1mm_Latente	tmed_media_10dias	dias_tmin_menor15_20dias
prcpt_maior_Latente	prcpt_media_10dias	tmax_media_20dias
tmin_media_5dias	prcpt_acum_mm_10dias	dias_tmax_maior30_20dias
dias_tmin_menor15_5dias	dias_prcpt_maior1mm_10dias	tmed_media_20dias
tmax_media_5dias	prcpt_maior_10dias	prcpt_media_20dias
dias_tmax_maior30_5dias	tmin_media_15dias	prcpt_acum_mm_20dias
tmed_media_5dias	dias_tmin_menor15_15dias	dias_prcpt_maior1mm_20dias
prcpt_media_5dias	tmax_media_15dias	prcpt_maior_20dias

3.5 Modelagem

A técnica de modelagem escolhida foi a indução de árvores de decisão (QUINLAN, 1986) para classificação, por meio do classificador J48 do Weka, que é uma versão adaptada do algoritmo clássico C4.5, desenvolvido por Quinlan (1993).

A partir do conjunto de dados final, também chamado de conjunto de treinamento, foram realizados diversos experimentos a fim de se avaliar e caracterizar o desempenho dos

modelos gerados a partir da aplicação do classificador J48. Esses experimentos são caracterizados por diferenças nos subconjuntos de treinamento utilizados, tanto em relação aos registros quanto nas diferenças nos parâmetros utilizados para a geração dos modelos.

Os primeiros experimentos foram realizados para avaliar a influência dos *outliers* e do número mínimo de objetos por folha no modelo. Inicialmente, para identificar a influência dos *outliers*, foi aplicado o filtro *InterquartileRange*, método não supervisionado existente no Weka, baseado em análise interquartílica que classifica registros como *outliers* ou "valores extremos", que no caso, também foram tratados como *outliers*.

Na maioria dos casos, os *outliers* podem ser considerados como ruído no conjunto de dados, o que pode influenciar o algoritmo de modelagem e trazer resultados que não representam as principais características do conjunto de dados.

Após a aplicação desse filtro, os registros identificados como *outliers* foram removidos do conjunto de treinamento original, obtendo assim um novo conjunto de treinamento. A partir de cada um desses dois conjuntos de treinamento (o original e o sem *outliers*) foram realizados experimentos a fim de verificar a diferença de desempenho dos modelos gerados, variando-se também o número mínimo de objetos por folha, a partir do parâmetro "minNumObj" do classificador J48.

O número mínimo de objetos por folha é um parâmetro utilizado como técnica de pré-poda a fim de deixar o modelo mais genérico ou mais específico, uma vez que altera o critério de divisão de atributos, transformando-os em novos ramos ou folhas, o que faz com que o número de regras seja menor, para um modelo mais simples e genérico, ou um número de regras maior, para um modelo mais complexo e específico. Os principais objetivos dessa estratégia foram deixar o modelo mais genérico e, eventualmente, permitir a visualização e interpretação dos modelos obtidos.

A partir dos resultados, foi selecionado o melhor modelo considerando a configuração entre conjunto de treinamento e parâmetros utilizados. Esse modelo foi denominado modelo preditivo, pois no caso de uma aplicação do modelo, ele poderia ser utilizado automaticamente, sem a interpretação das regras obtidas, apenas com a obtenção da classe "Oc" para ocorrência ou "NaoOc" para não ocorrência.

Com o objetivo de se obter um modelo em árvore de decisão que fosse possível de ser interpretado visualmente e assim identificar os fatores, cenários ou períodos que classificaram casos de ocorrência e casos de não ocorrência da ferrugem asiática da soja, foi necessário realizar mais uma etapa de modelagem.

Para interpretar visualmente o modelo preditivo foi necessário aplicar uma técnica capaz de reduzir o número de folhas (regras). Uma das técnicas que pode gerar esse efeito é a eliminação de registros considerados inconsistentes, ruídos e *outliers*, pois, no momento da indução do modelo de árvore de decisão, a presença desses dados pode fazer com que regras específicas para esses registros sejam criadas, aumentando o número de regras (HAN et al., 2011).

Diferentemente da técnica de detecção de *outliers* chamada *InterquartileRange*, que avalia os atributos de acordo com a distribuição dos valores e elimina os registros que apresentaram valores distantes da média (de acordo com parâmetros previamente estabelecidos), para essa etapa foi utilizado o filtro *RemoveMisclassified*, embutido no Weka, que, ao encontrar registros que não obedecem o padrão principal obtido pelo filtro classificador e foram classificados incorretamente, são removidos do conjunto de treinamento.

Um dos parâmetros deste filtro foi a escolha do classificador utilizado, e neste caso, foi utilizado o mesmo classificador J48 que gerou o modelo preditivo, e que também foi utilizado na etapa seguinte, na obtenção do modelo para interpretação de árvore de decisão. Os parâmetros do classificador J48 para o filtro foram os mesmos que geraram o modelo preditivo.

3.6 Softwares e parâmetros

Os principais softwares utilizados para a produção deste trabalho foram o Weka (HALL et al., 2009) versão 3.6.4 e o software R versão 2.13. Os algoritmos desenvolvidos na fase de preparação de dados foram feitos na linguagem *Python* 2.6. A plataforma na qual foi realizado este trabalho foi a distribuição Ubuntu do sistema operacional Linux. A principal característica dos softwares utilizados é que todos são livres e gratuitos, sem a necessidade de licenças especiais para realizar este trabalho.

A programação em *Python* na fase de preparação dos dados facilitou a formatação e a integração dos dados do Consórcio Antiferrugem e do Agritempo. Essa linguagem foi escolhida em função de experiência prévia com a mesma.

O software estatístico R, que abrange técnicas estatísticas e gráficas para análise de dados, também é um software livre, gratuito (http://www.r-project.org/) e distribuído sob a licença de uso GNU GPL (*General Public License*).

O R foi utilizado, principalmente, na fase de entendimento dos dados, para a produção de gráficos e tabelas. Todas as funções e os pacotes utilizados já estavam instalados por padrão, exceto o pacote "maptools", necessário para produzir a Figura 4.6, no Capítulo 4.

O software Weka (*Waikato Environment for Knowledge Analysis*), desenvolvido na Universidade de Waikato, Nova Zelândia, é uma coleção de algoritmos de aprendizado de máquina e outras ferramentas de análise, que oferece suporte ao processo completo de mineração de dados (WITTEN et al., 2011). Ele é um software livre, gratuito (http://www.cs.waikato.ac.nz/ml/weka/) e distribuído sob a licença de uso GNU GPL. Foi escolhido por sua facilidade de uso e ter sido apontado como um dos principais softwares livres utilizados para mineração de dados em uma pesquisa, como mostra a Figura 3.4.

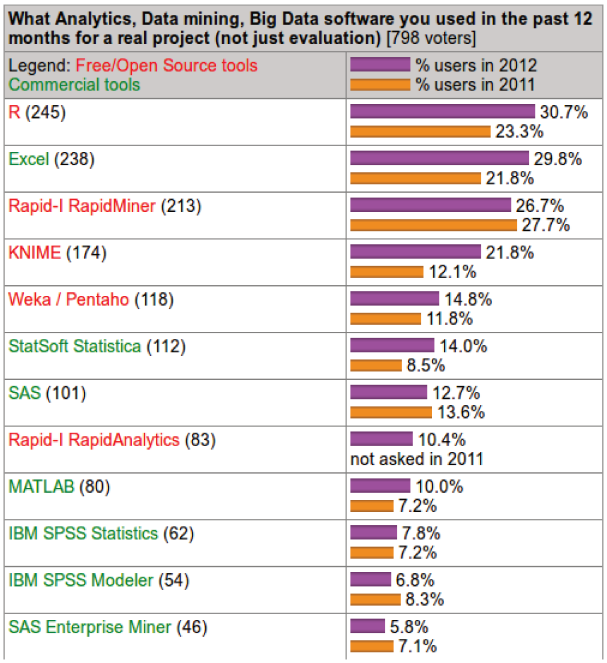


Figura 3.4: Levantamento publicado em maio de 2012 que apresenta os doze softwares mais utilizados em 2011 e 2012 com 798 participantes.
Fonte: http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html

No Weka, existem diferentes ambientes que permitem o uso das ferramentas para mineração de dados, são eles: “Explorer”, “Experimenter”, “KnowledgeFlow” e “Simple CLI”. Para este trabalho foi utilizado o ambiente gráfico interativo “Explorer”. Neste ambiente, assim como nos outros, porém cada um com características próprias, é possível personalizar diversos parâmetros dos algoritmos e ferramentas, dependendo da necessidade de cada problema.

Em cada método ou algoritmo do Weka, diversos parâmetros podem ser ajustados, dependendo de cada problema. Para este trabalho, os parâmetros dos filtros e dos métodos de seleção de atributos utilizados foram mantidos com os valores padrão (*default*) para maioria dos casos.

Na fase de modelagem, ao utilizar o classificador J48, dentre os parâmetros que podiam ser ajustados, apenas o número mínimo de objetos por folha, representado por “minNumObj” na Figura 3.5, foi alterado para uma sequência de experimentos. Os filtros e métodos de seleção de atributos que fizeram uso do classificador J48 também tiveram o valor desse parâmetro alterado. Os resultados da variação desse parâmetro podem ser encontrados na Seção 4.2.

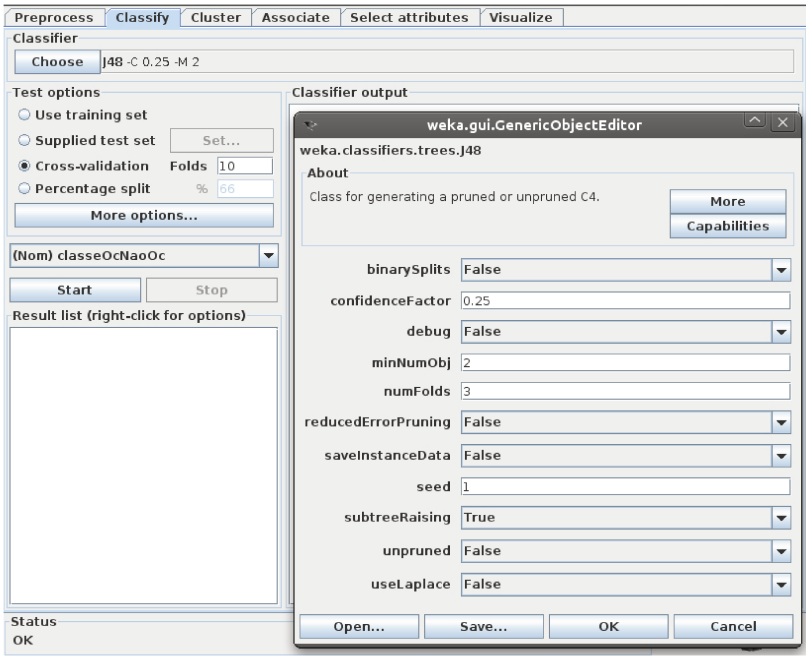


Figura 3.5: Janela de opções do classificador J48 do Weka.

4 RESULTADOS E DISCUSSÃO

4.1 Análise exploratória

4.1.1 Exploração dos dados do Consórcio Antiferrugem

O conjunto de dados utilizado para gerar os resultados deste trabalho foi o conjunto de dados proveniente do Consórcio Antiferrugem de quatro safras, 2007/2008 a 2010/2011. As ocorrências da ferrugem asiática foram caracterizadas de acordo com a distribuição da safra, mês, estádio de desenvolvimento da planta e localização geográfica. A distribuição de ocorrências conforme a safra está representada pela Figura 4.1.

