UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS AGRÁRIAS E ENGENHARIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS FLORESTAIS

RONIE SILVA JUVANHOL

APRENDIZADO DE MÁQUINA NA MODELAGEM DE INCÊNDIOS FLORESTAIS NO ESTADO DO ESPÍRITO SANTO

JERÔNIMO MONTEIRO - ES
2017

RONIE SILVA JUVANHOL

APRENDIZADO DE MÁQUINA NA MODELAGEM DE INCÊNDIOS FLORESTAIS NO ESTADO DO ESPÍRITO SANTO

Tese apresentada ao Programa de Pós-Graduação em Ciências Florestais do Centro de Ciências Agrárias e Engenharias da Universidade Federal do Espírito Santo, como parte das exigências para obtenção do Título de Doutor em Ciências Florestais na Área de Concentração Ciências Florestais.
Orientador: Prof. Dr. Nilton Cesar Fiedler
Coorientador: Prof. Dr. Alexandre Rosa dos Santos

JERÔNIMO MONTEIRO - ES

2017

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial de Ciências Agrárias, Universidade Federal do Espírito Santo, ES, Brasil)
Bibliotecário: Felicio Gomes Corteletti – CRB-6 ES-000646/O

Juvanhol, Ronie Silva, 1988-

J97a Aprendizado de máquina na modelagem de incêncios florestais no estado do Espírito Santo / Ronie Silva Juvanhol. – 2017.

115 f. : il.

Orientador: Alexandre Rosa dos Santos.

Coorientador: Nilton Cesar Fiedler.

Tese (Doutorado em Ciências Florestais) – Universidade Federal do Espírito Santo, Centro de Ciências Agrárias e Engenharias.

1. Florestas - Proteção. 2. Incêndios florestais. 3. Inteligência computacional. 4. Algoritmos. I. Santos, Alexandre Rosa dos. II. Fiedler, Nilton Cesar Gilson. III. Universidade Federal do Espírito Santo. Centro de Ciências Agrárias e Engenharias. IV. Título.

CDU: 630

APRENDIZADO DE MÁQUINA NA MODELAGEM DE INCÊNDIOS FLORESTAIS NO ESTADO DO ESPÍRITO SANTO

Ronie Silva Juvanhol

Tese apresentada ao Programa de Pós-Graduação em Ciências Florestais do Centro de Ciências Agrárias e Engenharias da Universidade Federal do Espírito Santo, como parte das exigências para obtenção do Título de Doutor em Ciências Florestais na Área de Concentração Ciências Florestais.

Aprovada em 19 de Dezembro de 2017.

Prof. Dr. Wellington Betencurte da Silva (Examinador externo)
Universidade Federal do Espírito Santos

Profª. Drª. Telma Machado de Oliveira Peluzio (Examinadora externa)
Instituto Federal do Espírito Santo/Alegre

Prof. Dr. Christiano Jorge Gomes Pinheiro (Examinador externo)
Universidade Federal do Espírito Santos

Prof. Dr. Alexandre Rosa dos Santos (Coorientador)
Universidade Federal do Espírito Santo

Prof. Dr. Nilton Cesar Fiedler (Orientador)
Universidade Federal do Espírito Santo

AGRADECIMENTOS

Primeiramente a Deus, pelo dom da vida, e pela graça concedida, de todos os dias encontrar motivos para agradecer e amá-lo em todas as circunstâncias.

Meus sinceros agradecimentos ao meu orientador, professor Dr. Nilton Cesar Fiedler e coorientador, professor Dr. Alexandre Rosa dos Santos. Seus ensinamentos, que transpõem às salas de aula, me acompanharão pela minha vida. Agradeço pelos exemplos de vida, dedicação em orientar e pelo apoio durante esta jornada de desafio, construção e amadurecimento.

Aos meus pais, pelo carinho incondicional e pela perseverança em amar, sempre. Palavras não mensuram meus agradecimentos aos senhores.

À minha irmã, pelos incentivos, pela ajuda financeira e por acreditar neste projeto de vida. Aos entes queridos, por estarem sempre próximos de coração, mesmo quando os impedimentos tornam raros os encontros e os abraços.

À Universidade Federal do Espírito Santo, ao Programa de Pós-Graduação em Ciências Florestais, em especial a todos os professores do Departamento de Ciências Florestais e da Madeira da UFES, por tanto contribuir em minha formação acadêmica e profissional. ”Ensinar não é transferir conhecimento, mas criar possibilidades para a sua própria produção ou a sua construção” (Paulo Freire).

Ao professor Dr. Christiano Jorge Gomes Pinheiro, por tão bem representar meus mentores, no momento de suas ausências durante o Pós-Doutorado. Agradeço por sua amizade, confiança e valores transmitidos durante nossas conversas.

Aos professores Dr. Wellington Betencurte da Silva e Dra. Telma Machado de Oliveira Peluzio, por terem aceitado a participar deste momento importante de minha trajetória acadêmica, compondo a banca avaliadora.

À FAPES pela concessão da bolsa de doutorado, tão importante para a realização deste projeto de vida. Aos órgãos e agências governamentais por cederem alguns dos dados utilizados neste estudo.

Ao colega Thiago Tuler pela fundamental ajuda na implementação das regras da árvore de decisão, geradas neste estudo.

Aos amigos de longa data e àqueles feitos durante a vivência acadêmica, em especial, aos amigos do Laboratório de Incêndios, LABCELF, LaMFlor e GAGEN. Obrigado pela convivência e pelos bons momentos durante toda essa jornada.

“Porque dele, e por meio dele, e para ele são todas as coisas.
A ele, pois, a glória eternamente. Amém!”
(Romanos 11:36)

RESUMO

JUVANHOL, Ronie Silva. **APRENDIZADO DE MÁQUINA NA MODELAGEM DE INCÊNDIOS FLORESTAIS NO ESTADO DO ESPÍRITO SANTO.** 2017. Tese (Doutorado em Ciências Florestais) – Universidade Federal do Espírito Santo, Jerônimo Monteiro, ES. Orientador: Prof. Dr. Nilton Cesar Fiedler. Coorientador: Prof. Dr. Alexandre Rosa dos Santos.

O principal problema encontrado quando da aplicação de técnicas de sistemas de informações geográficas e sensoriamento remoto para a predição de incêndios florestais é a necessidade de integrar diferentes fontes de dados. Os métodos aplicados, geralmente são baseados em técnicas de regressão ou em coeficientes que dependem de conhecimentos dos especialistas. Esta pesquisa objetivou testar a capacidade da árvore de classificação e regressão (CART) em avaliar o risco de incêndios florestais no estado do Espírito Santo. A análise CART é uma técnica estatística não paramétrica que gera regras de decisão na forma de uma árvore binária, para um processo de classificação ou de regressão. O produto MCD45A1 de área de queima, relativo a um período de 16 anos (2000-2015), foi utilizado para, a partir dos pontos centrais da célula de grade, obter um mapa de ocorrência de incêndio por meio de uma abordagem de densidade *Kernel*. O mapa resultante foi então utilizado como variável de entrada para a análise CART com variáveis de influência de incêndios usados como preditores. Um total de 12 variáveis preditoras foram determinadas de diversas bases de dados, abrangendo aspectos ambientais, físicos e socioeconômicos. As regras induzidas pelo processo de regressão permitiram a definição de diferentes níveis de risco, expressa em 35 unidades de gestão, utilizado para a produção de um mapa de predição de fogo. De acordo com os resultados, as áreas de maiores riscos no estado são representadas pela Região Nordeste, Vale do Rio Doce e Sudeste (Costa Sul). Os resultados do processo de regressão ($r=0,94$ e $r^2=0,88$), a capacidade de análise do algoritmo CART para destacar as relações hierárquicas entre as variáveis preditoras e a interpretabilidade fácil das regras de decisão, representam uma ferramenta possível para melhor abordar o problema de avaliar e representar o risco de incêndios florestais.

Palavras-chave: Estatística não paramétrica, Densidade *Kernel*, Algoritmo CART, Regras de decisão, Mapa de predição do fogo.

ABSTRACT

JUVANHOL, Ronie Silva. **MACHINE LEARNING IN THE MODELING OF FOREST FIRES IN THE STATE OF ESPÍRITO SANTO.** 2017. Thesis (Doctor of Forest Science) – Federal University of Espírito Santo, Jerônimo Monteiro, ES. Advisor: Prof. Dr. Nilton Cesar Fiedler. Co Advisor: Prof. Dr. Alexandre Rosa dos Santos.

The main problem encountered when applying geographic information systems and remote sensing techniques for the prediction of forest fires is the necessity to integrate different data sources. The methods applied are usually based on regression techniques or on coefficients that depend on expert knowledge. The objective of this study was to test the capacity of the classification and regression tree (CART) to assess the regional fire risk. The CART analysis is a non-parametric statistical technique that generates decision rules in the form of a binary tree, for a classification or regression process. The MCD45A1 product of burn area, relative to 16-year (2000-2015) was used to obtain a fire occurrence map, from the center points of the grid cell, using a kernel density approach. The resulting map was then used as input response variable for the CART analysis with fire influence variables used as predictors. A total of 12 predictors were determined from several databases, covering environmental physical and socioeconomic aspects. The rules induced by the regression process allowed the definition of different risk levels, expressed in 35 management units, used to produce a fire prediction map. According to the results, the Northeast Region, sweet river and Southeast represent the major risk areas in the state (South Coast). The results of the regression process ($r = 0.94$ and $r^2 = 0.88$), the capability of the CART algorithm analysis to highlight the hierarchical relationships between the predictor variables and the easy interpretability of the decision rules represent a possible tool to better approaching the problem of assessing and representing forest fires.

Keywords: Non-parametric statistics, Kernel density, CART algorithm, Decision rules, Fire prediction map.

SUMÁRIO

1. INTRODUÇÃO

Ao longo das últimas décadas, os incêndios florestais no Brasil têm recebido maior atenção devido à grande variedade de impactos ecológicos, econômicos, sociais e políticos; embora estatísticas de ocorrência e efeitos do fogo ainda sejam incientes.

O fogo desempenha um importante papel na criação e manutenção da estrutura da paisagem, composição, função e integridade ecológica, podendo influenciar as taxas e os processos de sucessão ecológica (COVINGTON; MOORE, 1994; MORGAN et al., 2001).

O impacto dos incêndios em escalas locais, regionais e globais foi revisado em Stolle e Lambin (2003) e Lentile et al. (2006). Em escala local, o fogo pode estimular processos microbianos do solo e a combustão da vegetação, alterando a estrutura e a composição de solos e vegetação (LENTILE et al., 2006). Em escalas regionais e globais, a combustão da floresta e vegetação de pastagem liberam grandes volumes de gases radiativamente ativos, aerossóis pirogênicos e outros compostos que significativamente influenciam o orçamento radiativo da terra e a química atmosférica (ANDREAE; MERLET, 2001), afetando a qualidade do ar (HARDY et al., 2001) e aumentando a preocupação sobre os riscos para a saúde humana (STEFANIDOU et al., 2008).

Compreender a distribuição espacial dos incêndios florestais e conhecer os principais fatores para sua ocorrência baseia-se principalmente na análise de locais históricos de ocorrências (SYPHARD et al., 2008). A importância dos fatores antropogênicos na regulação dos eventos de incêndios, além do clima, vegetação e fatores topográficos faz a predição do fogo altamente desafiadora (PERRY, 1998). O desenvolvimento e utilização de modelos de predição de fogo pode auxiliar o manejo florestal na tomada de decisão ativa e preventiva (GONZÁLEZ et al., 2006).

Uma grande variedade de técnicas tem sido utilizadas para o modelo de risco de incêndio. Os modelos mais complexos de fogo requerem informação espacial que é fornecida por sensoriamento remoto e Sistemas de Informação Geográfica (SIG) (BONAZOUNTAS et al., 2005). O uso de dados de várias fontes, implica a presença de variação local e relações multivariadas entre as variáveis preditoras, e requer modelos flexíveis, imparciais, objetivos e consistentes. Apesar disso, os

modelos comumente utilizados sugerem uma configuração a priori dos parâmetros de modelagem. Essa configuração preliminar é muitas vezes baseada no conhecimento dos especialistas em incêndios (CAETANO et al., 2004; CHUVIECO; CONGALTON, 1989; EUGENIO et al., 2016; SEMERARO et al., 2016; VADREVU et al., 2010) ou na implementação da análise de regressão, onde os coeficientes representam os pesos das variáveis preditoras consideradas.

Os modelos estatísticos propostos em trabalhos anteriores para implementar essa análise variam desde regressão linear múltipla (CHUVIECO; CONGALTON, 1989) e regressão logística (RL) (CATRY et al., 2009; CHOU et al., 1993; VASCONCELOS et al., 2001; PREISLER et al., 2004). A regressão linear múltipla lida com uma variável de resposta contínua, enquanto a RL admite a existência de variáveis binárias. Além disso, a RL não precisa de hipóteses sobre a distribuição dos dados, e as variáveis preditoras podem ser contínuas e/ou categóricas. Entretanto, a RL dá uma saída como uma simples resposta "sim / não". Tal limitação, não permite a sua utilização para a predição de variáveis contínuas como a densidade de fogo. Ao contrário, a regressão linear múltipla pode lidar com estas questões, mas a sua aplicação exclui variáveis categóricas, tais como o uso da terra e tipo de combustível, muito essencial na modelagem de risco de incêndio.

A fim de superar estas desvantagens, alguns autores utilizaram redes neurais para prever o risco de incêndio, com resultados satisfatórios, especialmente em áreas onde a ocorrência de incêndios revela um comportamento muito complexo e heterogêneo (BART, 1998; CHUVIECO, 1999; VASCONCELOS, 2001; VEGA-GARCIA, 1996). No entanto, as redes neurais apresentam algumas desvantagens, tal como a computação lenta durante as fases de treinamento. Na prática, a rede neural calcula mapas de risco de incêndio precisos, mas não fornece informações sobre as regras de hierarquia e de regressão das variáveis (AMATULLI et al., 2006).

Como apresentado, diferentes técnicas foram testadas e desenvolvidas, mas cada uma com suas lacunas e desvantagens. Muitas vezes, os coeficientes de regressão obtidos por técnicas estatísticas comuns, são aplicados a toda a área de estudo, sem considerar a importância da variação espacial de cada variável no processo de regressão. Esta suposição, também conhecido como estacionaridade, é frequentemente violado em situações do mundo real (MARTÍNEZ-FERNÁNDEZ et al., 2013). Isto significa que as técnicas tradicionais comumente usadas para a

integração e avaliação das relações entre as variáveis preditoras não oferecem uma visão realista dos incêndios e, portanto, não são capazes de oferecer uma ajuda prática aos gestores para a tarefa de sistema de apoio à decisão (DSS).

Os gestores de incêndio enfrentam dificuldades crescentes ao planejar as atividades de gestão. Portanto, antes do início da temporada do fogo, são necessárias ferramentas operacionais mais eficientes do que as propostas até o momento, a fim de fornecer informações úteis para a fase de planejamento prévio. (SAN-MIGUEL et al., 2003). Os modelos aplicáveis devem ser capazes não só de lidar com diferentes fontes de dados ou de delinear relacionamentos não estacionários, mas também de fornecer um resultado mais detalhado do que um único mapa de risco de saída.

Contra às lacunas mencionadas dos sistemas atuais de predição de incêndio, os algoritmos de aprendizado de máquina apresentam uma abordagem interessante para tratar o problema. Alguns algoritmos são encontrados na literatura para avaliar a capacidade preditiva do mapa de fogo, tais como *Random Forest* (ARPACI et al., 2014; OLIVEIRA et al., 2012; WU et al., 2014), *MaxEnt* (ARPACI et al., 2014) e *Boosted Regression Trees* (ARGAÑARAZ et al., 2015). Entretanto, tais abordagens não possui o atrativo da interpretabilidade proporcionada pela estrutura da árvore oferecida pelo algoritmo *Classification and Regression Trees* (CART).

Proposto por Breiman et al. (1984), a árvore de decisão CART é capaz de processar atributos contínuos e nominais como alvos e preditores por meio de um procedimento recursivo binário para construir uma árvore ideal. Considerado como um dos dez melhores algoritmos de mineração de dados segundo Wu e Kumar, (2009), a árvore de classificação prevê probabilidades de associação para variáveis de resposta categóricas, enquanto a árvore de regressão prevê valores médios para as variáveis de resposta em escala intervalar ou de razão (MICHAELSEN et al., 1994), útil para a predição do fogo.

Esta tese se insere no âmbito da modelagem e análise espacial pelo uso de técnicas de aprendizado de máquina para avaliar a predição de fogo em escala regional; um campo em que o potencial desta abordagem estatística ainda não foi bem explorado. A técnica proposta visa fornecer saídas compreensíveis, na forma de regras de decisão, capazes de predizer os valores de risco médio para cada

célula de grade, definindo assim as unidades de manejo do fogo no estado do Espírito Santo.

2. REVISÃO DE LITERATURA

Incêndio florestal é um fenômeno global, com relevante pesquisa mundial ligada aos impactos sobre a vida humana, os ecossistemas e outros recursos. A gestão do fogo está sujeita a fontes múltiplas de incertezas. A incerteza decorre de dados imprecisos ou ausentes, da compreensão científica incompleta da resposta ecológica ao fogo, compreensão científica incompleta da resposta do comportamento do fogo para intervenções de gestão (supressão, redução de combustível, entre outros), e medidas de valor de recursos limitados para orientar instrumentos de prevenção de recursos em risco. A gestão estratégica está sujeita a significativa incerteza adicional ao considerar a dinâmica espaço-temporal das mudanças climáticas, a sucessão vegetativa, a migração de espécies, e regimes de perturbação.

As avaliações do risco de incêndios florestais são ferramentas de apoio à decisão que integram informações referentes à probabilidade e magnitude da resposta do recurso a fatores de risco, a fim de sintetizar uma conclusão sobre o risco que pode informar a tomada de decisão (SIKDER et al., 2006). As avaliações de risco informam a tomada de decisão estratégica, tática e operacional, e são usadas em modelos de apoio à decisão desenvolvida por pesquisadores e especialistas em gestão (BAR-MASSADA et al., 2009).

Cada vez mais, a gestão do incêndio florestal está sendo vista como uma forma de gestão de risco, com um correspondente aumento no rigor analítico e alinhamento com os princípios de gestão de risco. O desenvolvimento e utilização de modelos de avaliação de risco de incêndio pode auxiliar o manejo florestal na tomada de decisão ativa e preventiva (GONZÁLEZ et al., 2006). A literatura continua a expandir-se com exemplos de análises de risco de incêndio em todo o mundo (ATKINSON, 2010; DLAMINI, 2010; KALOUDIS et al., 2010; ZHIJUN et al., 2009). As escalas de planejamento para a gama de gestão de um incêndio incluem a prevenção de incêndios, detecção de incêndio, expedição de ataque inicial, e planejamento estratégico e gestão de combustível (MARTELL, 2007). O gerenciamento do risco de incêndio envolve a análise de exposição e efeitos (probabilidade e magnitude de potenciais efeitos benéficos ou nocivos), e o desenvolvimento de respostas de gestão apropriadas para reduzir a exposição e/ou mitigar os efeitos adversos (FAIRBROTHER; TURNLEY, 2005; FINNEY, 2005).

O documento dissertativo a seguir revisa o estado de avaliação da gestão, com uma abordagem sobre as incertezas em avaliações de risco integrados que consideram um conjunto de valores humanos e ecológicos. Dar-se início ao descrever a tipologia das incertezas enfrentadas em recursos naturais e de tomada de decisão ambiental. Uma análise é realizada dos métodos de modelagem de ocorrência de incêndios florestais, incluindo aspectos sobre o banco de dados e os fatores que afetam a ocorrência de fogo. Também são descritos os fundamentos da árvore de decisão e do algoritmo CART, considerado nesta tese como técnica para predição de ocorrência de incêndios florestais.

2.1. Caracterização e incerteza na gestão

Ascough et al. (2008), sintetiza várias caracterizações existentes de incerteza no contexto de tomada de decisão ambiental. Quatro grandes categorias de incerteza existem: (i) incerteza linguística, (ii) incerteza de variabilidade, (iii) incerteza do conhecimento e (iv) incerteza de decisão.

(i) A incerteza linguística refere-se a questões de indefinição, ambiguidade, dependência contextual de palavras e dificuldade em explicar resultados. Brugnach et al. (2011) descrevem a ambiguidade como uma fonte de incerteza em que podem existir mais de uma maneira válida de compreender o sistema a ser gerenciado. No contexto de incêndio florestal, o termo "risco" tem sido usado de modo indefinido, causando certa incompreensão. Hardy (2005), por exemplo, define o risco como a probabilidade de ocorrência de incêndios, enquanto Finney (2005) adota uma abordagem atual, definindo risco como a expectativa probabilística da alteração do valor de recursos em resposta ao fogo.

(ii) A incerteza de variabilidade refere-se à variabilidade inerente que se manifesta em sistemas naturais. A frequência e padrão espacial de locais de ignição, ou as condições meteorológicas que influenciam no comportamento do fogo são exemplos de incerteza de variabilidade. As abordagens probabilísticas são mais frequentemente utilizadas para lidar com este tipo de incerteza, tais como a modelagem de ocorrência de incêndio, propagação, e intensidade (AGER et al., 2010; CARMEL et al., 2009; KROUGLY et al., 2009; PODUR; WOTTON, 2010). As abordagens na ciência de gestão que incorporam elementos probabilísticos

incluem programação dinâmica estocástica, análise de cenários e modelos de cadeia de Markov (WEINTRAUB; ROMERO, 2006).

(iii) A incerteza de conhecimento refere-se aos limites de nosso conhecimento e limites da nossa compreensão científica. Este tipo de incerteza está presente, no que diz respeito, à forma como conceituar os processos naturais que ocorrem ao nosso redor, a natureza e a qualidade dos dados que usamos para informar esses modelos e a incerteza propagada nos resultados do modelo. Jones et al. (2004), por exemplo, identificam a necessidade de gerar estimativas de incerteza para as camadas de dados espaciais ao mapear combustíveis, e Bachmann e Allgöwer (2002) descrevem a propagação de incerteza na modelagem do comportamento do fogo devido à incerteza em relação às variáveis de entrada. Mais recentemente, Cruz e Alexander (2010) destacam as lacunas de conhecimento e erros comuns a muitos sistemas de modelagem de fogo em relação ao comportamento do fogo. A incerteza de conhecimento é considerada reduzida, na medida em que podemos reduzir o escopo dessa incerteza por meio de pesquisa adicional e investigação empírica.

Em contraste com a incerteza de variabilidade, as abordagens não-probabilísticas são, em geral, mais adequadas para gerir a incerteza do conhecimento (KANGAS; KANGAS, 2004). A gestão na incerteza do conhecimento, normalmente, envolve alguma forma de sistema pericial. Em problemas de gestão de recursos complexos, tais como a gestão de um incêndio, um reconhecimento formal de incerteza emparelhado com o julgamento especialista é muitas vezes a melhor abordagem (BORCHERS, 2005). O conhecimento especializado e o julgamento podem ser incorporados em uma série de maneiras, incluindo sistemas baseados em conhecimento, modelos hierárquicos multiatributo, modelos lógicos, teoria dos conjuntos *fuzzy*, e seus híbridos (GONZÁLEZ et al., 2007; HESSBURG et al., 2007; KALOUDIS et al., 2005; VADREVU et al., 2010).

(iv) A incerteza de decisão refere-se à informação imperfeita envolvida na análise custo/benefício social, também referida como a incerteza no valor ou incerteza na preferência. Na medida em que não sabemos completamente as preferências/valores sociais, a nossa capacidade de gerir para o bem-estar social é limitada. Este tipo de incerteza é geralmente tratada com alguma forma de um método de mensuração do valor (DIAZ-BALTEIRO; ROMERO, 2008; MENDOZA;

MARTINS, 2006). Nestes, vemos a interseção de economia de recursos com a literatura de ciência da decisão. As abordagens de avaliação não-mercantis econômicos incluem preços hedônicos, modelos custo-viagem, avaliação contingente e modelagem de escolha (VENN; CALKIN, 2009). As abordagens na literatura de apoio à decisão florestal para lidar com a incerteza de decisão incluem o *Analytic Hierarchy Proces* (AHP), a teoria da utilidade e teoria da escolha social (MENDOZA; MARTINS, 2006).

Gerir a incerteza de decisão pode ser bastante desafiador, devido à presença de múltiplas partes interessadas com diferentes perspectivas, percepções e objetivos, e o fato de que as preferências podem mudar com o tempo ou quando houver mais informações disponíveis (BRAGA; STARMER, 2005; BROWN et al., 2008; MAGUIRE; ALBRIGHT, 2005; REISKAMP et al., 2006).

O Quadro 1 ilustra um mapeamento de fontes de incerteza no contexto dos incêndios florestais para o tipo de incerteza comum, juntamente com as abordagens de apoio à decisão mais utilizados.