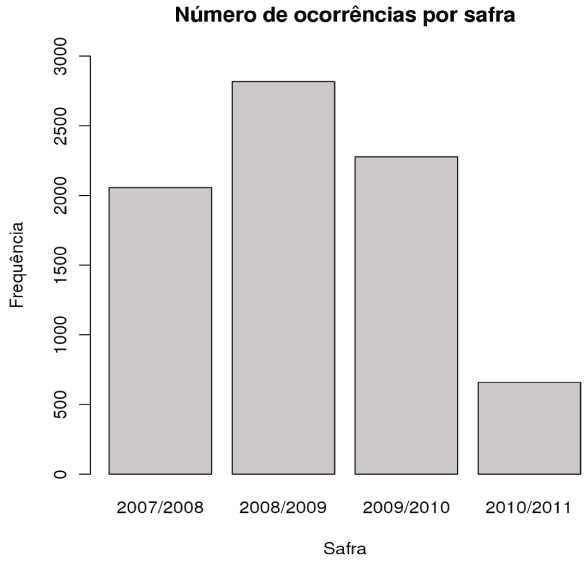
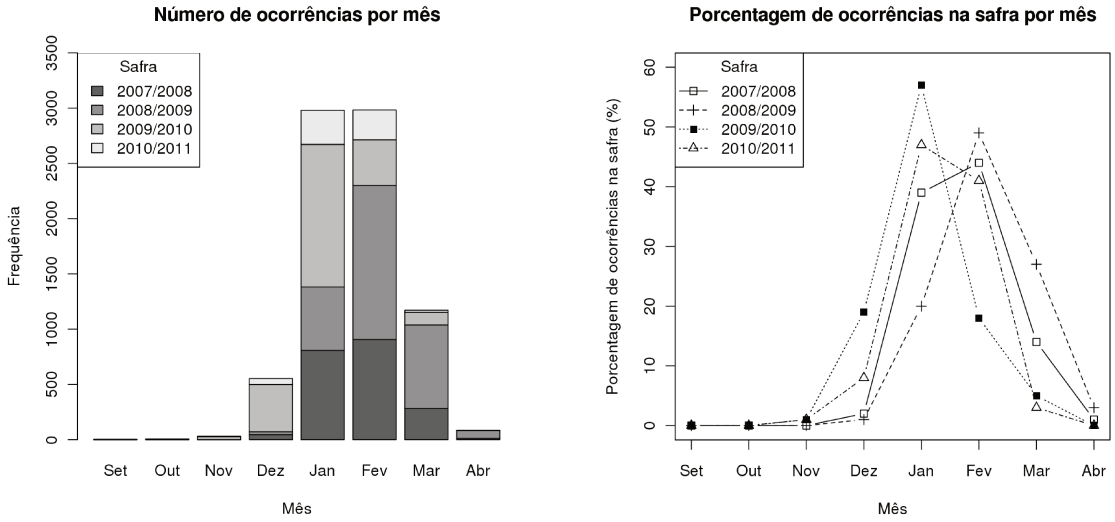


Figura 4.1: Distribuição do número de ocorrências da ferrugem asiática da soja por safra.

De acordo com a Figura 4.1, observa-se um número menor de registros de ocorrências na safra 2010/2011. Alguns motivos para a diminuição de ocorrências foram a estiagem, o vazio sanitário e o monitoramento das plantações (CASTRO, 2011). Comportamento que contrasta com a safra 2008/2009, na qual houve um número maior de ocorrências, fato vinculado a um certo descuido dos produtores e ao aumento dos custos de combate à doença (MARQUES, 2011).

A distribuição de ocorrências da ferrugem de cada safra pode ser visualizada de duas maneiras diferentes na Figura 4.2, de acordo com o mês e a safra.

Ao analisar a Figura 4.2(a), destaca-se a safra 2009/2010, em relação ao número de ocorrências em dezembro e em janeiro, pois são maiores que nas outras safras. Na Figura 4.2(b) onde não há a influência do número absoluto de registros, percebe-se dois tipos de comportamento de ocorrências da doença: nas safras 2007/2008 e 2010/2011; e nas safras 2008/2009 e 2009/2010.



(a) (b)

Figura 4.2: Distribuição de ocorrências da ferrugem asiática da soja em relação à safra e ao mês. 4.2(a) Distribuição de ocorrências por mês e safra, em números absolutos. 4.2(b) Porcentagem de ocorrências em relação à safra, com distribuição nos meses.

Para comparar o comportamento do número de ocorrências em cada safra, foi realizado o teste de correlação de Pearson, utilizando dados de cada mês em cada safra. A matriz de correlação entre o número de ocorrências de cada mês está descrita na Tabela 4.1.

Tabela 4.1: Coeficiente de correlação entre as safras, considerando o número de ocorrências por mês.

Safras	**2007/2008**	**2008/2009**	**2009/2010**	**2010/2011**
2007/2008	1,00	0,88	0,74	0,96
2008/2009	0,88	1,00	0,38	0,72
2009/2010	0,74	0,38	1,00	0,86
2010/2011	0,96	0,72	0,86	1,00

As safras 2007/2008 e 2010/2011 apresentaram comportamentos similares, com poucas ocorrências em dezembro e março, sendo que na safra 2010/2011, houve mais registros de ocorrências em dezembro e janeiro, e menos em fevereiro e março, em relação aos mesmos períodos da safra 2007/2008. Porém, ainda há uma relação forte, como indica o coeficiente de correlação de 0,96.

As safras 2008/2009 e 2009/2010 também apresentaram comportamentos similares, porém, com um aparente deslocamento de um mês, conforme a Figura 4.2(b). Na Tabela 4.1 pode ser observado o baixo valor do coeficiente de correlação (0,38), porém, ao considerar os dados de outubro a abril na safra 2008/2009 e de setembro a março na safra 2009/2010, o coeficiente de correlação foi de 0,96, dando subsídios a hipótese de deslocamento de registro de ocorrências em um mês entre essas duas safras.

Também foi realizada uma análise sobre a distribuição de ocorrências da doença de acordo com o estádio de desenvolvimento da planta, ilustrada pelas Figuras 4.3 e 4.4. Embora a planta possa ser infectada em qualquer estádio de desenvolvimento (EMBRAPA, 2010), observa-se um número crescente de registros de ocorrências conforme o estádio de desenvolvimento da planta avança, com um pico de ocorrências no estádio R5, equivalente à fase de enchimento de grãos, segundo a Tabela 3.2. Após o pico de registros de ocorrência, há uma diminuição do número de registros de ocorrências até o momento de colheita.

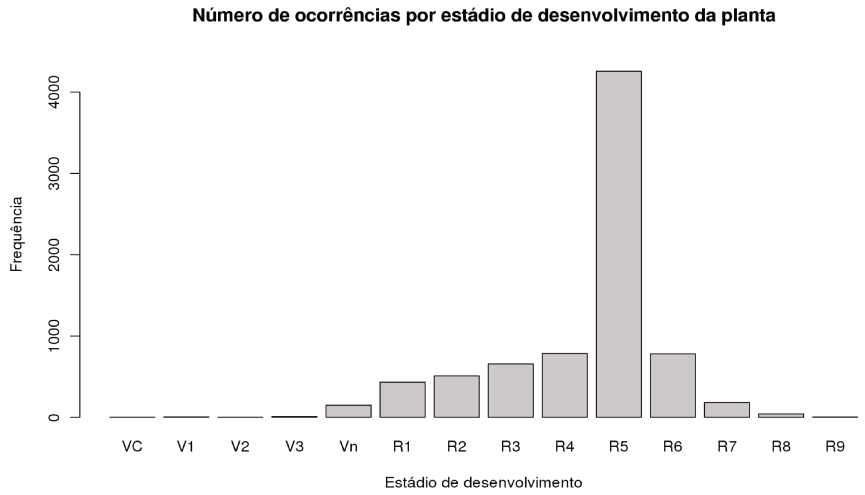


Figura 4.3: Distribuição de ocorrências por estádio de desenvolvimento da planta.

Para complementar a Figura 4.2, a Figura 4.4 apresenta a distribuição das ocorrências de acordo com o estádio, mês e safra. Um dos destaques é a semelhança de comportamento entre as safras 2008/2009 e 2009/2010, porém, com um aparente deslocamento de um mês entre o número de registros de ocorrências, como também foi observado na Figura 4.2(b).

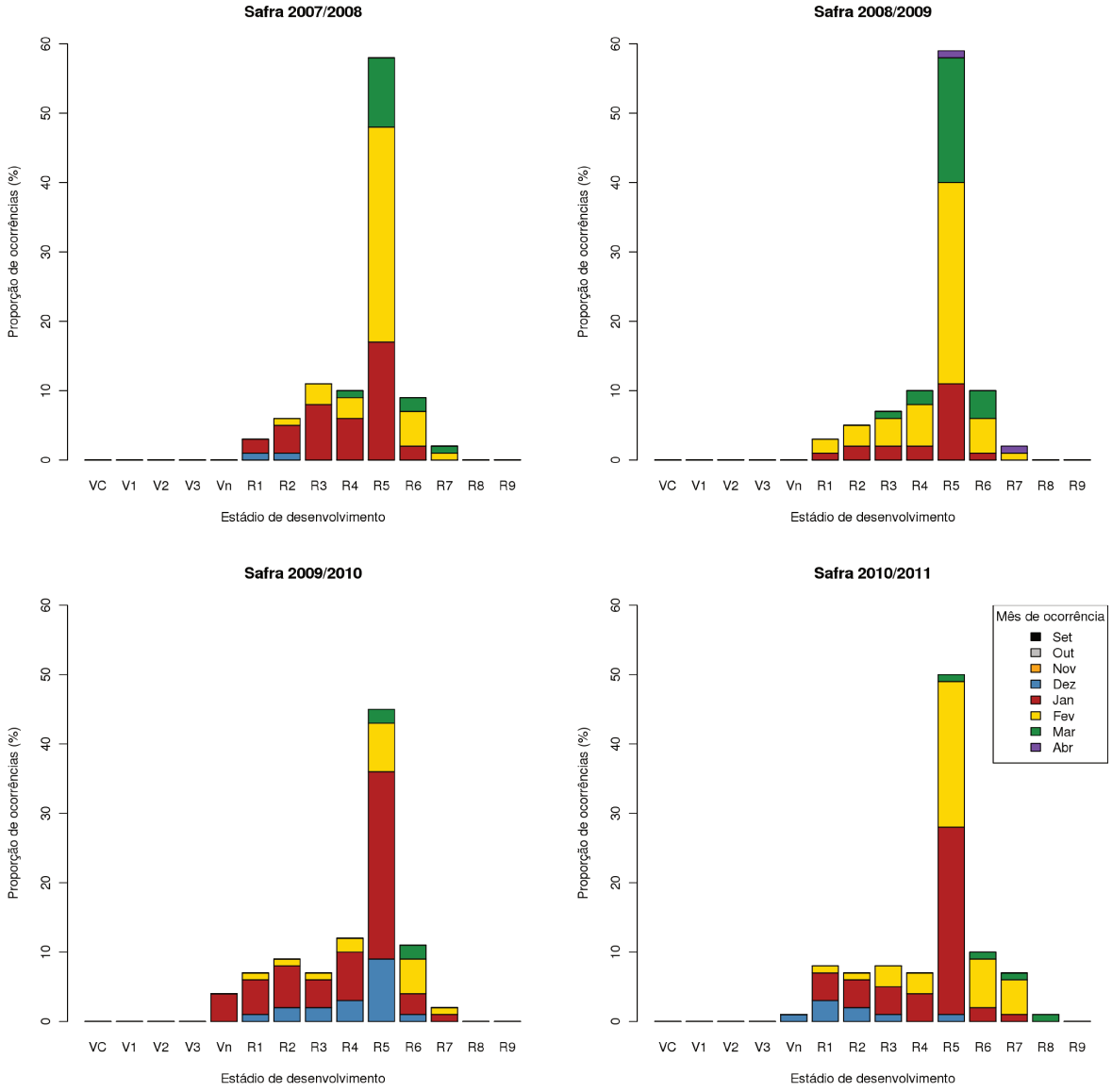


Figura 4.4: Distribuição da frequência relativa de ocorrências em cada safra, distribuídas pelos estádios de desenvolvimento da planta e mês.