Quadro 1 – Incertezas no planejamento estratégico do fogo florestal para a tipologia de incerteza de Ascough et al. (2008)

Fontes de Incerteza	**Tipo de Incerteza**	**Metodologia**
Ocorrência de incêndio	Variabilidade	Baseado em probabilidade
Comportamento do fogo	Variabilidade; conhecimento	Baseado em probabilidade; sistema especialista
Resposta ecológica ao fogo	Conhecimento	Sistema especialista
Eficácia dos tratamentos de gestão	Conhecimento	Sistema especialista
Valorização dos recursos não-mercantis	Decisão	Mensuração do valor
Interação do fogo com outro distúrbio	Variabilidade; conhecimento	Baseado em probabilidade; sistema especialista
Dinâmica temporal de vegetação e combustível	Variabilidade; conhecimento	Baseado em probabilidade; sistema especialista
Contabilização do papel das mudanças climáticas	Variabilidade; conhecimento	Baseado em probabilidade; sistema especialista

Fonte: Thompson e Calkin, 2011, p. 1897.

A modelagem do comportamento do fogo, por exemplo, implica na incerteza da variabilidade em termos de processos, tais como frequência/localização das ignições e vento padrões, bem como a incerteza do conhecimento em termos de como nós representamos as condições dos combustíveis, a qualidade dos dados de entrada da paisagem e características dos combustíveis, e como modelar o

movimento de incêndio em toda a paisagem. Os sistemas de apoio à decisão ajudam a reduzir o âmbito de aplicação destas incertezas e facilita o conhecimento dos riscos na tomada de decisão.

2.2. Análise de ocorrência de incêndio e métodos de modelagem

A pesquisa de ocorrência de incêndio foi amplamente realizada para obter uma melhor compreensão dos fatores espaciais e temporais que influenciam as ignições de incêndios florestais e para desenvolver modelos que possam ser usados para predizer a probabilidade de ignição em diferentes áreas e com diferentes condições climáticas.

A pesquisa de ocorrência de incêndio baseada espacialmente está principalmente preocupada com a distribuição espacial das ignições em relação às variáveis geográficas, como o terreno e as características da paisagem humana. Foram utilizadas uma série de metodologias para determinar os fatores espaciais que influenciam a ocorrência de incêndio. Estas incluem testes de hipóteses estatísticas tradicionais (CARDILLE; VENTURA, 2001; MAINGI; HENRY, 2007; MERCER; PRESTEMON, 2005), análise de regressão linear (DONOGHUE; MAIN, 1985) e análise de regressão logística (CARDILLE et al., 2001; CATRY, 2009; KALABOKIDIS et al., 2007; MARTÍNEZ, 2009; PEW; LARSEN, 2001; SYPHARD et al., 2008), árvore de classificação e regressão (AMATULLI et al., 2006) e métodos de rede bayesiana (DILTS et al., 2009; DLAMINI, 2010). Algumas pesquisas investigaram o agrupamento de pontos de ignição e usaram a função K de Ripley e função-L para avaliar o agrupamento e a densidade *kernel* para fornecer representações gráficas (GENTON et al., 2006; HERING et al., 2009; PODUR et al., 2003; TURNER, 2009; YANG et al., 2007). Uma lista contendo os detalhes de pesquisas publicadas de ocorrência espacial de incêndio é apresentada no Quadro 2.

A pesquisa de ocorrência de incêndio baseada temporalmente foi realizada para modelar a probabilidade de ocorrência de fogo, definida como um ou mais incêndios que ocorrem dentro de limites temporais e espaciais definidos. A pesquisa de ocorrência temporal de incêndio também foi realizada para estimar o número de ignições que podem ocorrer em um determinado dia. Esses modelos usam principalmente variáveis dinâmicas do tempo e índices climáticos de fogo.

Quadro 2 – Exemplos de modelos e análises de ocorrência de incêndios baseados espacialmente

Referência	Origem	Tipo de ignição	Método	Fatores significativos
Donoghue e Main (1985)	Leste dos EUA	Antropogênico	Regressão linear	Latitude, clima (precipitação), densidade populacional (não urbana) e número de processos judiciais e condenações
Chou (1993)	Sul da Califórnia, EUA	Não especificado	Regressão logística	Topografia, vegetação, temperatura, precipitação, proximidade com edifícios e transportes
Cardille e Ventura (2001)	Alto Centro-Oeste dos EUA	Todos (97% antropogênico)	Análise estatística (Teste-Z)	Posse de terra
Cardille et al. (2001)	Alto Centro-Oeste dos EUA	Todos (96% antropogênico)	Regressão logística	Densidade de estradas e demográfica, temperatura e precipitação
Pew e Larsen (2001)	Ilha de Vancouver, Canadá	Antropogênico	Regressão logística	Temperatura, precipitação, distância das cidades, rodovias e ferrovias
Podur et al. (2003)	Ontário, Canadá	Raio	Função K	Tempo seco localizado e tempestade
Mercer e Prestemon (2005)	Flórida, USA	Não especificado	Análise estatística (Estimativa de verossimilhança)	Desemprego, pobreza e número de policiais
Amatulli et al (2006)	Sudeste da Itália	Não especificado	Árvore de decisão	Cobertura da terra, temperatura dos meses mais quentes e mais frios, declividade e altitude
Genton et al. (2006)	Flórida, USA	Todos (75% antropogênico)	Função K	Raio, incêndios criminosos e de ferrovia
Kalabokidis et al. (2007)	Norte da Grécia	Não especificado	Regressão logística	Cobertura da vegetação, declividade, altitude, pecuária
Maingi e Henry (2007)	Kentucky, EUA	Antropogênico (incêndio criminoso)	Análise estatística (teste de Kruskal-Wallis, correlação)	Distância a estradas e lugares povoados, altitude e declividade
Yang et al. (2007)	Missouri, EUA	Antropogênico	Função K	Posse de terreno, distância a cidades e estradas, tipo de floresta e declividade
Romero-Calcerrada et al. (2008)	Região de Madri, Espaha	Antropogênico	Estatística bayesiana (pesos de evidência)	Proximidade a áreas urbanas e estradas
Syphard et al (2008)	California, USA	Principalmente antropogênico	Regressão logística	Distância a estradas e trilhas, tipo de vegetação, interface floresta-meio urbano e temperatura mínima média de janeiro
Dilts (2009)	Nevada, EUA	Raio	Modelagem bayesiana	Densidade de raio, rugosidade topográfica
Martinez et al. (2009)	Espanha	Antropogênico	Regressão logística	Fragmentação da paisagem agrícola, abondono de terras agrícolas e processos de desenvolvimento
Catry et al. (2009)	Portugal	Principalmente antropogênico	Regressão logística	Densidade populacional, cobertura da terra, altitude e distância de estradas
Dlamini (2010)	Suazilândia	Não especificado	Modelagem bayesiana	Cobertura da terra, altitude, precipitação média anual e temperatura média anual

Algumas análises e modelos de ocorrência de incêndios temporais foram especificamente realizados para selecionar o índice de perigo de incêndio mais apropriado para uma área (ANDREWS et al., 2003; HAINES et al., 1983; PADILLA; VEGA-GARCIA, 2011; VASILAKOS et al., 2009; VIEGAS et al., 1999). Da mesma forma, outros artigos compararam a ocorrência de fogo com o teor de umidade de combustível vivo e morto (CHUVIECO et al., 2009; VIEGAS et al., 1992). A influência de variáveis geográficas foi minimizada na maioria dos trabalhos de ocorrência de fogo temporal, dividindo a paisagem em unidades relativamente homogêneas e tratando estas individualmente.

A grande maioria dos artigos considerou a probabilidade de ocorrência de fogo na escala diária (Quadro 3), modelada principalmente por meio de regressão logística (ANDREWS et al., 2003; MARTELL et al., 1989; PADILLA; VEGA-GARCIA, 2011; PREISLER, 2004; VEGA-GARCIA et al., 1995; VILAR, 2010). Outros autores utilizaram métodos como redes neurais artificiais (VASILAKOS et al., 2009) e árvores de classificação e regressão (KRUSEL et al., 1993).

Quadro 3 – Exemplos de modelos temporais e análises de ocorrência de incêndios (probabilidade de um ou mais incêndios em um determinado dia)

Referência	Origem	Tipo de ignição	Método	Fatores significativos
Martell et al. (1989)	Ontário, Canadá	Antropogênico	Regressão logística	Umidade do combustível fino[1] e dia da estação
Krusel et al. (1993)	Noroeste de Victória, Austrália	Não especificado	Árvore de decisão	Variáveis meteorológicas (temperatura, dias desde a última chuva, índice de seca de Keetch-Byram, velocidade do vento e humidade relativa)
Vega-Garcia et al. (1995)	Alberta, Canadá	Antropogênico	Regressão logística	Distrito, Índice de acumulação[1] e Índice de propagação inicial[1]
Andrews et al. (2003)	Arizona, USA	Todos	Regressão logística	Componente de liberação de energia[2]
Preisler et al. (2004)	Oregon, USA	Todos	Regressão logística	Localização espacial, dia no ano, altitude, umidade do combustível de 1000horas[2], temperatura do bulbo seco e estado do tempo
Albertson et al. (2010)	Reino Unido	Antropogênico	Modelo probit	Chuva, temperatura, feriado público, dia da semana e mês
Vasilakos et al. (2009)[1]	Ilha de Lesbos, Grécia	Principalmente antropogênico	Redes neurais	Chuva, umidade de combustível de 10horas[2], mês, umidade relativa, altitude e dia da semana
Chuvieco et al. (2009)	Espanha Central	Todos	Regressão de Poisson	Umidade do combustível vivo
Vilar et al. (2010)	Região de Madri, Espanha	Antropogênico	Regressão logística	Dia do ano, densidade urbana, distância de rodovia e ferrovia, altitude e temperatura máxima

[1] Componente do Canadian Forest Fire Weather Index System – CFFWIS (WAGNER, 1987).
[2] US National Fire Danger Rating System – NFDRS (DEEMING et al., 1997).
Fonte: Plucinski (2012), adaptado pelo autor.

Alguns trabalhos de ocorrência de fogo temporal apresentaram modelos para predizer o número de incêndios que ocorrem em um único dia (Quadro 4). Estes usaram a regressão de poisson (CUNNINGHAM; MARTELL 1973; WOTTON et al., 2003) e modelagem bayesiana (TODD; KOURTZ, 1991).

Quadro 4 – Exemplos de modelos temporais de predição do número de incêndios diários

Referência	Origem	Tipo de ignição	Método	Fatores significativos
Cunningham e Martell (1973)	Ontário, Canadá	Não especificado	Regressão de Poisson	Umidade do combustível fino[1]
Haines et al. (1983)	Nordeste dos EUA	Não especificado	Regressão linear	Componente de ignição[2]
Todd e Kourtz (1991)	Quebec, Canadá	Antropogênico	Modelagem bayesiana	Velocidade do vento, umidade do combustível fino[1] e umidade da turfa[1]
Garcia Diez et al. (1994)	Galícia, Espanha	Não especificado	Autoregressivo	Nível de estabilidade atmosférica e déficit de saturação
Mandallaz e Ye (1997)	Sul da Suécia	Todos	Regressão de Poisson	Região, dia da semana, histórico de fogo recente, umidade relativa e ETP[3]
	Sul da França	Todos	Regressão de Poisson	ICONA[3], histórico de fogo recente e IP[3]
	Norte da Itália	Todos	Regressão de Poisson	Região, histórico de fogo recente, precipitação, velocidade do vento, umidade do combustível fino[2]
	Portugal	Todos	Regressão de Poisson	RN[3], IREPI[3] e histórico de fogo recente
Garcia Diez et al. (1999)	Galícia, Espanha	Não especificado	Autoregressivo	Estabilidade atmosférica e umidade
Anderson (2002)	Canadá	Raio	Modelo probabilístico	Número de raios, teor de umidade do combustível, chuva e tipo de floresta
Wotton et al. (2003)	Ontário, Canadá	Antropogênico	Regressão de Poisson	Umidade do combustível fino[1], umidade da turfa[1] e grau de secura[1]
Wotton at al. (2010)	Canadá	Antropogênico	Regressão de Poisson	Ecoregião, umidade de combustível fino[1], umidade da turfa[1], grau de secura[1], estação do ano
		Raio		

[1] Componente do Canadian Forest Fire Weather Index System – CFFWIS (WAGNER, 1987).
[2] US National Fire Danger Rating System – NFDRS (DEEMING et al., 1997).
[3] Índices de perigo de incêndio indefinidos e não referenciados – suiço (ETP), espanhol (ICONA), português (IP), italiano (IREPI) e francês (RN) utilizados por Mandallaz e Ye (1997).
Fonte: Plucinski (2012), adaptado pelo autor.

Uma variedade de técnicas de análise e modelagem foram aplicadas aos dados de ocorrência de incêndio, embora a regressão logística tenha sido utilizada em grande parte das pesquisas e para a maioria dos modelos prevendo a ocorrência de um ou mais incêndios dentro de limites espaciais e temporais definidos.

A pesquisa de ocorrência de incêndio com base espacial foi utilizada para identificar áreas com alto risco de ignição. Esta informação foi utilizada em análises de risco de incêndio mais amplas e em operações de gerenciamento de incêndio, onde ajudou a segmentação de tratamentos de combustível e a alocação de recursos de supressão.

2.2.1. Banco de dados de ocorrência de incêndios florestais

Toda pesquisa de ocorrência de incêndio baseia-se em registros de fogo e requer dados que abranjam várias estações de fogo. Os bancos de dados de ocorrência de incêndio mantidos pelas agências de combate ao fogo são a fonte mais comum desses dados. Embora ocasionalmente, quando esses registros não estão disponíveis, os arquivos de imagens de satélite foram usados para pesquisa espacial de ocorrência de incêndio (DLAMINI, 2010; MAINGI; HENRY, 2007; PRASAD et al., 2008). Os documentos de pesquisa baseados em dados de satélite dependem da identificação de cicatrizes de fogo, e é improvável que detectam todos os incêndios, particularmente incêndios menores. Embora os registros oficiais de agências de bombeiros sejam uma fonte de dados mais confiável, pode haver variabilidade significativa nos padrões de relatórios mantidos pelas agências (ANDREWS et al., 2003).

A disponibilidade de registros de fogo restringiu a pesquisa de ocorrência de incêndio no passado. As tendências recentes para o aumento da captura de dados, dentro de agências de combate a incêndios em muitos países e os avanços em programas de computador projetados para análise e modelagem de dados possibilitaram a realização de mais pesquisas neste campo.

Muitas áreas do mundo afetadas por incêndios não possuem registros confiáveis de longo prazo de ocorrência de fogo. Desde o advento da observação da Terra a partir do espaço, o sensoriamento remoto tornou-se uma ferramenta valiosa para a comunidade científica e gestores de recursos naturais, pois permite a coleta periódica de dados em diferentes faixas do espectro eletromagnético em zonas muito vastas e inacessíveis da Terra (KENNEDY et al., 2009).

O mapeamento de área queimada pelos sistemas globais de satélites fornece informações sobre a sazonalidade do fogo, frequência de ocorrência,

localização e quantificação da área queimada, que é essencial para o desenvolvimento de políticas de gestão ambiental. Anteriormente e na ausência de produtos precisos de área queimada, as avaliações da área queimada foram criadas com base na calibração dos dados de fogo ativos disponíveis do *Advanced Very High Resolution Radiometer* – AVHRR regionais e dados *Moderate Resolution Imaging Spectroradiometer* – MODIS globais. No entanto, vários fatores de controle remoto, ambientais e de comportamento do incêndio limitaram a precisão de tais conjuntos de dados da área afetada pelo fogo.

A disponibilidade de dados com localização geográfica robustamente calibrados, atmosfericamente corrigidos, fornecidos pela última geração de sistemas de detecção remota de resolução moderada, permitem grandes avanços no mapeamento por satélites, da área afetada pelo fogo.

O produto MCD45A1 é distribuído mensalmente e é parte da coleção MODIS 5 (ROY et al., 2008) desenvolvido para mapear a área afetada pelo fogo. Este produto utiliza dados diários de ambas as plataformas Terra e Aqua e é gerado a partir de uma série temporal de dados de reflectância de superfície da terra. O algoritmo *Burned Area* (BA) é um algoritmo que detecta a mudança da estrutura da vegetação, baseado no modelo de reflectância bidirecional e no uso de uma medida estatística, para detectar a probabilidade de mudança de um estado observado anteriormente. O algoritmo mapeia a extensão espacial (500m de resolução) dos últimos incêndios e não de incêndios que ocorreram em épocas ou anos anteriores.

O método de mudança é aplicado a pixels georreferenciados, em longas séries de observações de reflectância. O algoritmo inicia pelo pré-processamento da reflectância da superfície da terra, detectada pelas sete bandas do sensor MODIS. Uma reflectância diária é prevista com base na reflectância detectada nos 16 dias anteriores. Esta informação é usada para modelar uma Função de Distribuição de Reflectância Bidirecional (BRDF), que permite o gerenciamento das variações angulares de reflectância. Este modelo é usado para prever mudanças na reflectância estimando observações subsequentes no tempo. Em seguida, uma medida estatística (*Z-score*) é calculada a partir das bandas 2 e 5 do sensor MODIS, para cada pixel georreferenciado, para determinar a queima do pixel com a diferença entre a reflectância observada e prevista. Esta decisão é baseada em um limite (*Zthre*). Finalmente, o algoritmo reduz os erros de

comissão ao selecionar, a partir de pixels candidatos, os que fornecem evidências persistentes de ocorrência de fogo. Uma descrição mais detalhada do algoritmo aplicado para se obter os produtos BA MODIS pode ser encontrada em Roy et al. (2005).

2.2.2. Fatores que afetam a ocorrência de fogo

Uma grande variedade de fatores que afetam a ocorrência de fogo foram usados em modelos de predição (Quadros 2-4). Estes dependem dos tipos de ignição que ocorrem nos locais considerados e das variáveis disponíveis para análise. Muitos trabalhos de ocorrência de fogo consideraram incêndios por raio e incêndios antropogênicos como causas separadamente. Todavia, a maioria dos trabalhos apenas considerou incêndios de uma dessas categorias de ignição.

Algumas análises espaciais de ignição de fogo por raio associaram-se a características do terreno (DILTS et al., 2009; KILINC; BERINGER, 2007; McRAE; 1992; VÁZQUEZ; MORENO, 1998) e áreas com combustíveis mais secos (PODUR et al., 2003). Modelos e análises temporais de ocorrência de incêndio por raio evidenciaram a importância da umidade do combustível e das chuvas para a predição. Muitos modelos usaram índices dentro do *Canadian Forest Fire Weather Index System* – CFFWIS (WAGNER, 1987), particularmente o *Duff Moisture Code* (ANDERSON, 2002; MARTELL et al., 1989; PODUR et al., 2003), que está associado ao teor de umidade de camadas de combustível mais profundas.

A maioria das pesquisas de ocorrência espacial de incêndio foram realizadas em áreas dominadas por ignições antropogênicas. Estas relacionaram a ocorrência de fogo antropogênico com uma série de variáveis geográficas associadas à densidade populacional (CARDILLE et al., 2001; CATRY, 2009; PRASAD et al., 2008; MERCER; PRESTEMON, 2005; ROMERO-CALCERRADA et al., 2008); proximidade de estradas, cidades e infraestrutura (CATRY et al. 2009; CHOU et al., 1993; MAINGI; HENRY, 2007; MARTINEZ et al., 2009; PEW; LARSEN; 2001; ROMERO-CALCERRADA et al., 2008; SYPHARD et al., 2008; VILAR et al., 2010) e uso da terra (CARDILLE; VENTURA; 2001; ROMERO-CALCERRADA et al., 2008). Algumas investigações de ocorrência de incêndios antropogênicos também identificaram variáveis socioeconômicas, como a pobreza e as taxas de desemprego como tendo alguma influência (MAINGI; HENRY, 2007; MARTINEZ et

al. 2009; MERCER; PRESTEMON, 2005). As variáveis de combustível foram consideradas em alguns trabalhos, principalmente em termos de tipo de vegetação (PADILLA; VEGA-GARCÍA, 2011; SYPHARD et al., 2008; VEGA-GARCIA et al., 1996), mas foram consistentemente menos significativos do que as variáveis humanas e; as variáveis climáticas foram consideradas com base em interpolações geográficas de valores anuais ou mensais médios (DLAMINI, 2010; PEW; LARSEN, 2001; PRASAD et al., 2008).

Os modelos de ocorrência temporal de incêndio que consideram as ignições antropogênicas assumem que as medidas de prevenção de incêndio, uso da terra e as variáveis socioeconômicas permanecem constantes durante o período de coleta de dados. Muitos desses modelos relacionaram a ocorrência de incêndio com o teor de umidade dos combustíveis de superfície, estimado pela umidade de combustível fino no CFFWIS (MARTELL et al., 1989; PADILLA; VEGA-GARCÍA, 2011; VEGA-GARCIA et al., 1995; WOTTON et al., 2003). Alguns modelos incluíram variáveis associadas à data (dia do ano ou estação) para explicar a distribuição de incêndios que ocorrem durante a estação de fogo (ALBERTSON et al., 2010; MARTELL et al., 1989; PREISLER et al., 2004; VILAR et al., 2010).

Algumas pesquisas de ocorrência de incêndio consideraram a influência de diferentes tipos de ignição. Martell et al. (1989) e Wotton et al. (2003) dividiram os incêndios antropogênicos em dois grupos com base em suas distribuições anuais de ocorrência. Os estudos das ignições provocadas por incêndios criminosos de outras ignições antropogênicas na predição de fogo são importantes, pois observam-se tendências espaciais distintas relacionadas à acessibilidade (MAINGI; HENRY, 2007) e tendências temporais associadas aos fins de semana e feriados (MANDALLAZ; YE, 1997)

A pesquisa de ocorrência de incêndios florestais foi realizada em muitas partes do mundo para melhorar o conhecimento de fatores que afetam a distribuição de ignição de incêndios e desenvolver modelos preditivos. Estes fatores refletem a variedade de clima, paisagem e culturas nas regiões onde foram realizadas. Em geral, as ignições antropogênicas foram influenciadas por variáveis de geografia humana, bem como a ocorrência de fins de semana e feriados públicos e o teor de umidade dos combustíveis de superfície. As ignições do fogo por raio foram associadas à precipitação e ao teor de umidade das camadas de combustível mais profundas.

Os principais fatores influentes na ocorrência de incêndios florestais foram descritos na literatura, pelas variáveis topográficas, climáticas, socioeconômicas e de vegetação.

As características topográficas afetam a distribuição da vegetação, composição, inflamabilidade e também influenciam as variações climáticas locais (SYPHARD et al., 2008; WHELAN, 1995). A altitude é um importante fator fisiográfico que está relacionado ao comportamento do vento e, portanto, afeta a propensão do fogo (ROTHERMEL, 1983). A declividade é um indicador de taxa de mudança de elevação (em graus). A inclinação da declividade afeta a intensidade da radiação e a umidade do combustível. À medida que um incêndio se move sobre a paisagem, seu comportamento pode mudar com a hora do dia e as características topográficas, devido às variações na intensidade de radiação solar recebida (PYNE, et al., 1996). Portanto, a declividade tem influência nas condições de pré-aquecimento dos combustíveis e modifica a taxa de propagação e direção do fogo (BROW; DAVIS, 1973). As encostas íngremes aumentam a velocidade do fogo, porque o pré-aquecimento convectivo e a taxa de ignição são mais efetivos (KUSHLA; RIPPLE, 1997; ROTHERMEL, 1983; TROLLOPE et al., 2002).

A ocorrência de fogo, frequência, bem como a intensidade dependem principalmente do clima (FLANNIGAN; WOTTON, 2001). Os fatores climáticos e do tempo, também desempenham um importante papel no comportamento e propagação do fogo.

A temperatura do ar desempenha um papel importante no comportamento do fogo. Seu efeito direto é influenciar a temperatura do combustível e, portanto, a quantidade de calor necessário para o elevar ao ponto de ignição. A temperatura do ar também tem efeitos indiretos por meio da sua influência sobre a umidade relativa do ar e perdas de umidade da vegetação por evaporação (YAKUBU et al., 2015).

A precipitação é fator fundamental na definição da estação do fogo. Segundo estudos (BRAVO et al., 2010; MORENO et al., 2011; PAUSAS, 2004; PEREIRA et al., 2005) foi sugerida a hipótese de que a precipitação ocorrida fora da estação de incêndio pode afetar a ocorrência de incêndios, favorecendo o crescimento sazonal da vegetação, resultando em aumento da disponibilidade de combustíveis finos (principalmente em pastagens), onde os incêndios podem iniciar e se espalhar mais facilmente durante a principal estação de fogo; pelo contrário, a precipitação

durante a estação de fogo pode dificultar a ocorrência de incêndios pelo aumento do teor de umidade dos combustíveis e limitar a ignição e propagação do fogo.

A radiação solar tem influência sobre a umidade e o tipo de material combustível pela disponibilidade de energia.

A deficiência hídrica também é uma variável climática que existe uma forte relação com incêndios florestais. Soares (1985) afirma que em prolongados períodos de seca, o material combustível cede umidade ao ambiente, tornando as condições favoráveis à ocorrência de incêndios.

Os padrões espaciais de ocorrência de incêndios estão fortemente associados com o acesso humano à paisagem. Em geral, o risco de incêndio aumenta em áreas mais próximas a estradas e com interface entre o meio urbano e a floresta (PEW; LARSEN, 2001; RODRÍGUEZ-SILVA et al., 2010; SOTO, 2012). Nos últimos anos, as medidas de influência humana (distância a estradas, distâncias de área de lazer, densidade populacional, entre outros) foram consideradas em pesquisas no campo de risco de incêndios florestais. Contudo, o problema na prevenção de incêndios é, que muitas vezes, existem grandes lacunas na informação disponível sobre a distribuição espacial dos recursos humanos, sendo um grande desafio para o desenvolvimento de planos de manejo (ROMERO-CALCERRADA et al., 2010).