Conforme o Consórcio Antiferrugem (2010a), as chuvas abundantes (acumulado entre 01 de novembro e 30 de dezembro, em milímetros) na safra 2009/2010 dificultaram o

controle químico com antecedência e contribuíram para o espalhamento da doença, principalmente para a região Centro-Oeste e para os estados do Paraná e Rio Grande do Sul, o que pode indicar o aumento do número de ocorrências nesse mês em relação às outras safras.

Ao comparar as safras 2007/2008 e 2010/2011, também foram observados registros de ocorrências no mês de dezembro, porém, em estádios precedentes ao de enchimento de vagens (Vn, R1 e R2), diferentemente da safra 2009/2010, quando também houve registros em dezembro, mas no R5.

Os resultados a seguir apresentam o período de plantio da soja. A unidade utilizada no banco de dados foi a frequência quinzenal e mensal de plantio. As Tabelas 4.2 e 4.3 apresentam a distribuição dos registros de ocorrências da doença de acordo com o período no qual houve o plantio.

Tabela 4.2: Mês de plantio da soja conforme o conjunto de dados do Consórcio Antiferrugem.

Mês de plantio	Ausente	1	2	6	9	10	11	12
Frequência	2872	23	17	2	79	2193	2267	357

Tabela 4.3: Quinzena do mês de plantio da soja conforme o conjunto de dados do Consórcio Antiferrugem.

Quinzena de plantio	1	2
Frequência	7513	297

De acordo com a Tabela 4.2, mais de 36% dos registros apresentaram dados faltantes para o mês de plantio da soja, e dos dados não faltantes, mais de 90% se concentraram nos meses de outubro e novembro. Já para o atributo “quinzenaPlantio”, representado na Tabela 4.3, em mais de 96% dos casos refere-se à primeira quinzena do mês, fazendo com que este atributo não apresente uma variação considerável ou possa abrigar algum conhecimento oculto. Assim, ambos os atributos foram descartados para a composição do conjunto final de dados utilizado na fase de modelagem.

Nas próximas análises foram consideradas as informações de localização das ocorrências. Essas informações foram concentradas em estados, municípios e suas respectivas latitude e longitude. A Figura 4.5 apresenta a distribuição de ocorrências nos estados do Brasil, separadas por safra.

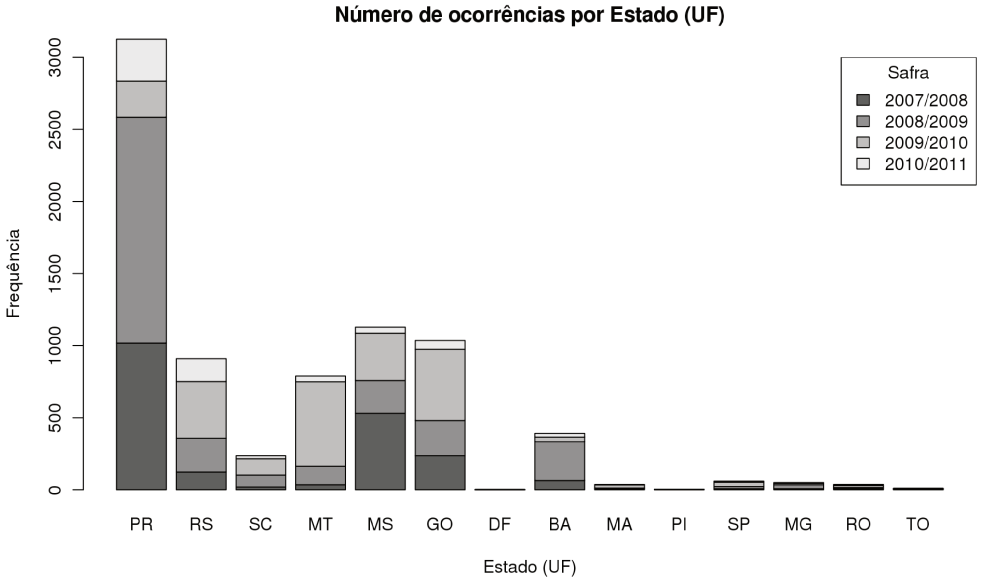


Figura 4.5: Distribuição do número de ocorrências em relação ao estado do Brasil.

A partir do gráfico da Figura 4.5, pode-se perceber que a safra 2009/2010 apresentou um aumento de ocorrências nas regiões Sul e Centro-Oeste, exceto no estado do Paraná. A região Sul é uma das mais afetadas pela presença do El Niño, principalmente pelo aumento das chuvas nos meses de outubro, novembro e dezembro (CPTEC, 2012).

Buscando-se compreender a distribuição espacial das ocorrências do conjunto de dados, a Figura 4.6 mostra a representação cartográfica do Brasil e os municípios nos quais foram detectadas as ocorrências, nas diferentes safras, considerando a latitude e a longitude de cada município provenientes do banco de dados do Consórcio Antiferrugem.

Vale ressaltar que os municípios podem apresentar mais de uma ocorrência por safra, assim, alguns pontos no mapa podem ser sobrepostos.

A presença da ferrugem asiática da soja é marcante em todas as regiões produtoras, variando o número de focos e número de municípios atingidos.

A exploração dos dados do Consórcio Antiferrugem proporcionou uma visão geral do comportamento das ocorrências da ferrugem asiática nas principais regiões produtoras de soja do Brasil. Os diversos atributos analisados serviram para definir o escopo deste trabalho, que é a modelagem baseada em elementos meteorológicos, buscando-se características intrínsecas à

doença e sua interação com o ambiente, não utilizando atributos relacionados à localização especial ou temporal dos registros no conjunto de dados final.

Figura 4.6: Distribuição espacial das ocorrências em cada safra.

4.1.2 Exploração dos dados meteorológicos

Para visualizar os resultados da correlação entre os dados provenientes de estações reais de superfície e de estações virtuais, foram elaborados mapas do Brasil contendo pontos

representando os municípios que apresentaram dados onde houve a comparação. Para correlacionar os dados, foram obtidos valores acumulados em 5 dias de cada variável meteorológica.

Foram elaborados três mapas, um para a temperatura mínima, um para a temperatura máxima e um para a precipitação. Para facilitar a visualização, foram definidos três limites de correlação, sendo abaixo de 0,5, entre 0,5 e 0,8 e acima de 0,8.

Foi observado que haviam mais dados faltantes de precipitação do que os de temperaturas. Os coeficientes de correlação entre as estações virtuais e reais para dados de precipitação também apresentaram valores menores. Também foi observado por Romani et al. (2003b; 2007) que os dados de precipitação são mais difíceis de apresentar boas estimativas.

Alguns municípios que apresentaram baixos coeficientes de correlação entre os dados foram avaliados separadamente, e foram observados comportamentos irregulares em alguns municípios, tanto em estações reais, quanto virtuais, para as temperaturas mínima e máxima, e precipitação. Algumas situações encontradas estão ilustradas nas Figuras 4.8, 4.9 e 4.10.

A Figura 4.8 ilustra os dados da temperatura mínima para o município de São Luiz Gonzaga, RS, provenientes da estação meteorológica real e da estação meteorológica virtual. Os dados provenientes da estação virtual acompanham as estações do ano, com um comportamento dentro do esperado, diferentemente dos dados provenientes da estação real, o que pode indicar alguma falha na aquisição dos dados, por falha de sensores ou humana.

Em relação à temperatura máxima, a Figura 4.9 ilustra o comportamento para o município de Santa Vitória do Palmar, RS. Aparentemente, este município também apresenta problemas com as medições na estação real, pelo menos até o ano de 2010. Como a estação virtual utiliza os dados de temperaturas de municípios vizinhos também, as simulações para a temperatura apresentam comportamento mais próximo ao esperado, porém, próximo ao ano de 2008, os dados ainda apresentam um comportamento irregular.

A Figura 4.10 apresenta os dados de precipitação do município de Luís Eduardo Magalhães, BA. Observa-se que entre os anos de 2008 e 2009 pode ter havido algum problema com a estação meteorológica real, pois os registros de precipitação permaneceram no zero. Pela estimativa do TRMM, houve eventos de precipitação, porém, houve falhas no ano de 2009, assim como na estação real.

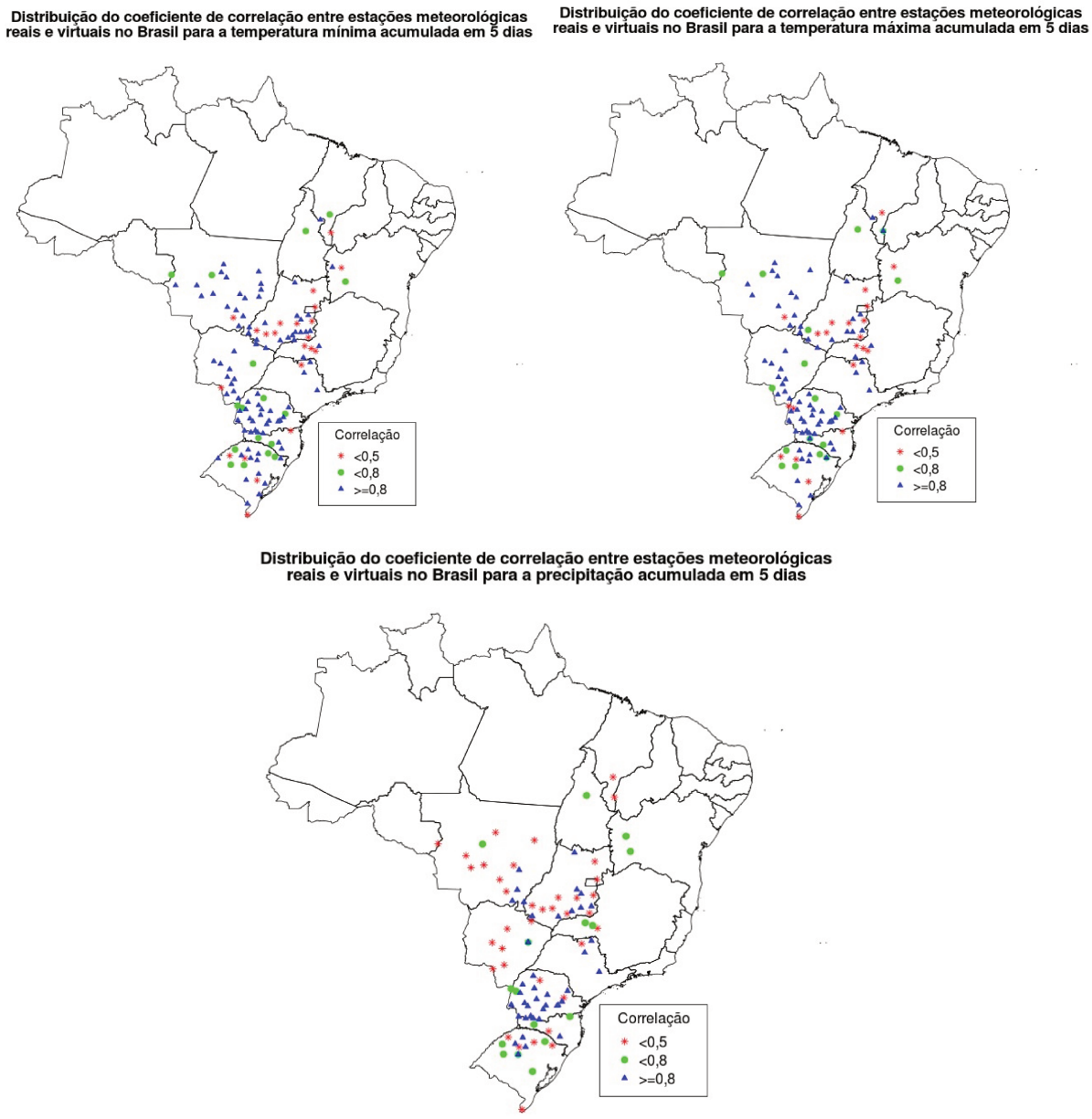


Figura 4.7: Distribuição espacial das estações meteorológicas que tiveram os dados reais e virtuais acumulados em 5 dias e comparados a partir do coeficiente de correlação. Acima e à esquerda, para a temperatura mínima. Acima e à direita, para a temperatura máxima. Abaixo, para a precipitação.