O combustível é um importante elemento no triângulo do fogo. Influencia na facilidade de ignição, bem como o tamanho e a intensidade do fogo (PYNE et al., 1996). A carga de combustível é considerada como um dos fatores mais importantes que influenciam o comportamento do fogo, porque a quantidade total de energia calorífica disponível para liberação durante o incêndio está relacionada com a quantidade de combustível. Assumindo uma produção de calor constante, a intensidade de um incêndio é diretamente proporcional à quantidade de combustível disponível para combustão, a qualquer taxa de propagação da frente do fogo (YAKUBU et al., 2015).

2.3. Árvore de decisão

A partir da ideia inicial de Hunt (QUINLAN, 1993), no final da década de 50, as árvores de decisão foram usadas com sucesso em sistemas de aprendizado de máquina e têm sido estudadas tanto na área de reconhecimento de padrões quanto

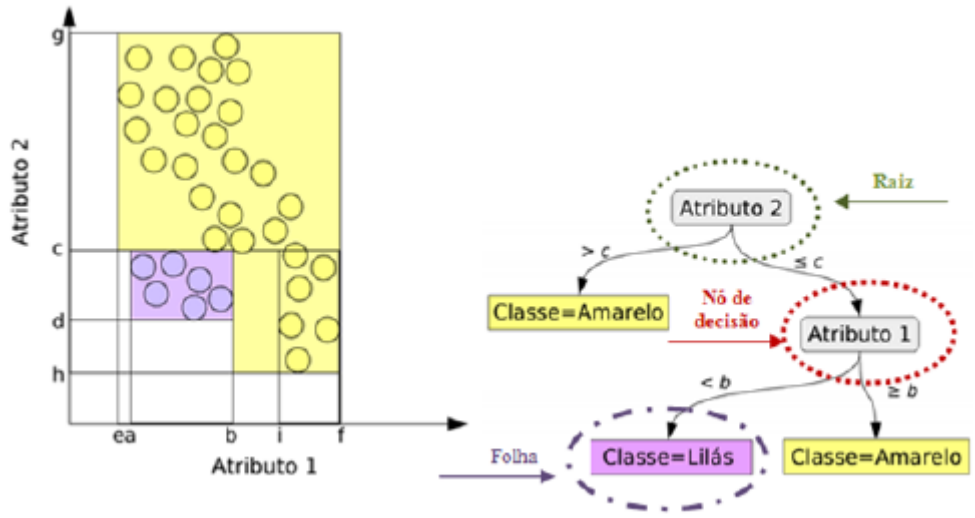
na área de aprendizado de máquina em diversos campos de pesquisa. Os modelos estruturados em árvores são modelos não paramétricos, cuja filosofia é utilizar modelos simples para sub-amostras dos dados, dividindo o problema em partes. Frequentemente, esses modelos são estimados por meio de algoritmos recursivos de particionamento.

A árvore de decisão consiste de uma hierarquia de nós internos e externos que são conectados por ramos. O nó interno, também conhecido como nó decisório ou nó intermediário, é a unidade de tomada de decisão que avalia por meio de teste lógico qual será o próximo nó descendente ou filho. Em contraste, um nó externo, também conhecido como folha ou nó terminal, está associado a um rótulo ou a um valor.

Em geral, o procedimento de uma árvore de decisão é o seguinte: apresenta-se um conjunto de dados ao nó inicial (ou nó raiz que também é um nó interno) da árvore; dependendo do resultado do teste lógico usado pelo nó, a árvore ramifica-se para um dos nós filhos e este processo é repetido até que um nó terminal seja alcançado. A repetição deste procedimento caracteriza a recursividade da árvore de decisão.

No caso das árvores de decisão binária, cada nó intermediário divide-se exatamente em dois nós descendentes: o nó esquerdo e o nó direito. Quando os dados satisfazem o teste lógico do nó intermediário seguem para o nó esquerdo e quando não satisfazem seguem para o nó direito. Logo, uma decisão é sempre interpretada como verdadeira ou falsa. A descrição de divisão para árvores binárias é descrita nesta tese. Contudo, na literatura, árvore de decisão com outras divisões pode ser encontrada em Zighed e Rakotomalala (2000). O nível de um nó é a distância do mesmo até a raiz, ou seja, o número de ligações percorridas até chegar a raiz. A profundidade de uma árvore de decisão é definida pela maior distância entre uma folha e a raiz, existindo árvores com profundidade uniforme em todas as folhas e outras não (MURTHY, 1997). A representação de uma árvore de decisão pode ser visualizada na Figura 1.

Figura 1 – Exemplo de partição dos espaços atributos (esquerda) e exemplo de uma árvore de decisão (direita).



Fonte: Santos (2011).

O aprendizado de uma árvore de decisão é supervisionado, ou seja, compreende a abstração de um modelo de conhecimento a partir dos dados apresentados na forma de pares ordenados (entrada, saída desejada). As árvores treinadas podem ser representadas como um conjunto de regras “Se-Então” para melhoria da compreensão e interpretação (GOLDSCHMIDT; PASSOS, 2005).

2.3.1 O processo de crescimento da árvore

A tarefa de construção de uma árvore de decisão é chamada de indução. O processo de indução de árvores de decisão é realizado por meio da estratégia *Top Down*, que inicia a geração da árvore pela raiz e continua por seus filhos até que um critério de parada seja encontrado. Existem várias técnicas para a eleição do melhor preditor a ser utilizado em um nó de uma árvore de decisão, constituindo a função de avaliação de cada partição e o segredo para o sucesso do algoritmo de indução. Assim, a função de avaliação verifica cada preditor candidato e seleciona aquele que maximiza (ou minimiza) alguma função heurística sobre os subconjuntos.

Entre as principais funções para a escolha da melhor divisão do preditor em cada nó da árvore estão:

a) Escolha randômica: citado por Baranauskas e Monard (2000), este método não utiliza nenhuma heurística, sendo escolhido qualquer preditor entre os preditores disponíveis.

b) Critério de entropia: desenvolvido por Quinlan em 1993, este critério mede a quantidade de informação necessária para codificar a classe do nó. Assim, um conjunto de entrada S que pode ter c classes distintas, a entropia de S será dada por:

$$E(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \quad (1)$$

Em que: p_i é a proporção de dados em S que pertence à classe i.

A entropia é uma medida aplicável à partição de um espaço de probabilidade, medindo quanto esse espaço é homogêneo, ou por outro lado, quanto maior a entropia maior a desordem. A entropia atinge seu valor máximo, igual a 1, quando o conjunto de dados é heterogêneo (MITCEHELL, 1997).

O ganho de informação para um atributo A, de um conjunto de dados S, nos informa a medida da diminuição da entropia esperada quando utilizamos o atributo A, para fazer a partição do conjunto de dados.

Seja $P(A)$, o conjunto de valores que o atributo A pode ter, seja x, um elemento desse conjunto, e seja S_x, um subconjunto de S formado pelos dados em que $A = x$, a entropia que se obtêm ao particionar S em função do atributo A é dada por:

$$E(A) = \sum_{x \in P(A)} \frac{|S_x|}{|S|} E(S_x) \quad (2)$$

O ganho de informação será dado por:

$$Ganho(S, A) = E(S) - E(A) \quad (3)$$

Desta forma, o ganho de informação, mede a eficácia de um atributo em classificar os dados de treino, a escolha do atributo mais eficaz, que mais reduz a entropia, faz com que a tendência seja a de gerar árvores, que são, em geral, menos profundas, com menos nós e ramificações.

c) Técnica de Laplace: citado por Fonseca (1994), esta técnica escolhe e usa, para realizar a partição do conjunto de treinamento em cada nó, o atributo ou preditor que minimizar o valor do erro esperado:

$$E = \sum_{i=1}^{m} \frac{n_i}{n} \sum_{j=1}^{K} p_{i,j}\left(1 - p_{i,j}\right) \tag{4}$$

Em que: m é o número de subconjuntos da partição; n é o número de exemplos de partição; n_i é o número de exemplos que possuem o *i-ésimo* valor; K é o número total de classe e; $p_{i,j}$ é a probabilidade de um exemplo da classe C_j ter o *i-ésimo* valor, expressa como:

$$p_{i,j} = \frac{n_{i,j} + 1}{n_i + k} \tag{5}$$

Neste caso, *o i-ésimo* valor se refere a cada um dos m valores do atributo em questão.

d) Índice Gini: desenvolvido por Conrado Gini em 1912, mede o grau de heterogeneidade dos dados. Logo, é utilizado para medir a impureza de cada nó (ONODA, 2001). A impureza de um nó é máxima quando os registros estão igualmente distribuídos entre todas as classes ($1\text{-}1/n_k$), e mínima quando todos os registros pertencem a uma classe. Considerando um conjunto de dados S, que contém n registros, cada um com uma classe C_i, o índice Gini é definido por:

$$G(S) = 1 - \sum_{i=1}^{k} p\left(C_i | n\right)^2 \tag{6}$$

Em que: p é a probabilidade relativa da classe C_i em S; n é o número de registros em S e; k é o número de classes.

Se S for particionado em dois subconjuntos S_1 e S_2, um para cada ligação, o índice Gini dos dados particionados será dado por:

$$G(S,A) = \frac{n_1}{n} G(S_1) + \frac{n_2}{n} G(S_2) \tag{7}$$

Em que: n_1 é o número de exemplos de S_1 e; n_2 é o número de exemplos de S_2.

e) Método da paridade ou critério de Twoing: aplicado em árvores binárias para definir a medida de impureza, sendo demonstrado por Fonseca (1994) como:

$$\Delta i(S) = \frac{p_E p_D}{4} \left[\sum_{i=1}^{k} \left| p(C_i | n_E) - p(C_i | n_D) \right| \right]^2 \tag{8}$$

Em que: p_E é a probabilidade do nó descendente esquerdo; p_D é a probabilidade do nó descendente direito; $p(C_i|n_E)$ é a probabilidade da classe C_i do nó descendente esquerdo e; $p(C_i|n_D)$ a probabilidade da classe C_i do nó descendente direito.

Neste método, o atributo que minimiza a impureza do nó é escolhido como melhor atributo para divisão. Assim, os critérios apresentados selecionam o melhor atributo para um determinado subconjunto de treinamento, em que a sua medida indica quão bem este atributo discrimina a classe.

2.3.2. Definição da partição

Ao estabelecer a variável a ser utilizada no nó, é necessário definir o particionamento das ligações do mesmo, pois muitos critérios de escolha do nó utilizam a distribuição da partição em seus cálculos. É importante ressaltar, que o tipo de partição realizada em cada nó afeta de forma decisiva o desempenho da

árvore. O tipo das variáveis também é um importante fator, pois o tratamento dado às variáveis discretas é diferente em relação às variáveis contínuas.

Dessa forma, para definir o particionamento de cada nó para variáveis discretas, pode-se utilizar os seguintes critérios:

a) Criação de uma ligação para cada valor da variável: é atribuído uma ligação para cada valor da variável. Embora permita extrair da variável todo o seu conteúdo informativo, a principal desvantagem do método é a criação de um número grande de ramificações, o que ocasiona a geração de árvores complexas e com ligações desnecessárias. No entanto, é o método mais simples.

b) Criação de nós binários: esta solução foi criada por Hunt (1966), em que é atribuído a uma das ligações um dos valores da variável eleita e à outra ligação todos os outros valores. Esta solução é bastante simples e inteligível, porém não aproveita todo o poder de discriminação da variável.

c) Ordenação dos valores: também cria duas ligações, sendo atribuídos a uma ligação, os valores de $x_n \leq A$, em que x_n é uma variável e A é um valor, e à outra ligação os valores de $x_n > A$. Este método, desenvolvido por Breiman et al. (1984), somente pode ser utilizado para atributos que possuam uma relação de ordem entre os seus possíveis valores. Para um atributo ou variável com cardinalidade n, serão possíveis $n-1$ diferentes partições (ou valores de A), tornando-se necessário testar todas para escolher a melhor delas. Proporciona uma árvore bastante compreensível para o usuário, apesar de não utilizar a capacidade total de cada característica.

d) Agrupamento de valores em dois conjuntos: apresentado por Breiman et al. (1984), este método também propõe a criação de duas ligações, associando um subconjunto de valores do atributo a uma das ligações e um outro subconjunto à outra ligação. Como todos os subconjuntos possíveis serão testados, totalizando $(2^{n-1}-1)$ partições, de modo a garantir a seleção da melhor partição, possui a grande desvantagem de necessitar de um número muito grande de testes, principalmente quando a cardinalidade do atributo é elevada.

e) Agrupamento de valores em vários conjuntos: proposto por Quinlan (1993), o método permite agrupar valores da variável em várias partições, não somente em duas. Para tal, é calculado o valor da solução, atribuindo a cada diferente característica da variável uma ligação própria. Em seguida, são testadas todas as combinações possíveis de dois valores. Se nenhuma dessas soluções for considerada boa pelo critério de avaliação (discutido mais adiante), o processo é interrompido, sendo a solução anterior adotada como solução. Caso contrário, repete-se o processo de testar as diversas combinações, tendo como base a melhor das soluções anteriores. Possui uma complexidade de cálculo razoável, mas a solução encontrada não é, necessariamente, a melhor possível.