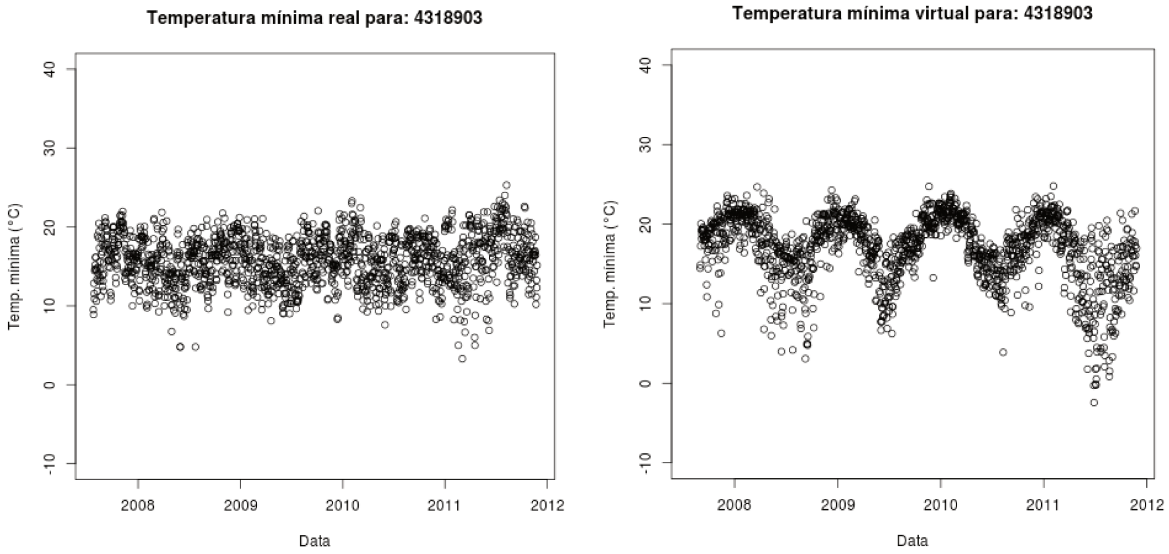


Figura 4.8: Gráficos da temperatura mínima para o município de São Luiz Gonzaga, RS (código IBGE 4318903), provenientes das estações meteorológicas real (esquerda) e virtual (direita).

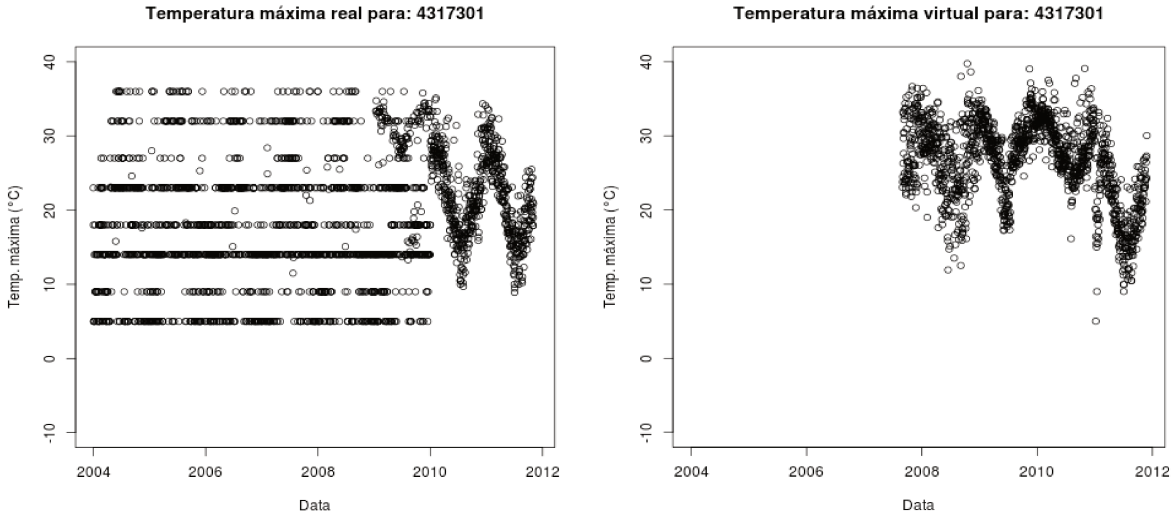


Figura 4.9: Gráficos da temperatura máxima para o município de Santa Vitória do Palmar, RS (código IBGE 4317301), provenientes das estações meteorológicas real (esquerda) e virtual (direita).

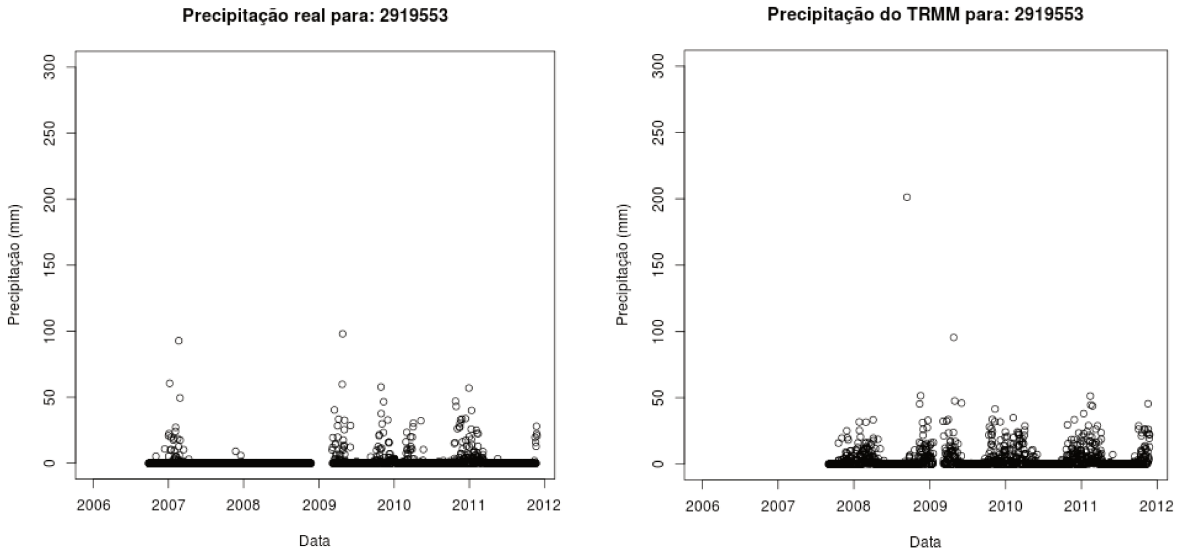


Figura 4.10: Gráficos da precipitação para o município de Luís Eduardo Magalhães, BA (código IBGE 2919553), provenientes da estação meteorológica real e do radar do satélite TRMM.

Para contornar a possível presença de valores estranhos no conjunto de dados final, a estratégia utilizada foi aplicar um filtro de identificação e remoção de *outliers*, descrita na seção a seguir.

4.2 Modelo de predição de ocorrências da ferrugem asiática da soja

Os resultados dos primeiros experimentos com o objetivo de selecionar o melhor modelo preditivo estão representados na Tabela 4.4. O conjunto de dados final obtido após a fase de preparação de dados foi composto por 12.591 registros, entre ocorrências e não ocorrências, pelos atributos meteorológicos e pelo atributo meta, somando-se 46 atributos.

A Tabela 4.4 apresenta a taxa de acerto, a estatística kappa e o número de folhas de cada modelo gerado, utilizando o classificador J48 com o método de avaliação de validação cruzada, variando-se o número mínimo de objetos por folha e considerando dois conjuntos de treinamento: (i) conjunto original (12.591 registros) e (ii) conjunto após a remoção de *outliers* (8.895 registros) utilizando o filtro *InterquartileRange*. Os outros parâmetros do classificador foram mantidos em seus valores padrão (*default*).

Ao analisar os três gráficos nas Figuras 4.11, 4.12 e 4.13, pode ser observado que a presença de *outliers* não afetou significativamente o desempenho dos modelos. Esse resultado é um exemplo que corrobora o conceito de que algoritmos de árvores de decisão são robustos o suficiente para minimizar o efeito da presença de ruídos nos dados (TAN et al., 2009).

Tabela 4.4: Avaliação baseada na taxa de acerto, estatística kappa e número de folhas, dos modelos em árvore de decisão sob os efeitos da variação do número mínimo de objetos por folha, utilizando um conjunto com todos os registros e um subconjunto excluindo-se os *outliers.*

Número mínimo de objetos por folhas	**Conjunto original: 12.591 registros**			**Conjunto sem *outliers*: 8.895 registros**		
	Taxa de acerto (%)	**Kappa**	**Número de folhas**	**Taxa de acerto (%)**	**Kappa**	**Número de folhas**
2	87,00	0,74	620	86,03	0,71	454
5	86,01	0,72	455	84,80	0,69	338
10	84,13	0,68	305	83,21	0,65	234
15	82,94	0,66	238	82,39	0,63	171
20	82,40	0,65	187	81,12	0,61	141
30	80,99	0,62	145	80,09	0,59	103
40	79,52	0,59	116	79,45	0,57	87
50	79,21	0,58	85	78,36	0,55	70
75	77,72	0,55	71	77,40	0,53	49
100	77,17	0,54	46	76,18	0,50	38

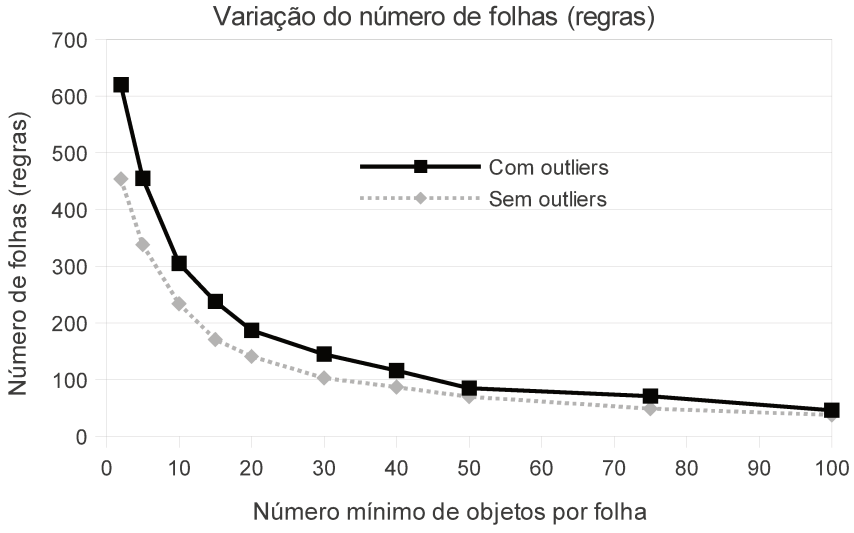


Figura 4.11: Variação do número de folhas em relação ao número mínimo de objetos por folha utilizando validação cruzada.

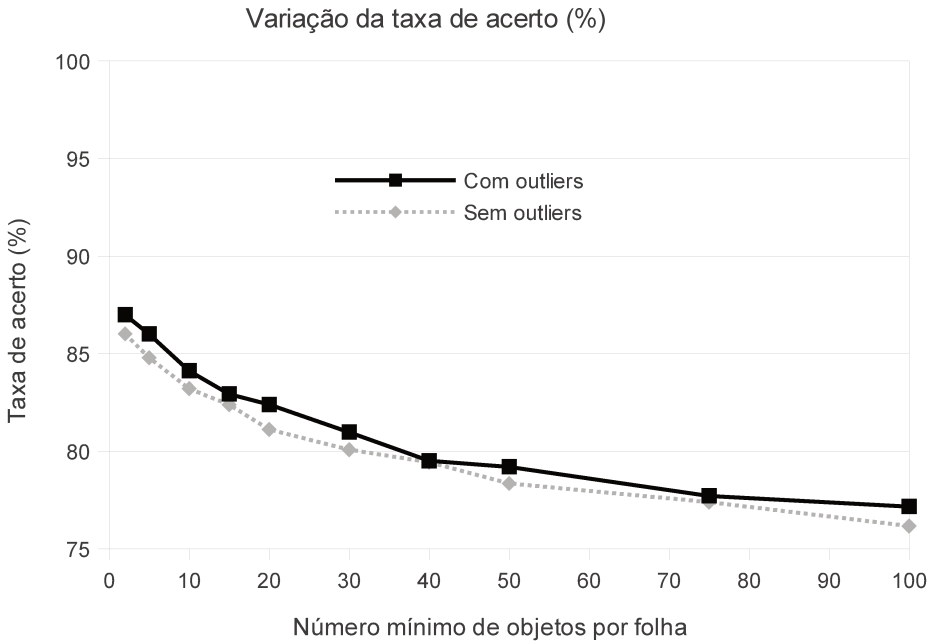


Figura 4.12: Variação da taxa de acerto em relação ao número mínimo de objetos por folha utilizando validação cruzada.

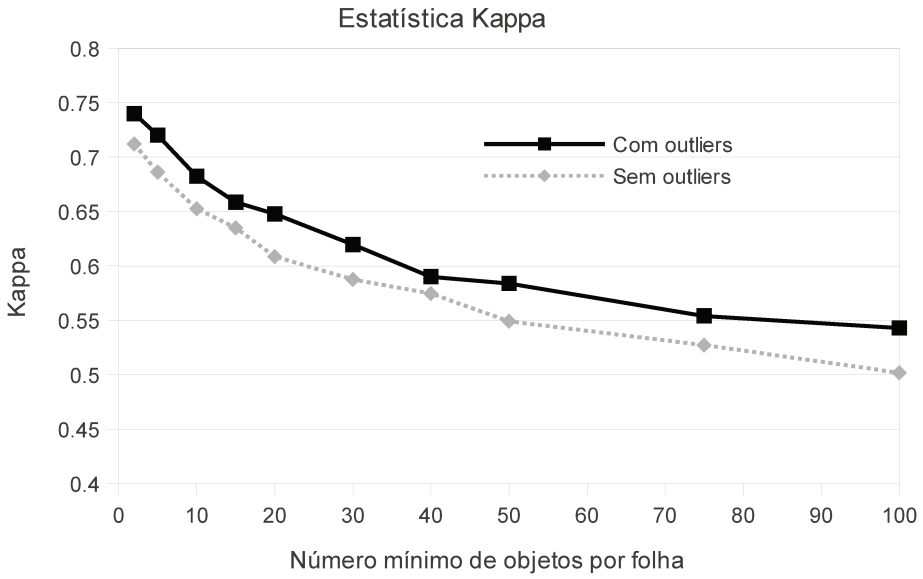


Figura 4.13: Variação da estatística kappa em relação ao número mínimo de objetos por folha utilizando validação cruzada.

Por não apresentarem diferenças significativas nos resultados em relação aos três aspectos avaliados, foi escolhido o conjunto com o maior número de registros (com *outliers*).

A principal motivação de utilizar o parâmetro de pré-poda que varia o número mínimo de objetos por folha foi definir um número de registros que pudessem ser avaliados ao mesmo tempo sem que houvesse grandes perdas no desempenho e abrangência do modelo.

Assim, para este conjunto de dados, a faixa de valores entre 30 e 50 objetos por folha foi considerada como ideal, pois não há grande variação no número de folhas (regras), apresenta um valor de kappa próximo a 0,6, que representa um bom índice de concordância para o modelo (Tabela 2.2), e a taxa de acerto (acurácia) se mantém próxima a 80%, o que foi considerado adequado para este trabalho. Inicialmente, o número mínimo de objetos por folha foi definido como 40, com a possibilidade de usar outros valores (30 ou 50) em trabalhos futuros.

Assim, o resultado do modelo preditivo para ocorrências e não ocorrências da ferrugem asiática da soja apresentou uma taxa de acerto de 79,52%, com a estatística kappa de 0,59 e um total de 116 folhas do modelo.

Porém, a partir dos resultados mostrados até o momento, não foram obtidas as informações sobre quais os fatores que favoreceram ou não a ocorrência da doença. Para isso seria necessário visualizar o modelo em formato de árvore de decisão e interpretá-lo, uma vez que esse tipo de modelagem é capaz de apresentar os atributos mais importantes e os valores que influenciam a decisão do algoritmo para classificar uma situação de ocorrência ou de não ocorrência da ferrugem asiática.

4.3 Modelo para interpretação das ocorrências da ferrugem asiática da soja

A partir do modelo preditivo, o filtro *RemoveMisclassified* foi utilizado para obtenção do modelo interpretativo. Para interpretar o modelo, as decisões que o algoritmo realizou foram analisadas de forma geral, buscando compor um cenário capaz de explicar a influência de fatores meteorológicos nas ocorrências, ou não ocorrências, da doença.

Para dar subsídios à interpretação visual do modelo, foi composto um resumo com medidas estatísticas de posição dos atributos utilizados no conjunto de treinamento, na Tabela

4.5. Em seguida, o modelo em árvore de decisão foi apresentado na Figura 4.14, interpretado e discutido com apoio na literatura.

Tabela 4.5: Medidas de posição da distribuição dos atributos do conjunto de treinamento, com mínimo, primeiro quartil, mediana, média, terceiro quartil, máximo e número de dados faltantes.

(continua)

Atributo	Mínimo	1ºQ.	Mediana	Média	3ºQ.	Máximo	Faltantes
tmin_media_Latente	14	18,1	19,14	19,16	20,35	23,11	0
dias_tmin_menor15_Latente	0	0	0	0,1332	0	5	0
tmax_media_Latente	23,19	28,12	29,21	29,13	30,27	34,43	0
dias_tmax_maior30_Latente	0	0	2	3,07	5	10	0
tmed_media_Latente	19,42	23,14	24,16	24,14	25,18	28,53	0
prcpt_media_Latente	0	3,4	6,35	7,301	10,03	39,72	5
prcpt_acum_mm_Latente	0	27,63	52,77	59,93	83,11	357,45	5
dias_prcpt_maior1mm_Latente	0	3	5	4,797	7	9	5
prcpt_maior_Latente	0	12,96	22,15	25,88	33,63	157,59	5
tmin_media_5dias	12,25	17,97	19,07	19,04	20,19	22,89	0
dias_tmin_menor15_5dias	0	0	0	0,1014	0	5	0
tmax_media_5dias	19,24	28,06	29,26	29,13	30,33	35,09	0
dias_tmax_maior30_5dias	0	0	1	1,747	3	5	0
tmed_media_5dias	15,74	23,11	24,11	24,08	25,25	28,64	0
prcpt_media_5dias	0	2,35	6,47	7,597	11,28	86,97	4
prcpt_acum_mm_5dias	0	11,67	32,34	37,9	56,42	434,87	4
dias_prcpt_maior1mm_5dias	0	2	3	2,905	4	5	4
prcpt_maior_5dias	0	7,65	16,96	20,88	30,35	235,78	4
tmin_media_10dias	14,13	18	19,05	19,03	20,07	23,03	0
dias_tmin_menor15_10dias	0	0	0	0,2228	0	8	0
tmax_media_10dias	22,98	28,11	29,18	29,16	30,31	34,18	0
dias_tmax_maior30_10dias	0	1	3	3,66	6	10	0
tmed_media_10dias	18,8	23,14	24,09	24,1	25,13	28,26	0
prcpt_media_10dias	0	4,04	6,93	7,699	10,46	43,49	4
prcpt_acum_mm_10dias	0	40,05	69,1	76,61	104,36	434,87	4
dias_prcpt_maior1mm_10dias	0	4	6	5,822	8	10	4
prcpt_maior_10dias	0	16,24	26,66	30,11	41,01	235,78	4
tmin_media_15dias	14,02	17,99	18,97	19	20,09	22,91	0
dias_tmin_menor15_15dias	0	0	0	0,3657	0	10	0

Tabela 4.5: Medidas de posição da distribuição dos atributos do conjunto de treinamento, com mínimo, primeiro quartil, mediana, média, terceiro quartil, máximo e número de dados faltantes.

(conclusão)

Atributo	Mínimo	1ºQ.	Mediana	Média	3ºQ.	Máximo	Faltantes
tmax_media_15dias	22,65	28,25	29,2	29,21	30,28	33,72	0
dias_tmax_maior30_15dias	0	2	5	5,634	9	15	0
tmed_media_15dias	18,92	23,19	24,03	24,1	25,16	27,88	0
prcpt_media_15dias	0	4,64	7,17	7,737	10,28	32,8	4
prcpt_acum_mm_15dias	0	68,31	106,64	115,28	153,07	480,76	4
dias_prcpt_maior1mm_15dias	0	6	8	8,637	12	15	4
prcpt_maior_15dias	0	21,71	32,04	36,2	47,27	235,78	4
tmin_media_20dias	14,04	17,97	18,9	18,97	20,09	23,11	0
dias_tmin_menor15_20dias	0	0	0	0,5045	1	14	0
tmax_media_20dias	23,57	28,32	29,22	29,23	30,3	35,01	0
dias_tmax_maior30_20dias	0	3	7	7,64	12	20	0
tmed_media_20dias	19,59	23,19	24	24,1	25,17	27,94	0
prcpt_media_20dias	0	5,09	7,18	7,784	10,06	33,88	4
prcpt_acum_mm_20dias	0	101,6	142,9	154,8	200,7	677,6	4
dias_prcpt_maior1mm_20dias	0	8	11	11,51	15	20	4
prcpt_maior_20dias	0	26,53	37,28	41,13	49,98	235,78	4

O atributo *dias_tmin_menor15_20dias*, que avalia o número de dias com temperatura mínima abaixo de 15°C no período de 20 dias antes da possível data de infecção, foi considerado o atributo mais importante pelo modelo (nó 1). A partir deste nó, foram derivadas duas sub-árvores, sendo uma para o valor menor ou igual a zero, nesse caso, nenhum dia, e outra para valores maiores que zero, um ou mais dias.

Seguindo a sub-árvore para nenhum dia com temperatura mínima abaixo de 15°C no período de 20 dias antes da possível data de infecção, o atributo avaliado na sequência foi o *dias_tmax_maior30_20dias* (nó 2), e para menos de três dias nos quais a temperatura máxima foi superior a 30°C nos 20 dias anteriores à possível data de infecção, foi detectada a ocorrência da doença em 1370 casos (nó folha 4). Vale ressaltar que, de maneira geral, para este conjunto de dados e de acordo com a Tabela 4.5, registros com menos de três dias com temperaturas acima de 30°C em 20 dias ocorrem em aproximadamente 25% dos casos.