O particionamento de variáveis contínuas implica em uma maior complexidade de cálculo. Destacam os seguintes critérios:

a) Testes simples ou pesquisa exaustiva: citado por Fonseca (1994), este critério escolhe para a partição do nó o ponto de cisão considerado de melhor valor pelo critério de avaliação adotado, testando e avaliando todas as partições possíveis com base em cada uma das características. O ponto de cisão consiste em um teste binário com resultados $x_n \leq A$ e $x_n > A$, em que x_n é uma variável e A é um valor de limiar. Para encontrar este ponto de cisão, ou seja, o valor de A, primeiramente ordenam-se todos os valores da variável de forma crescente: $\{v_1, v_2, ..., v_n\}$. A seguir, o ponto médio de cada dois valores consecutivos, dado por $A = (v_i + v_{i+1})/2$, é um dos possíveis valores do ponto de cisão. Este valor, então será avaliado pela função de avaliação utilizada para a escolha do nó. O ponto de cisão que obtiver o melhor resultado será utilizado para efetuar a partição da variável contínua. Assim, para um problema com N exemplos e M características tem-se um total de $(N-1)*M$ partições possíveis.

b) Testes múltiplos: a definição de múltiplos intervalos de partição do valor de uma característica pode aumentar a capacidade discriminativa de cada medida, levando à construção de árvores de menores dimensões sem perda de desempenho, conforme citado por Fonseca (1994). Para definir estes múltiplos intervalos, uma das soluções possíveis é utilizar a segmentação global ou a segmentação a nível de nó.

A segmentação global, transforma o problema em atributos apenas discretos e utiliza a estratégia de *Bayes* para calcular a segmentação com base na qual as medidas contínuas serão discretizadas. A segmentação a nível de nó segmenta os valores contínuos em cada nó. Um exemplo deste algoritmo é apresentado em Moura-Pires (1991).

Fonseca (1994) destaca que a importância da segmentação das variáveis em cada nó está na quantificação das probabilidades marginais das características em cada nó, de modo a conhecer nos intervalos de segmentação, as diferentes situações resultantes das partições anteriormente efetuadas.

c) Combinação linear de características: a combinação linear de características permite contornar a limitação apresentada pelas árvores baseadas em testes efetuados sobre uma só característica. Segundo Baranauskas e Monard (2000), este método permite a criação de árvores multivariadas, combinando linearmente características em cada nó. Neste tipo de teste, o espaço de busca não é particionado em regiões retangulares e sim em hiperplanos.

Supondo o problema M dimensional, a pesquisa do hiperplano que fará a partição do conjunto de treino inicia-se com a geração aleatória de um conjunto de coeficientes $a = (a_1, a_2, ..., a_M)$, tal que se verifica:

$$\|a\|^2 = \sum_{i=1}^{M} a_i{}^2 = 1 \qquad (9)$$

Será então procurado o valor de $c - (c^*)$ que permitir a melhor partição[1].

$$\sum_{i=1}^{M} a_i x_i \leq c \qquad (10)$$

Denote-se a melhor partição assim encontrada como $s^*(a)$ e o ganho que ocorre da sua utilização no nó t por $\Delta(s^*(a^*), t)$. O melhor conjunto de coeficientes de $a - (a^*)$ será aquele que permita maximizar o valor de Δ, ou seja:

[1]Identificam-se com $*$ os valores que conduzem ao resultado considerado ótimo.

$$\Delta\left(s^*\left(a^*\right),t\right)=\max_a \Delta\left(s^*(a),t\right) \tag{11}$$

A melhor partição será, portanto, dada por:

$$\sum_{i=1}^{M} a_i^* x_i \le c^* \tag{12}$$

Para problemas envolvendo um número significativo de medidas, a árvore assim gerada apresenta grande complexidade envolvendo combinações lineares de todas as medidas em todos os nós. De modo a contornar este problema é sugerido o seguinte processo:

1º Passo. Para todas as medidas i com $i=1,...,M$ variar o patamar c e calcular a melhor partição do tipo:

$$\sum_{j=1, j\neq i}^{M} a_j^* x_j \le c_i \tag{13}$$

Calcular a qualidade da partição, ou seja, o decréscimo na função de avaliação (Δ_i) assim encontrada.

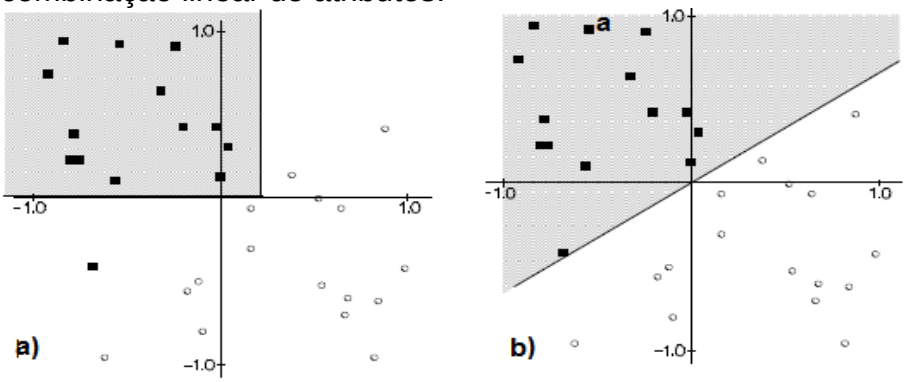
2º Passo. Calcular a degradação na solução proveniente da anulação da medida mais importante, ou seja, aquela que se retirada, conduz à pior solução implicando, portanto, um valor mínimo de Δ. Denote-se este valor por $\Delta^* - \min_i \Delta_i$. De uma forma equivalente, a degradação introduzida pelo retirar da medida menos importante será dada por $\Delta^* - \max_i \Delta_i$. Enquanto se verificar:

$$\Delta^* - \max_i \Delta_i < \beta\left(\Delta^* - \min_i \Delta_i\right) \tag{14}$$

retire a medida menos importante (β é uma constante pré-definida, de valor geralmente 0,1 ou 0,2).

3º Passo. Utilizar o algoritmo de pesquisa para encontrar a melhor combinação utilizando agora apenas as características não retiradas no passo anterior. Breiman et al. (1984) apresenta um algoritmo de pesquisa para os valores de a^* e c^*. Outros algoritmos também são propostos como o OC1, descrito em Murthy (1997). A Figura 2 mostra a vantagem de ser efetuada uma partição do espaço em um caso particular. A árvore gerada pelo algoritmo CART melhor generaliza a partição do espaço conseguida pela solução com combinação linear de atributos.

Figura 2 – Partições resultantes das árvores geradas pelo algoritmo CART a) sem e b) com combinação linear de atributos.



Fonte: Fonseca, 1994, p. 114.

2.3.3. Definição da folha e da classe

Estabelecida a partição das variáveis em cada nó, o próximo passo é verificar se o respectivo nó de partição será um nó terminal. Segundo Esposito et al (1997), a decisão para um nó terminal ou folha, pode ser realizada quando: 1) todos os exemplos atingem um nó pertencendo à mesma classe; 2) todos os exemplos atingem um nó com o mesmo vetor de características; 3) o número de exemplos em um nó é menor que certo percentual definido e; 4) o resultado da medida de qualidade do melhor atributo, para todos os possíveis testes que particionam o conjunto de observações é muito baixo.

A definição de qual classe atribuir a uma determinada folha pode ser realizada pela atribuição da classe de maior probabilidade ou pela determinação na noção de custos (FONSECA, 1994):

a) Atribuição da classe mais provável: é atribuída a classe mais frequente dentro dos exemplos que se encontram nesta folha:

$$\max_j(p_j) = \max_j \frac{n_j}{N} \qquad \text{com } j = 1,\ldots,k. \tag{15}$$

Em que: N é o número total de exemplos na folha; n_j é o número de exemplos da classe j.

b) Determinação baseada na noção de custos: é atribuída uma classe que, baseada na matriz de custo, minimiza os custos provenientes desta classificação:

$$Custo(j) = \sum_{i=1}^{k} p_i C_{i,j} \tag{16}$$

Em que: k é o número de classes; p_i é a probabilidade da classe i e; $C_{i,j}$ é o valor da linha i, coluna j da matriz de custos.

2.3.4. Limitação na dimensão das árvores de decisão

Segundo Fonseca (1994) uma das operações mais importantes para o desempenho das árvores de decisão é a limitação das suas dimensões, de acordo com o conhecimento alcançado do conjunto de treino.

As árvores de decisão estão propensas ao *overfitting*, que ocorre quando o modelo aprende detalhadamente os padrões ao invés de generalizar. Este fato, pode gerar árvores demasiadamente extensas, com desempenho real longe do ideal e pouco eficientes em termos de classificação, pois dando demasiada importância a casos particulares surgidos no conjunto de treino, prejudica a inferência da verdadeira estrutura do problema.

Para contornar o crescimento excessivo da árvore de decisão, pode-se substituir nós profundos, que fornecem pouco poder de previsão; por folhas, processo conhecido como poda da árvore. Esta técnica é realizada analisando a taxa de erro do nó e a taxa de erro que ocorre quando se poda o mesmo. A taxa de

erro- $E(T)$ de uma folha representa a razão entre o número de casos com classificação errada- ce e o número de casos classificados corretamente- cc pela partição:

$$E(T) = \frac{ce}{ce + cc} \tag{17}$$

A poda na árvore de decisão causa um erro de classificação em alguns exemplos do conjunto de treinamento. Porém, sua vantagem surge quando é feita a classificação de novos exemplos que não foram usados no processo de construção da árvore, conduzindo a um erro de generalização menor. Assim, o objetivo da poda é remover partes da árvore que não contribuem para a precisão da classificação, produzindo árvores menos complexas e, consequentemente, mais compreensíveis.

Segundo Gama (1999), os métodos de poda podem ser divididos em duas abordagens principais: pré-poda, quando critérios de parada no processo de indução são satisfeitos, ou seja, alguma condição é satisfeita, parando a geração da árvore; e pós-poda, que constrói toda a árvore para depois podá-la, reduzindo a mesma para dimensões ótimas.

Fonseca (1994), destaca alguns critérios de parada. As técnicas de pré-poda são descritas:

a) Parada baseada na informação mútua: se o ganho de informação obtido com o melhor atributo pelo método de entropia for inferior a um determinado δ, o nó é considerado folha e o crescimento da árvore de decisão finaliza.

$$Ganho(S, A) = E(S) - E(A) < \delta \tag{18}$$

Em que: δ é o parâmetro a ser ajustado.

b) Parada baseada no teste de independência: também conhecido como teste do qui-quadrado (χ^2), este método foi desenvolvido por Quinlan (1986). Baseia-se na presença de um conjunto de treinamento S e de uma característica A que possibilita a partição $\{S_1, S_2, ..., S_m\}$. Sendo p e n as frequências das classes P e

N no conjunto de treinamento e p_i e n_i as frequências da partição S_i. Assim, as frequências de p'_i e n'_i são dadas por:

$$\begin{cases} p'_i = p * \dfrac{p_i + n_i}{p + n} \\ n'_i = n * \dfrac{p_i + n_i}{p + n} \end{cases} \tag{19}$$

O valor aproximado do χ^2 com $n-1$ graus de liberdade pode ser usado para a determinação da confiança com que se rejeita a hipótese da característica A ser independente da classe dos componentes em S:

$$\chi^2 = \sum_{i=1}^{m} \frac{(p_i - p'_i)^2}{p'_i} + \frac{(n_i - n'_i)^2}{n'_i} \tag{20}$$

As técnicas de pós-poda consistem em calcular o erro de uma árvore e de todas as suas subárvores, examinando cada um dos nós não folhas da árvore, começando pelos nós mais próximos das folhas e subindo de forma *bottom-up*. Se a substituição do nó por uma folha ou pelo ramo mais frequente conduzir a um menor erro, é realizada a substituição. Dado que o erro da árvore global decresce sempre que o erro de qualquer das suas subárvores decresce, este processo conduz a uma árvore para o qual o erro é mínimo em relação a todas as técnicas de poda. Ademais, a poda da árvore reduz o número de nós internos e, consequentemente, a complexidade da árvore, enquanto ganha uma melhor performance em relação à árvore original (BARANAUSKAS; MONARD, 2000).

Os métodos de poda para árvores de decisão são uma das áreas mais pesquisadas no aprendizado de máquinas. Foram publicados vários estudos de métodos de indução para árvores de decisão (BRESLOW; AHA, 1997; ESPOSITO et al., 1997; SAFAVIAN; LANDGREBE, 1991; MURTHY, 1998), e todos eles discutem diferentes estratégias de poda.

Considere J_T ser o conjunto de nós internos (não terminais); L_T o conjunto de folhas em T e N_T o conjunto de nós de t. Assim, $N_T = J_T \cup L_T$. O ramo de

T que contém um nó t e todos os seus descendentes será indicado como T_T. O número de folhas de T vai ser indicado pela cardinalidade $L_T . |L_T|$. O número de exemplos de treinamento da classe i que compreende um nó t serão denotados por $n_i(t)$, o número total de exemplos em t, por $n(t)$ e o número de exemplos não pertencentes à classe majoritária por $e(t)$.

As técnicas de poda eliminam algumas partições da árvore de decisão de acordo com um dos seguintes critérios:

a) Poda de erro reduzido: proposto por Quinlan (1987), é conceitualmente mais simples. Utiliza um conjunto de poda para avaliar a eficácia de uma subárvore de T_{max}. O processo inicia com a árvore completa T_{max} e, para cada nó interno t de T_{max}, o algoritmo compara o número de erros de classificação no conjunto de poda e na subárvore T_t, quando t é transformado em uma folha e associado à melhor classe. Às vezes, a árvore simplificada tem um desempenho melhor do que a original. Neste caso, é aconselhável podar T_t. Esta operação da poda do ramo é repetida na árvore simplificada até que a poda adicional aumente a taxa de classificação errada.

Quinlan restringe a condição de poda dada acima com outra restrição: T_t pode ser podado apenas se não contém uma subárvore que resulte em uma taxa de erro menor do que a própria T_t. Isso significa que os nós a serem podados são examinados de acordo com uma estratégia de baixo para cima.

TEOREMA. O algoritmo encontra a menor versão da subárvore mais precisa em relação ao conjunto de poda.

PROVA. Seja T^* a árvore podada otimamente com respeito ao conjunto de poda, e t_0 seu nó raiz. Então, T^* é a raiz da árvore $\{t_0\}$ associada à classe mais prevalente, ou, se $t_1, t_2, ..., t_s$, são os filhos de t_0, T^* é a árvore enraizada em t_0 com subárvores $T_{t1}^*, T_{t2}^*, ..., T_{ts}^*$. A primeira parte da afirmação é evidente, enquanto a segunda parte é baseada na propriedade aditiva da taxa de erro para árvores de decisão, segundo a qual uma otimização local em cada ramo $T_{t_i}^*$ conduz a uma otimização global em T.

Outra propriedade positiva deste método é a sua complexidade computacional linear, uma vez que cada nó é visitado apenas uma vez para avaliar o momento da poda. Contudo, um problema com o algoritmo é o seu viés com respeito a poda excessiva. Este problema é particularmente notável quando o conjunto de poda é muito menor do que o conjunto de treinamento, mas torna-se menos relevante à medida que a porcentagem de casos no conjunto de poda aumenta.

b) Poda do erro pessimista: este método é caracterizado por usar o mesmo conjunto de treinamento no crescimento e poda de uma árvore. Todavia, como a taxa de erro aparente (do conjunto de treinamento) é otimista e não pode ser usada para escolher a melhor árvore podada, a poda do erro pessimista considera uma constante ao erro de treinamento de uma subárvore, assumindo que cada folha classifica automaticamente uma determinada fração de uma instância incorretamente. Essa fração é dividida em 1/2 pelo número total de casos cobertos pela folha e é chamado de correção de continuidade na estatística (WILD; WEBER, 1995). Nesse contexto, é usado para tornar a distribuição normal mais próxima da distribuição binomial no pequeno caso de amostra. Seja:

$$r(t) = e(t)/n(t) \tag{21}$$

Em que: $r(t)$ é a taxa de erro aparente em um único nó t quando o nó é podado, e:

$$r(T_t) = \frac{\sum_{s \in L_{T_t}} e(s)}{\sum_{s \in L_{T_t}} n(s)} \tag{22}$$

Em que: $r(T_t)$ a taxa de erro aparente para a subárvore inteira T_t. Então, a correção de continuidade para a distribuição binomial é dada por:

$$r'(t) = [e(t) + 1/2]/n(t) \tag{23}$$

Ao estender a aplicação da correção de continuidade à estimativa da taxa de erro de T_t, temos:

$$r'(T_t) = \frac{\sum_{s \in L_{T_t}} [e(s) + 1/2]}{\sum_{s \in L_{T_t}} n(s)} = \frac{\sum_{s \in L_{T_t}} e(s) + \frac{|L_{T_t}|}{2}}{\sum_{s \in L_{T_t}} n(s)} \quad (24)$$

Por simplicidade, o número de erros é dado por:

$$e'(t) = [e(t) + 1/2] \quad (25)$$

para um nó t, e:

$$e'(T_t) = \sum_{s \in L_{T_t}} e(s) + \frac{|L_{T_t}|}{2} \quad (26)$$

para uma subárvore T_t.

Deve-se observar que, quando uma árvore continua a desenvolver-se até que nenhuma das suas folhas cometa erros de conjunto de treinamento, então $e(s) = 0$ se s é uma folha. Como consequência, $e'(T)$ representa apenas uma medida de complexidade da árvore que associa cada folha com um custo igual a 1/2. Isso, porém não é verdade para árvores parcialmente podadas ou quando ocorrem discordâncias (observações iguais pertencentes a classes distintas) no conjunto de treinamento.

Como esperado, a subárvore T_t produz menos erros no conjunto de treinamento do que o nó t, quando t se torna uma folha, mas às vezes pode acontecer que $n'(t) \leq n'(T_t)$ devido à correção de continuidade, caso em que o nó t é podado. No entanto, isso raramente ocorre, uma vez que a estimativa $n'(T_t)$ do número de classificações erradas feitas pela subárvore ainda é bastante otimista. Por este motivo, Quinlan reduz a condição, exigindo que:

$$e'(t) \leq e'(T_t) + SE(e'(T_t)) \quad (27)$$

Onde

$$SE(e'(T_t)) = [e'(T_t) \cdot (n(t) - e'(T_t))/n(t)]^{1/2} \quad (28)$$

Em que: $SE(e'(T_t))$ é o erro padrão para a subárvore T_t, calculado considerando uma distribuição binomial para os erros, mesmo que a propriedade de independência dos eventos não seja considerada, porque T_{max} foi construída para se adequar aos dados de treinamento.

O algoritmo avalia cada nó a partir da raiz da árvore e, se um ramo T_t é podado, os descendentes de t não são examinados. Esta abordagem de cima para baixo confere à técnica de poda uma alta velocidade de execução.

A introdução da correção de continuidade na estimativa da taxa de erro não possui justificativa teórica. Na verdade, a correção de continuidade é útil apenas para introduzir um fator de complexidade da árvore. No entanto, esse fator é incorretamente comparado com uma taxa de erro, o que pode levar a uma poda reduzida ou ainda excessiva. Por certo, se T_{max} classifica corretamente todos os exemplos de treinamento, então:

$$e'(T_t) + SE(e'(T_t)) \approx \frac{1}{2}\left(|L_{T_t}| + \sqrt{|L_{T_t}|}\right) \quad (29)$$

e, desde $e'(t) \approx e(t)$, então o método irá podar se:

$$|L_{T_t}| + \sqrt{|L_{T_t}|} \geq 2e(t) \quad (30)$$

Ou seja, a poda ocorre se T_t tiver um número suficientemente alto de folhas com relação ao número de erros a cobrir. A constante 1/2 indica simplesmente a contribuição de uma folha para a complexidade da árvore.

Por fim, esse método possui uma complexidade linear no número de nós internos. Por certo, no pior caso, quando a árvore não precisa de poda, cada nó será visitado uma vez (ESPOSITO et al., 1997).

c) Poda do erro mínimo: Niblett e Bratko (1986) propuseram uma abordagem de poda ascendente buscando por uma única árvore que minimize a taxa de erro esperada em um conjunto de dados independente. Na verdade, tanto a versão original quanto a melhorada relatada em Cestnik e Bratko (1991) exploram apenas informações no conjunto de treinamento. No entanto, a implementação da versão melhorada requer um conjunto de poda independente pelos motivos explicados a seguir.

Para um problema de classe k, a probabilidade esperada de que uma observação que alcança o nó t pertença à *i-ésima* classe é a seguinte:

$$p_i(t) = \frac{n_i(t) + p_{ai} \cdot m}{n(t) + m} \tag{31}$$

Em que: p_{ai} é a probabilidade a priori da *i-ésima* classe, e m é um parâmetro que determina o impacto da probabilidade a priori na estimação da probabilidade a posteriori $p_i(t)$.

Por simplicidade, o parâmetro m é considerado igual para todas as classes. Cestnik e Bratko nomeiam $p_i(t)$ como a estimativa de probabilidade m. Quando uma nova observação alcançando t é classificada, a taxa de erro esperada é dada por:

$$\begin{aligned} EER(t) &= \min_i \{1 - p_i(t)\} \\ &= \min_i \{[n(t) - n_i(t) + (1 - p_{ai}) \cdot m] / [n(t) + m]\} \end{aligned} \tag{32}$$

Esta fórmula é uma generalização da taxa de erro esperada calculada por Niblett e Bratko (1986). Por certo, quando $m = k$ e $p_{ai} = 1/k, i = 1,2\ldots,k$, isto é, a distribuição de probabilidade a priori é uniforme e igual para todas as classes. Assim:

$$EER(t) = \min_i \{[n(t) - n_i(t) + k - 1]/[n(t) + k]\}$$
$$= [e(t) + k - 1]/n(t) + k \quad (33)$$

No método de poda de erro mínimo, a taxa de erro esperada para cada nó interno $t \in J_T$ é calculada. Isso é chamado de erro estático, STE (t). Então, a taxa de erro esperada de T_t, chamado erro dinâmico, DYE (t), é calculada como uma soma ponderada das taxas de erro esperadas dos seus descendentes.

Geralmente, quanto maior o m, mais severa é a poda. Quando m é infinito, $p_i(t) = p_{ai}$, e como p_{ai} é estimado como a porcentagem de exemplos da i-ésima classe no conjunto de treinamento, a árvore reduzida a uma única folha tem a menor taxa de erro esperada. No entanto, para m'> m, o algoritmo pode não retornar uma árvore menor que a obtida para um valor m. Esta propriedade de não-monotonicidade tem uma consequência grave em complexidade computacional: para aumentar o valor de m, o processo de poda deve sempre começar a partir de T_{max}.

Certamente, a escolha de m é crítica. Cestnik e Bratko sugerem a intervenção de um especialista que pode escolher o valor certo para m, de acordo com o nível de ruído nos dados ou mesmo estudar a seleção das árvores produzidas. Em estudos desenvolvidos para avaliar a acurácia da classificação das árvores podadas, a versão mais recente do algoritmo de poda de erro mínima parece ter superado dois problemas que afetaram a proposta original por Niblett e Bratko: viés otimista (WATKINS, 1987) e dependência da taxa de erro esperada no número de classes (MINGERS, 1989).

d) Poda de valor crítico: este método proposto por Mingers (1987) é uma técnica de baixo para cima, como a poda de erro reduzido. No entanto, faz decisões de poda de uma maneira fundamentalmente diferente. Considerando que a poda de erro reduzido usa o erro estimado nos dados de poda para julgar a qualidade de uma subárvore, a poda de valor crítico olha a informação coletada durante o crescimento da árvore. Um indutor de árvore de decisão de cima para baixo emprega recursivamente um critério de seleção para dividir os dados de treinamento em subconjuntos menores e mais puros. Em cada nó, ele se divide de forma a maximizar o valor do critério de divisão – por exemplo, o ganho de

informação. A poda de valor crítico usa esse valor para tomar decisões de poda. Quando uma subárvore é considerada para a poda, o valor do critério de divisão no nó correspondente é comparado a um limite fixo, e a árvore é substituída por uma folha se o valor for muito pequeno. No entanto, uma restrição adicional é imposta: se a subárvore contém pelo menos um nó cujo valor é maior do que o limite, ele não será podado. Isso significa que uma subárvore só é considerada para a poda se todos os seus sucessores forem nós de folha.