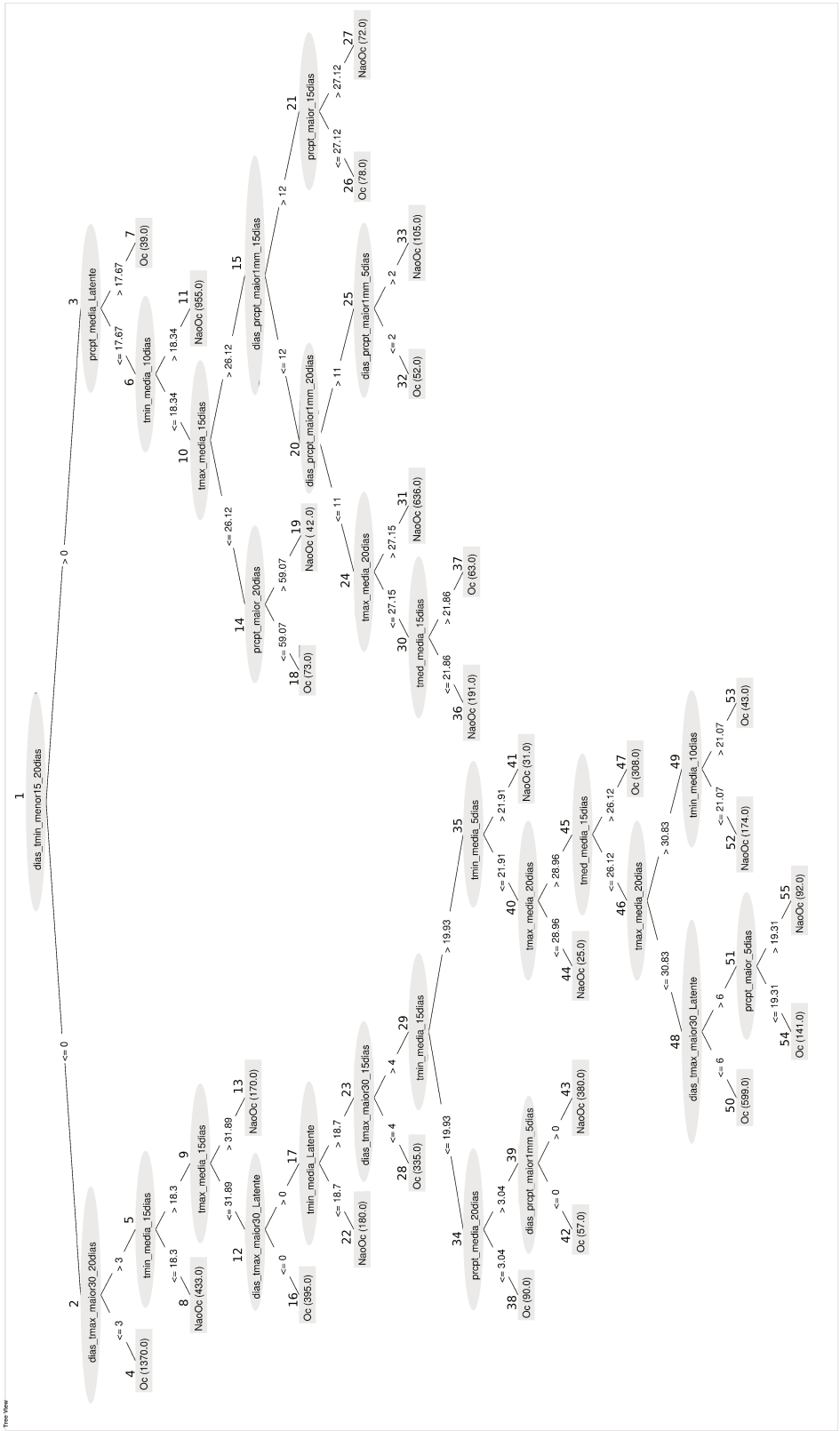


Figura 4.14: Modelo em árvore de decisão para interpretação.

Assim, considerando as condições de apenas três dias em um período de 20 dias com temperaturas acima de 30ºC e nenhum dia com temperaturas abaixo de 15ºC, pode haver um ambiente propício para a ocorrência da doença. Outros trabalhos podem corroborar esse ramo da árvore, sendo que no trabalho de Melching et al. (1989), foi apresentado que a faixa ótima de temperatura para a germinação do fungo foi de 18 a 26,5º e no de Marchetti et al. (1976), a temperatura ótima foi de 15 a 25ºC. Resultados similares foram encontrados por Kochman (1979), onde a temperatura ótima estava na faixa de 17 a 27ºC.

Para mais que três dias com temperatura máxima superior a 30°C nos 20 dias anteriores à possível data de infecção, foi avaliado o atributo *tmin_media_15dias* (nó 5) (média da temperatura mínima nos 15 dias anteriores à possível data de infecção) e para valores abaixo de 18,3°C, esse padrão foi detectado em 433 casos de não ocorrência da doença (nó folha 8). Valores abaixo de 18,3°C são abaixo da média para esse atributo, ou seja, são dias que atingem temperaturas mais baixas.

Considerando o outro ramo a partir do nó 5, onde a média da temperatura mínima nos 15 dias anteriores à possível data de infecção foi acima de 18,3°C, o nó seguinte avaliou o atributo *tmax_media_15dias* (nó 9) (média da temperatura máxima superior a 30°C nos 15 dias anteriores à possível data de infecção), no qual, para valores acima de 31,89°C houve 170 casos de não ocorrência da ferrugem asiática da soja (nó folha 13). Temperaturas acima de 31,89°C se encontram na porção de 75% dos maiores valores para esse atributo, assim, são dias que atingem temperaturas mais altas.

Para dar suporte a esses ramos, Kochman (1979) ressalta que temperaturas acima de 28,5°C são prejudiciais para a germinação do fungo. Alves et al. (2006) realizaram experimentos nos quais não houve germinação dos esporos em temperaturas acima de 30ºC. Em temperaturas abaixo de 15ºC ou acima de 30º, não houve infecção (CALDWELL et al., 2005 apud DEL PONTE e ESKER, 2008).

Com valores menores ou iguais a 31,89°C foi avaliado o atributo *dias_tmax_maior30_Latente* (nó 12) (número de dias com temperatura máxima acima de 30°C no período latente), onde, se não houve dias com temperatura máxima acima de 30°C no período latente, foi detectado esse padrão em 395 casos de ocorrência da doença (nó folha 16).

Vale ressaltar que o período latente considerado foi calculado de acordo com a Figura 3.3 e pode apresentar uma pequena variação no número de dias, cerca de 8 a 10 dias.

Para o período latente, o trabalho de Kochman (1979) evidencia a faixa ótima de 17 a 27ºC e a equação de segundo grau obtida por Alves et al. (2006), onde o mínimo é de 22,5ºC, ou seja, a temperatura ótima para o período latente (onde é o período mais curto). Em ambos os trabalhos, há evidências que quanto maior a temperatura além do ótimo, maior é o período latente, porém, não foram realizados experimentos de duração de período latente para temperaturas superiores a 30ºC, mesmo porque temperaturas acima de 30ºC já seriam desfavoráveis à germinação de esporos e infecção da planta, que são etapas do ciclo da doença anteriores ao período latente.

Em seguida ao nó 12, é avaliada a média da temperatura mínima também no período latente, pelo atributo *tmin_media_Latente* (nó 17). Para valores abaixo de 18,7°C, há 180 casos de não ocorrência da doença (nó 22). Como verificado por Alves et al. (2006), temperaturas abaixo do valor ótimo 22,5ºC também fazem com que o período latente seja maior.

A partir do nó 17, com valores superiores a 18,7°C, foi avaliado o atributo *dias_tmax_maior30_15dias* (nó 23) (número de dias que a temperatura máxima atingiu mais que 30°C no período de 15 dias antes da possível data de infecção), e, para valores menores ou iguais a quatro dias, foram detectados 355 casos de ocorrência da ferrugem (nó 28). Assim como no nó 2, com poucos dias com temperatura acima de 30ºC, há a tendência de detectar mais casos de ocorrência da doença.

Em relação à importância dos atributos avaliados, quanto mais o atributo está deslocado para a porção inferior do modelo, é considerado menos importante, uma vez que os melhores atributos são posicionados nos primeiros níveis da árvore (a partir da raiz), em algoritmos de indução de árvore de decisão (HAN et al., 2011). Outros destaques do modelo foram considerados a partir do número de casos avaliados em cada nó folha.

Considerando a sub-árvore a partir do nó 29, onde é avaliado o atributo *tmin_media_15dias* (média da temperatura mínima no período de 15 dias antes da possível data de infecção), o teste sobre o atributo foi com a temperatura de 19,93°C, e para valores abaixo dessa temperatura, foram classificados 380 casos de não ocorrência e 147 casos de

ocorrência, evidenciando o padrão também encontrado por outros ramos da árvore, onde, com menores temperaturas, há uma tendência para casos de não ocorrência da doença.

O nó 39 avalia o número de dias com precipitação acima de 1mm no período de 5 dias antes da possível infecção *(dias_prcpt_maior1mm_5dias)*. Nesse ponto o modelo apresenta ramos (nós folhas 42 e 43) contraditórios aos resultados encontrados na literatura, uma vez que há um conceito bem definido de deposição de esporos por meio de precipitação (ISARD et al., 2007, DEL PONTE et al., 2008; DUFAULT et al. 2010).

O esperado seria que, com alguma precipitação próxima à data de infecção, houvesse mais casos de ocorrência, não como mostrado pelo nó folha 43, onde foi verificado mais casos de não ocorrência. Esta aparente menor importância das variáveis de precipitação, no entanto, pode ter sido ocasionada pela inexistência de casos reais de não ocorrência da doença, uma vez que todo suposto evento de não ocorrência tenha precedido a ocorrência real correspondente. Nesse caso, faz sentido que eventos de temperatura tenham adquirido maior importância no estabelecimento das infecções.

Seguindo o lado da sub-árvore a partir do nó 29, para a média da temperatura mínima no período de 15 dias antes da possível data de infecção com valores acima de 19,93°C, foram classificados 1.091 casos de ocorrência da doença e 322 casos de não ocorrência, concordando novamente com outros trabalhos nos quais as faixas ótimas de temperatura para a germinação de esporos e infecção da planta têm início em 15°C (MARCHETTI et al., 1976), 17°C (KOCHMAN, 1979) ou 18,5°C (MELCHING et al.,1989).

O nó folha 47 apresentou 308 casos de ocorrências, também a partir da avaliação de diversos atributos relacionados com a temperatura, sendo o nó 45 o primeiro atributo relacionado com a temperatura média (temperatura média do período de 15 dias antes da possível data de infecção) desde o nó raiz.

O nó folha 50 foi um dos com maior número de casos (599) para a classe de ocorrência. Seguindo o ramo em direção ao nó raiz, são 13 restrições impostas pelos atributos que foram satisfeitas por esses casos classificados como ocorrência. Em todos esses 599 casos, a média da temperatura máxima no período de 20 dias antes da possível data de infecção está entre 28,96°C (nó 40) e 30,83°C (nó 46), a média da temperatura mínima no período de 15

dias antes da possível data de infecção foi maior que 19,93°C (nó 29) e no período de 5 dias antes da possível data de infecção foi menor que 21,91°C (nó 35).

A partir do nó raiz (nó 1), considerando a sub-árvore que se inicia com mais que um dia, com temperatura mínima menor que 15°C no período de 20 dias antes da possível data de infecção, há um maior número de casos classificados como não ocorrência (2001 casos) do que casos classificados como ocorrência (305 casos), evidenciando a característica que temperaturas abaixo de 15ºC são prejudiciais para a germinação de esporos ou infeção das plantas que provocam a ocorrência da doença (ALVES et al., 2006; BONDE et al., 2007; KOCHMAN, 1979; MARCHETTI et al., 1976; MELCHING et al.,1989).

O segundo atributo avaliado nessa sub-árvore foi o *prcpt_media_Latente* (nó 3) que considera a média da precipitação no período latente e, para valores acima de 17,67mm na média do período, houve 39 registros de ocorrência da doença (nó folha 7). Para o período latente, foi detectado o aumento da lesão para umidades relativas e período de molhamento foliar maiores (CALDWELL et al., 2005 apud DEL PONTE e ESKER, 2008), embora esse fato não evidencie uma relação direta com os registros de ocorrência ou com a germinação de esporos e infecção da planta.

Para valores menores que 17,67mm, foi avaliado o atributo *tmin_media_10dias* (nó 6) (média da temperatura mínima no período de 10 dias antes da possível data de infecção), onde, para valores acima de 18,34°C, foram detectados 955 casos de não ocorrência da doença (nó folha 11). Este ramo apresentou uma relação entre a temperatura mínima e a classe de não ocorrência diferente das outras relações obtidas neste modelo, onde, para este ramo, a média de temperaturas mínimas *acima* de 18,34°C foi a decisão do modelo, enquanto as outras relações obtidas foram para temperaturas mínimas *abaixo* de algum valor de teste, exceto para ramos onde são avaliados outros atributos na sequência.

O nó folha 31 foi um dos que classificou mais casos de não ocorrência (636) e teve sua origem a partir do nó 24, que avaliou a média da temperatura máxima no período de 20 dias antes da possível data de infecção, para valores acima de 27,15°C. Relacionando com outros ramos do modelo, há outros casos com temperaturas superiores a esse valor e ainda assim foram detectados casos de ocorrência, como por exemplo, a partir do nó 40. Porém, as condições geradas pelos outros atributos são diferentes, e, pelo ramo do nó raiz até o nó 31,

podem ter formado um conjunto de condições ambientais nas quais a temperatura máxima possa ser um pouco que mais baixa que em outras condições e mesmo assim provocar um ambiente desfavorável à ocorrência da doença.

Considerando toda a sub-árvore a partir do nó 3, foi possível verificar a maior quantidade de atributos relacionados com a precipitação. Foram 6 atributos relacionados com a precipitação e 4 com a temperatura, diferentemente da sub-árvore a partir do nó 2, onde foram utilizados 13 atributos de temperatura e 3 de precipitação. De forma geral, os testes sobre os atributos de precipitação relacionados na sub-árvore a partir do nó 3, tendem a estabelecer valores que indicam que, para *menores* quantidades de chuva, tanto em número de dias com precipitação acima de 1mm (nós 15, 20 e 25) quanto para o maior volume de chuva em um único dia (nós 14 e 21), há mais casos classificados como *ocorrência.*

Um dos efeitos prejudiciais da precipitação sobre a ocorrência da doença foi evidenciado por Dufault et al. 2010, onde a precipitação de intensidades de 45 ou 85mm.h^{-1}, em seguida (de 1 a 30 minutos) à deposição dos esporos causada pela chuva pode remover de 38 a 91% dos esporos depositados nas folhas. Porém, com as informações apresentadas pelos dados e pelo modelo não foi possível relacionar diretamente o conhecimento obtido no modelo com a literatura.

Em outros trabalhos, que consideram o desenvolvimento da ferrugem asiática, principalmente a severidade, a precipitação favorece o agravamento da doença (DEL PONTE et al., 2006; TCHANZ, 1984 apud DEL PONTE e ESKER, 2008). A proposta da modelagem deste trabalho buscou relacionar a presença da doença na propriedade do produtor, e não seu agravamento (desenvolvimento), o que pode não evidenciar as relações entre o aumento de chuvas e o aumento das ocorrências da doença.

De uma forma geral, para a ocorrência ou não ocorrência da doença, os atributos considerados mais importantes foram relacionados com a temperatura, utilizando como valores de decisão próximos às temperaturas consideradas críticas para a germinação de esporos e infecção das plantas encontradas na literatura.

Um fator a considerar são os erros envolvidos nos dados meteorológicos, principalmente relacionados com as estimativas de precipitação. Na Seção 4.1.2, foi verificado

que os dados de precipitação apresentaram correlações mais baixas entre estações virtuais e estações reais, quando comparado com dados de temperatura mínima ou máxima.

Há também uma dificuldade natural em se estimar a quantidade de chuva em um local específico, uma vez que a distribuição de chuva em uma área não é uniforme, podendo chover em uma determinada região e, em uma curta distância, não chover. Assim, ao utilizar dados de precipitação em uma escala de município, não é possível garantir que na região de ocorrência da doença tenha ocorrido chuva ou não, o que pode influenciar a detecção da ação da chuva sobre a ocorrência ou não da doença.

Os atributos relacionados com períodos mais longos, de 20 e 15 dias foram considerados mais importantes pelo modelo e foram mais utilizados (8 e 9 vezes, respectivamente), sendo que os períodos de 5 e 10 dias, e o período latente, foram utilizados poucas vezes (4, 2 e 4 vezes, respectivamente) para alguma decisão do modelo.

5 CONCLUSÕES

Concluiu-se ser possível gerar modelos por meio de indução de árvores de decisão para classificação de casos de ocorrência ou não da ferrugem asiática da soja, considerando variáveis derivadas da temperatura e da precipitação em períodos anteriores à ocorrência da doença.

O trabalho se constitui na primeira abordagem do uso de árvores de decisão para a modelagem de dados de epidemias de doenças, a partir de um banco de dados com milhares de registros de epidemias no campo, com o objetivo de identificar fatores que predispunham à ocorrência da doença, no caso variáveis meteorológicas.

O modelo preditivo apresentou uma taxa de acerto de 79,52%, com a estatística kappa de 0,59 e um total de 116 folhas do modelo, sendo considerado adequado para predizer eventos de ocorrência ou não ocorrência, considerando a modelagem adotada neste trabalho.

É importante considerar que o conjunto de dados possui diversas incertezas, como por exemplo, a data exata da detecção da doença bem como o seu nível de intensidade no momento da detecção, o que torna difícil se estimar o momento em que ocorreu a infecção.

Da mesma forma, os dados meteorológicos utilizados podem não ser representativos das condições microclimáticas que ocorreram em cada lavoura em que se relatou a ocorrência da doença. Por exemplo, a localização utilizada foi o centroide do município, o qual pode estar distante da lavoura onde a doença foi detectada. Além disso, os dados das estações virtuais possuem viés em relação ao dado observado, tal como foi verificado nesse estudo.

Outra limitação foi a falta de uma não ocorrência verdadeira, que seria uma lavoura monitorada onde não tivesse sido detectada a doença. Com isso, é possível que a maior importância das variáveis de temperatura seja devido ao fato de que as condições de precipitação não foram limitantes, já que o evento de não ocorrência precedeu ao de ocorrência. Nesse caso, a temperatura exerceu fator importante na duração do período latente ou mesmo da temperatura favorável para a infecção.

Este trabalho representa uma contribuição na utilização de um processo de mineração de dados para explorar fatores de influência nas ocorrências da ferrugem asiática da soja,

especificamente com a técnica de árvore de decisão, demostrando assim o potencial da técnica em futuros estudos na área.

A seguir são listadas algumas possibilidades de continuidade deste trabalho:

•Utilizar outros métodos e parâmetros na detecção de *outliers* e remoção de registros, como por exemplo, utilizar outros valores de número mínimo de objetos por folha no filtro *RemoveMissclassifeid* utilizado.

•Agregar ao conjunto de dados mais informações relacionadas aos fenômenos El Niño e La Niña., como por exemplo, um atributo que representa a presença do fenômeno na safra.

•Realizar uma preparação de dados que permita o uso do estádio de desenvolvimento da planta.

•Obter modelos específicos para cada região do Brasil.

•Utilizar diferentes métodos de seleção de atributos com a finalidade de se investigar modelos gerados a partir de conjuntos de atributos distintos. Com a redução de atributos que menos contribuem para os modelos, dois aspectos poderiam ser estudados: (a) comparação dos modelos sob as mesmas condições; e (b) análise dos modelos interpretativos mais concisos, o que supostamente seriam os modelos mais inteligíveis.

6 REFERÊNCIAS

AGRIOS, G. N. **Plant pathology**. 5ed. San Diego: Elsevier Academic Press, 2004.

ALVES, S. A. M.; FURTADO, G. Q.; BERGAMIN FILHO, A. Influência das condições climáticas sobre a ferrugem da soja. In: ZAMBOLIM, L. (Ed.). **Ferrugem Asiática da Soja**. Viçosa: Universidade Federal de Viçosa, Departamento de Fitopatologia, p.37-59. 2006.

APTÉ, C.; WEISS, S. Data mining with decision trees and decision rules. **Future Generation Computer Systems**, 13p, 1997.

BATCHELOR, W. D.; YANG, X. B.; TSCHANZ, A. T. Development of a neural network for soybean rust epidemics. **Transactions of the ASAE**, v. 40, n. 1, p. 247-252, 1997.

BERGAMIN FILHO, A. Epidemiologia comparativa: ferrugem da soja e outras doenças. In: ZAMBOLIM, L. (Ed.). **Ferrugem Asiática da Soja**. Viçosa: Universidade Federal de Viçosa, Departamento de Fitopatologia, 2006. p. 15-35.

BONDE, M. R.; BERNER, D. K.; NESTER, S. E.; FREDERICK, R. D. Effects of temperature on urediniospore germination, germ tube growth, and initiation of infection in soybean by phakopsora isolates. **Phytopathology**, v. 97, n. 8, p. 997-1003, ago 2007.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. Boca Raton: CRC Press, 1984. 358 p.

CANTERI, M. G.; DEL PONTE, E. M.; GODOY, C. V.; TSUKAHARA, R. Y. Emprego da tecnologia da informação para simulação de epidemias e zoneamento agroclimático aplicáveis no controle de doenças de plantas. **Summa Phytopathologica**, v. 33, n. supl., p. 121-124, 2007.

CANTERI, M.G.; CARAMORI, P.; TSUKAHARA, R.; SILVA, O.C.; FARIA, R.; GODOY, C.V. A system to map risk of infection by Phakopsora pachyrhizi for Parana State, Brazil. **Phytopathology** 95:S16. 2005 (Abstract)

CANTERI, M.G.; GODOY, C.V.; DEL PONTE, E.M.; FERNANDES, J.M.C.; PAVAN, W. Aplicações da computação na fitopatologia. **Revisão Anual de Patologia de Plantas**, 12:243-285. 2004.

CASTRO, M. Incidência de ferrugem na lavoura de soja é baixa. **Correio de Uberlândia,** 05 de abril 2011. Disponível em: <http://www.correiodeuberlandia.com.br/cidade-e-regiao/incidencia-de-ferrugem-na-lavoura-de-soja-e-baixa/ >. Acesso em: 23/02/2012.

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

COAKLEY, S. M. Variation in climate and prediction of disease in plants. **Annual Review of Phytopathology**, v. 26, n. 1, p. 163-181, 1988.

COAKLEY, S. M.; LINE, R. F.; MCDANIEL, L. R. Predicting stripe rust severity on winter wheat using an improved method for analyzing meteorological and rust data. **Phytopathology**, v. 78, p. 543-550, 1988.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**. v. 20, 1960. p.37-46.

COLLISCHONN, B; COLLISCHONN, W; TUCCI, C. Análise do campo de precipitação gerado pelo satélite TRMM sobre a bacia do São Francisco até Três Marias. In: SIMPÓSIO DE RECURSOS HÍDRICOS DO SUL SUDESTE, 1, 2006. 27- 29 de agosto, Curitiba, PR. **Anais...**, Curitiba: ABRH, 2006.

COLLISCHONN, B.; ALLASIA, D.; COLLISCHONN, W.; TUCCI, C. E. M. Desempenho do satélite TRMM na estimativa de precipitação sobre a bacia do Paraguai superior. **Revista Brasileira de Cartografia**, v. 59, n. 1, p. 93-99, 2007.

CONSÓRCIO ANTIFERRUGEM. **Informativo de risco 2009/2010 n.1**. 2010a. Disponível em: <http://www.consorcioantiferrugem.net>. Acesso em 27/02/2012.

CONSÓRCIO ANTIFERRUGEM. **Informativo de risco 2010/2011 n.1**.2010b. Disponível em: <http://www.consorcioantiferrugem.net>. Acesso em 27/02/2012.

CPTEC. Centro de Previsão de Tempo e Estudos Climáticos. **Informações sobre El Niño**. Disponível em: <http://enos.cptec.inpe.br/>. Acesso em: 16/02/2012.

DALL'AGNOL, A.; LAZAROTTO, J. J.; HIRAKURI, M. H. **Desenvolvimento, mercado e rentabilidade da soja brasileira**. Londrina: Embrapa Soja. 2010.

DEL PONTE, E. M.; ESKER, P. D. Meteorological factors and asian soybean rust epidemics: a systems approach and implications for risk assessment. **Scientia Agricola**, v. 65, n. spe, p. 88-97, dez 2008.

DEL PONTE, E. M.; GODOY, C. V.; CANTERI, M. G.; REIS, E. M.; YANG, X. B. Models and applications for risk assessment and prediction of Asian soybean rust epidemics. **Fitopatologia Brasileira**, v. 31, n. 6, p. 533-544, 2006a.

DEL PONTE, E. M.; GODOY, C. V.; LI, X.; YANG, X. B. Predicting severity of asian soybean rust epidemics with empirical rainfall models. **Phytopathology**, v. 96, n. 7, p. 797-803, 2006b.

DEL PONTE, E. M.; MAIA, A. D. H. N.; DOS SANTOS, T. V. ; MARTINS, E. J.; BAETHGEN, W. E. Early-season warning of soybean rust regional epidemics using El Niño Southern/Oscillation information. **International journal of biometeorology**, v. 55, n. 4, p. 575-83. 2011.

DEL PONTE, E. M.; MARTINS, E. J. Decifrando a complexa equação de risco da ferrugem da soja no Brasil. **Revista Plantio Direto**, p. 12-16, 2008.

DUFAULT, N. S.; ISARD, S. A.; MAROIS, J. J.; WRIGHT, D. L. Removal of wet deposited Phakopsora pachyrhizi urediniospores from soybean leaves by subsequent rainfall. **Plant Disease**, v. 94, n. 11, p. 1336-1340, 2010.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Tecnologias de Produção de Soja -** Região Central do Brasil 2011.Londrina: Embrapa Soja. 2010. (Sistemas de Produção, 14)

EVANGELISTA, S. R. M.; TERNES S.; SANTOS, E. H. dos; ASSAD, E. D.; ROMANI, L. A. S.; FRANZONI, A. Agroclima: sistema de monitoramento agroclimatológico. In: CONGRESSO BRASILEIRO DE AGROMETEOROLOGIA, 13., 2003, Santa Maria. Situação atual e perspectivas da agrometeorologia: **Anais...** Santa Maria: Unifra, SBA, UFSM, 2003, v. 1, p. 603-604.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, p. 37-54, 1996.

FEHR, W. R.; CAVINESS, C. E. Stages of soybean development. **Ames: Iowa State University of science and techonology**, Special Report 80, p. 11, 1977.

GLEASON, M. L.; DUTTWEILER, K. B.; BATZER, J. C. et al. Obtaining weather data for input to crop disease-warning systems: leaf wetness duration as a case study. **Scientia Agricola**, v. 65, n. spe, p. 76-87, dez 2008.

GODOY, C. V.; FLAUSINO, A. M.; SANTOS, L. C. M.; DEL PONTE, E. M. Eficiência do controle da ferrugem asiática da soja em função do momento de aplicação sob condições de epidemia em Londrina, PR. **Tropical Plant Pathology**, v. 34, n. 1, p. 056-061, 2009.

GOELLNER, K.; LOEHRER, M.; LANGENBACH, C.; CONRATH, U. W. E.; KOCH, E.; SCHAFFRATH, U. *Phakopsora pachyrhizi*: the causal agent of asian soybean rust. **Molecular Plant Pathology**, v. 11, p. 169-177, 2010.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v. 11, n. 1. 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

HARDWICK, N. V. Disease Forecasting. In: COOKE, B. M.; JONES, D. G.; KAYE, B. (Ed.). **The epidemiology of plant diseases**. AA Dordrecht, The Netherlands: Springer. p. 239-264. 2006.

HENNING, A. A.; GODOY, C. V. Situação da ferrugem da soja no Brasil e no mundo. In: ZAMBOLIM, L. (Ed.). **Ferrugem Asiática da Soja**. Viçosa: Universidade Federal de Viçosa, Departamento de Fitopatologia, p. 1-14. 2006.

INMET. Instituto Nacional de Meteorologia. Disponível em <http://www.inmet.gov.br>. Acesso em 21/08/2012.

ISARD, S. A.; GAGE, S. H.; COMTOIS, P.; RUSSO, J. M. Principles of the atmospheric pathway for invasive species applied to soybean rust. **BioScience**, v. 55, n. 10, p. 851-861, 2005.

ISARD, S. A.; RUSSO, J. M.; ARIATTI, A. The integrated aerobiology modeling system applied to the spread of soybean rust into the Ohio River valley during September 2006. **Aerobiologia**, v. 23, n. 4, p. 271-282, out 2007.

KIM, K. S.; WANG, T. C.; YANG, X. B. Simulation of apparent infection rate to predict severity of soybean rust using a fuzzy logic system. **Phytopathology**, v. 95, n. 10, p. 1122-31, 2005.

KOCHMAN, J. K. The effect of temperature on development of soybean rust (Phakopsora pachyrhizi). **Australian Journal of Agricultural Research**, n. 30, p. 273-277, 1979.

KUMMEROW, C.; SIMPSON, J.; THIELE, O. et al.. The status of the tropical rainfall measuring mission (TRMM) after two years in orbit. **Journal of Climate**, p. 1965-1982, 2000.

LANDIS, J.R.; KOCH, G.G. The measurement of observer agreement for categorical data. **Biometrics**. Michigan, v. 33, n. 1, p.159-174, 1977.

MADDEN, L. V.; HUGHES, G.; VAN DEN BOSCH, F. **The study of plant disease epidemics**. St Paul, MN: American Phytopathological Society. 2007.

MARCHETTI, M. A.; MELCHING, J. S.; BROMFIELD, K. R. The effects of temperature and dew period on germination and infection by uredospores of Phakopsora pachyrhizi. **Phytopathology**, n. 66, p. 461-463, 1976.

MARCHETTI, M. A.; UECKER, F. A.; BROMFIELD, K. R. Uredial development of Phakopsora pachyrhizi in soybeans. **Phytopathology**, p. 822-823, 1975.

MARQUES, L. C. Ferrugem reduz 92%. **Gazeta Digital**, 20 de março de 2011. Disponível em: <http://www.gazetamt.com/conteudo/show/secao/2/materia/268654>. Acesso em: 22/11/2011.

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. **Tropical Plant Pathology**, v. 33, n. 2, p. 114-124, 2008.

MEIRA, C. A. A.; RODRIGUES, L. H. A.; MORAES, S. A. D. Modelos de alerta para o controle da ferrugem-do-cafeeiro em lavouras com alta carga pendente. **Pesquisa Agropecuária Brasileira**, v. 44, n. 3, p. 233-242, mar 2009.

MELCHING, J. S.; DOWLER, W. M.; KOOGLE, D. L.; ROYER, M. H. Effects on duration, frequency, and temperature of leaf wetness periods on soybean rust. **Plant Disease**, v. 73, p. 117-122, 1989.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002a. p. 89-114.

MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002b. p. 115-139.

NEWTON, A.; MCROBERTS, N.; HUGHES , G. Information technology in plant disease epidemiology. In: COOKE, B. M.; JONES, D. G.; KAYE, B. (Ed.). **The epidemiology of plant diseases**. AA Dordrecht, The Netherlands: Springer. p. 335-356. 2006.

NICHOLSON, S. On the question of the "recovery" of the rains in the West African Sahel. **Journal of Arid Environments**, v. 63, p. 615-641, 2005.

NOAA. *National Oceanic and Atmospheric Administration.* **NOAA ARL HYSPLIT Model**. Disponível em: <http://www.arl.noaa.gov/ready/hysplit4.html>. Acesso em: 22/11/2011.

OTAVIAN, A. F. **Agritempo**. Campinas: Embrapa Informática Agropecuária (CNPTIA), 2011. (Comunicação oral).

PAN, Z.; YANG, X. B.; PIVONIA, S.; XUE, L.; PASKEN, R.; ROADS, J. Long-term prediction of soybean rust entry into the continental United States. **Plant Disease**, v. 90, n. 7, p. 840-846, 2006.

PIVONIA, S.; YANG, X. B. Relating epidemic progress from a general disease model to seasonal appearance time of rusts in the United States: implications for soybean rust. **Phytopathology**, v. 96, n. 4, p. 400-407, 2006.

QUINLAN, J. R. **C4.5**: Programs for machine learning. San Francisco: Morgan Kaufmann, 1993.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, n. 1, p. 81-106, out 1986.

REIS, E.M.; SARTORI, A.F.; CAMARA, R.K. Modelo climático para a previsão da ferrugem da soja. **Summa Phytopathologica**. v. 30, p. 290-292. 2004.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. de. Mineração de dados. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002. p. 307-335.

ROMANI, A. S.; EVANGELISTA, S. R. M.; SANTOS, E. H.; TERNES, S.; MONTAGNERS, A. J. Organização do banco de dados meteorológicos do Sistema Agritempo. In: IV Congresso Brasileiro Da Sociedade Brasileira De Informática Aplicada À Agropecuária e à Agroindustria. **Anais...** Porto Seguro, 2003a.

ROMANI, L. A. S.; OTAVIAN, A. F.; EVANGELISTA, S. R. M.; ASSAD, E. D. Modelo de estações virtuais com estimativa de precipitação e temperatura para aprimoramento dos mapas no Agritempo. XV Congresso Brasileiro de Agrometeorologia. **Anais**... Aracaju - SE , 2007.

ROMANI, L. A. S.; SANTOS, E. H. DOS; EVANGELISTA, S. R. M.; ASSAD, E. D.; PINTO, H. S. Utilização de estações vizinhas para estimativa de temperatura e precipitação usando o inverso do quadrado da distância. XIII Congresso Brasileiro de Agrometeorologia. **Anais...** p.717-718. Santa Maria – RS, 2003b.

ROZANTE, J. R.; MOREIRA, D. S.; GONCALVES, L. G. G. DE; VILA, D. A. Combining TRMM and surface observations of precipitation: Technique and validation over south america. **Weather and Forecasting**, v. 25, n. 3, p. 885-894, jun 2010.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining**: Mineração de dados. Rio de Janeiro: Ciência Moderna. 932p. 2009.

TRMM. ***Tropical Rainfall Measuring Mission***. Disponível em: <http://trmm.gsfc.nasa.gov>. Acesso em: 27/02/2012.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3ed. San Francisco: Morgan Kaufmann, 2011.

YANG, X. B.; DOWLER, W. M.; TSCHANZ, A. T. A simulation model for assessing soybean rust epidemics. **Journal of Phytopathology**, v. 133, n. 3, p. 187-200, nov 1991.

YORINORI, J. T.; PAIVA, W. M.; FREDERICK, R. D.; COSTAMILAN, L. M.; BERTAGNOLLI, P. F; HARTMAN, G. E.; GODOY, C. V.; NUNES JR., J. Epidemics of Soybean Rust (Phakopsora pachyrhizi) in Brazil and Paraguay from 2001 to 2003. **Plant Disease**, v. 89, n. 6, p. 675-677, jun 2005.

ZAMBENEDETTI E. B.; ALVES, E, POZZA, E.A .; ARAÚJO, D.V.; GODOY, C. V. Avaliação de parâmetros monocíclicos e da intensidade da ferrugem asiática (Phakopsora pachyrhizi) em diferentes genótipos de soja e posições de copa. **Summa Phytopathologica**, v.33, n.2, p.178-181, 2007.