O desempenho da poda de valor crítico depende do limiar utilizado para as decisões de poda. Supondo que o valor do critério de divisão aumenta com a qualidade da divisão, limiares maiores resultam em podas mais agressivas. A maioria dos critérios de divisão não leva em conta o número de instâncias que suportam a subárvore. Consequentemente, o procedimento superestima a qualidade das subárvores que cobrem apenas algumas instâncias (FRANK, 2000).

e) Poda baseada no erro: este é o método de poda implementado em C4.5 (QUINLAN, 1993). Uma proposta similar também foi proposta por Kalkanis (1993). Como a poda de erro pessimista, ele extrai estimativas de erro dos dados de treinamento, assumindo que os erros são distribuídos em binômio. No entanto, em vez da regra de erro único padrão empregada pela poda de erro pessimista, o algoritmo calcula um intervalo de confiança nas contagens de erros, com base no fato de que a distribuição binomial é aproximada pela distribuição normal no caso de amostra grande. Então, o limite superior deste intervalo de confiança é usado para estimar a taxa de erro de uma folha em novos dados. Em C4.5, o intervalor de confiança é definido como 25% por padrão. Como a poda de erro reduzido – e em contraste com a poda de erro pessimista – uma estratégia de poda ascendente é empregada; uma subárvore é considerada para substituição por uma folha depois que todos os seus ramos já foram considerados para poda. A substituição é realizada se a estimativa de erro para a folha prospectiva não for maior que a soma das estimativas de erro para os nós de folhas atuais da subárvore.

O uso de um intervalo de confiança é uma maneira heurística de reduzir o viés otimista na estimativa de erro derivada dos dados de treinamento, mas não é estatisticamente correto. Do ponto de vista estatístico, este procedimento de poda compartilha os problemas de poda de erro pessimista. O uso da suposição de normalidade também é questionável porque está correto apenas no limite. Para

pequenas amostras com menos de 100 casos, os estatísticos utilizam a distribuição de Student, em vez da distribuição normal (WILD; WEBER, 1995). Na indução da árvore de decisão, pequenas amostras são exatamente aquelas que provavelmente serão relevantes no processo de poda.

f) Poda por minimização do custo-complexidade: desenvolvido por Breiman et al (1984) foi introduzida no sistema clássico CART e baseia-se nos seguintes passos: **1)** Seleção de uma família paramétrica de subárvores de T_{max}, $\{T_0, T_1, T_2, ..., T_L\}$ de acordo com algumas heurísticas; **2)** Escolha da melhor árvore T_i de acordo com uma estimativa das taxas de erro verdadeiro das árvores na família paramétrica.
No que diz respeito ao primeiro passo, a ideia básica é que T_{i+1} é obtido a partir de T_i pela poda dos ramos que mostram o menor aumento na taxa de erro aparente por folha podada. Por certo, quando uma árvore T é podada no nó t, sua taxa de erro aparente aumenta pela quantidade $r(t) - r(T_t)$, enquanto o número de folhas diminui por $|L_{T_t}| - 1$ unidades. Assim, a seguinte relação

$$\alpha = (r(t) - r(T_t)) / (|L_{T_t}| - 1) \tag{34}$$

mede o aumento da taxa de erro aparente por folha podada. Então, T_{i+1} na família paramétrica é obtido pela poda de todos os nós em T_i com o menor valor de α. A primeira árvore T_0 é obtida pela poda T_{max} dos ramos cujo valor α é 0, enquanto a última árvore T_L é a árvore raiz. É possível provar que cada árvore T_i é caracterizada por um valor distinto α_i, tal que $\alpha_i < \alpha_{i+1}$. Portanto, o conjunto $\{T_0, T_1, T_2, ..., T_L\}$ é uma família paramétrica de árvores que denotaremos como $T_{max}(\alpha)$. A família paramétrica pode ser construída em um tempo quadrático no número de nós internos.

Na segunda fase, a melhor árvore em $T_{max}(\alpha)$ com relação à precisão preditiva é escolhida. Os autores propõem duas formas distintas de estimar a taxa de erro verdadeira de cada árvore na família, uma baseada em um conjunto de poda independente ou a validação cruzada. A validação cruzada apresenta o

problema adicional de relacionar os valores de α_j^k observados no conjunto de treinamento k para os valores de α_i da sequência original de árvores, pois esses valores geralmente são diferentes. O algoritmo CART resolve este problema calculando a média geométrica α_i^{av} de α_i e α_{i+1} para a árvore i da sequência original. Então, cada conjunto k da validação cruzada, o algoritmo escolhe a árvore que exibe o maior valor α_j^k menor do que α_i^{av}. A média das estimativas de erro para essas árvores é a estimativa de validação cruzada para a árvore i.

Algumas comparações dos métodos de pós-poda foram realizadas. Segundo Gama (1999), os métodos baseados no custo-complexidade e na redução do erro são bons, enquanto os outros podem causar eventuais problemas.

A técnica de pós-poda é a abordagem mais utilizada e confiável, porém, promove um processo mais lento, enquanto que a pré-poda tem a vantagem de não demandar tempo na construção de uma estrutura que não será utilizada no final da árvore. Nem sempre a árvore podada é mais precisa que a correspondente gerada, mas a poda permite simplificar a árvore de decisão, essencial em árvores complexas.

2.3.5. Qualidade do classificador

Para que o resultado da mineração de dados seja utilizado com segurança, o classificador deve obter o menor erro de classificação aceitável. Para tal, faz-se necessário utilizar métricas de avaliação para estimar este valor em função dos exemplos do conjunto de treinamento ou em dados ainda não apresentados.

A estimativa da qualidade de um classificador incide em estimar a percentagem de erro que se espera que o classificador venha a obter na classificação de exemplos futuros. De forma abrangente, a percentagem de erro pode ser demonstrada pela relação do número de erros e o número de casos testados. Assim, a medida mais utilizada é a taxa de erro de um classificador, também conhecida como taxa de classificação incorreta. Já o complemento da taxa de erro conhecida como acurácia ou precisão do classificador, nos informa a proporção de objetos corretamente classificados e determina quão bom um modelo será para dados não apresentados no conjunto de exemplos.

As medidas de desempenho de um classificador efetuadas sobre o conjunto de treinamento são comumente designadas como aparentes e as medidas realizadas sobre o conjunto de teste são conhecidas como reais ou verdadeiras (BARANAUSKAS; MONARD, 2001).

Um problema que se coloca na estimação da qualidade de um classificador está nos custos do erro. As técnicas de estimação de erro e de custos são apresentadas a seguir:

a) Estimação de custos: o custo de um erro pode ser considerado como a penalização imposta ao sistema, no caso deste cometer um dado tipo de erro de classificação. Se o objetivo do sistema for a minimização dos custos em lugar da minimização do erro faz-se necessário definir as penalizações a atribuir. É normalmente utilizada para este efeito a chamada matriz de custos com os N^2 elementos da matriz onde N representa o número de classes. Para o cálculo do desempenho do classificador em termos de custos considera-se a matriz de confusão que descreve os resultados do teste do classificador em termos do número de exemplos da classe i aos quais foi atribuída a classe j com i e j variando entre 1 a N. Portanto, o custo do classificador é dado por:

$$Custo = \sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j} M_{i,j} \qquad (35)$$

Em que: $C_{i,j}$ é o valor da linha i, coluna j da matriz de custos e; $M_{i,j}$ é o valor da linha i, i, coluna j da matriz de confusão.

b) Estimação por resubstituição ou erro aparente: desenvolvido por Breiman et al (1984), o método estima a proporção de resultados incorretos quando o conjunto de treinamento é novamente apresentado para a classificação, depois que o classificador foi construído. Ou seja, o conjunto de treinamento e de teste são idênticos. Assim, a proporção dos resultados incorretos nos informará o valor estimado para o erro de classificação. O erro de resubstituição pode, portanto, ser demonstrado por:

$$R(d) = \frac{1}{N}\sum_{i=1}^{N} X(d(x_n) \neq j_n) \quad (36)$$

Em que: N é o número de exemplos; x_n é o exemplo n; j_n é a classe correspondente ao exemplo n; $d(x)$ é a resposta do classificador diante do exemplo x; $X(A) =$ 1 se A é verdadeiro e 0 se A é falso.

Dado que é usado para a estimação do erro o mesmo conjunto de exemplos para o cálculo do classificador, pode-se destacar que, como qualquer classificador é construído com vista a minimizar o erro obtido com o conjunto de treino, este método de estimação é otimista.

Como exemplo, suponha a situação em que $d(x)$ constitui uma partição $A_1,...,A_j$ tal que A_j contenha todos os exemplos para os quais $j_n = j$ e que os vetores não presentes no conjunto de treino estejam distribuídos aleatoriamente por todas as classes. Por certo, neste caso tem-se que $R(d) = 0$, que estará certamente distante do valor real $R^*(d)$.

c) Estimação por utilização de um conjunto independente[2]: neste método, o conjunto de exemplos t é dividido em dois subconjuntos t_1 e t_2, respectivamente para o treino do classificador e para a estimação do erro $R^*(d)$ (BREIMAN et al., 1984).

Sendo N_2 o número de exemplos em t_2, temos para este tipo de estimação a taxa do erro, denotada por $R^{ts}(d)$:

$$R^{ts}(d) = \frac{1}{N_2}\sum_{(x_n, j_n)\in t_2} X(d(x_n) \neq j_n) \quad (37)$$

Como preceito, os exemplos de t_2 devem ser independentes dos casos de t_1 e seguir, aproximadamente a mesma distribuição de probabilidade. Normalmente, o conjunto de treinamento t_1 é escolhido de forma aleatória contendo 2/3 dos exemplos de t e o conjunto de teste t_2, 1/3 dos exemplos restantes.

[2] Do inglês *test sample estimation*.

Em situações em que a dimensão do conjunto de exemplos é reduzida, o método de validação cruzada ou por camadas é considerado preferencial.

d) Estimação por camadas ou validação cruzada[3]: na estimação por camadas, os exemplos do conjunto de treinamento t é dividido randomicamente em V subconjuntos $t_1,...,t_V$ contendo, aproximadamente, o mesmo número de elementos (BREIMAN et al., 1984).

Para v=1,...,V constrói-se o classificador $d^V(X)$ utilizando $t - t_v$, ou seja, todos os exemplos menos os que estão no subconjunto v e usa-se este subconjunto para estimar o erro de classificação. O algoritmo gera um modelo do conjunto de treinamento e utiliza o conjunto de teste para classificar os exemplos. Realiza-se, portanto, uma estimação por utilizar um conjunto independente para cada um dos classificadores parciais. O valor estimado para o erro do classificador final será então, a média dos erros estimados a cada um dos V classificadores parciais calculados.

Temos assim, que a estimação do erro $R^{ts}(d^v)$ para cada um dos V classificadores é dada por:

$$R^{ts}(d^v) = \frac{1}{N_v} \sum_{(x_n, j_n) \in t_V} X(d^v(x_n) \neq j_n) \tag{38}$$

Em que: $N_v \approx \frac{N}{V}$

O erro estimado para o classificador será então:

$$R^{cv}(d) = \frac{1}{V} \sum_{v=1}^{V} R^{ts}(d^v) \tag{39}$$

Este método de medida de desempenho de um classificador também é designado por camadas estratificadas (WEISS, 1990), quando a seleção dos

[3] Do inglês *cross-validation*.

exemplos é efetuada com a preocupação de manter a distribuição percentual das várias classes nas diversas partições de treinamento e teste.

Geralmente, o conjunto inicial de treinamento é dividido em dez subconjuntos. Assim, todos os exemplos são utilizados nove vezes para a geração de um classificador e uma vez para o teste da sua eficiência com resultados bastante realistas (FONSECA, 1994).

f) Estimativa por *bootstrapping*: a técnica de estimação por *bootstrapping* é indicada para os casos em que o conjunto de treino é pequeno, devido sua complexidade de cálculo. A variante mais comum é conhecida por e0 (WEISS, 1990). Para a estimação por *bootstrapping* e0 é formado um novo conjunto de treino a partir do conjunto de treino original, amostrando *n* vezes com reposição os *n* exemplos do conjunto original. Os exemplos repetidos são eliminados e a percentagem esperada de exemplos no novo conjunto de treino obtido é de 63,2% do número de exemplos do conjunto original. Os casos que não se encontrarem no conjunto de treinamento constituirão o conjunto de teste. O erro estimado para o classificador será a média das várias iterações deste algoritmo.

2.4. Algoritmo CART

Proposto por Breiman et al. (1984), o algoritmo *Classification and Regression Trees* (CART) é usado para prever variáveis dependentes contínuas (regressão) e categóricas (classificação) por meio do particionamento recursivo do espaço de variáveis de transição.

A metodologia CART tem como principal atrativo a interpretabilidade proporcionada pela estrutura de árvore de decisão obtida no modelo final que, também pode ser lido, como um conjunto de sentenças lógicas a respeito das variáveis explicativas. Esta é uma característica importante, pois permite ao analista acessar prontamente o que o modelo está fazendo e tomar decisões com mais clareza.

2.4.1. Modelo estruturado em árvore

Quando o espaço de saída é um conjunto finito de valores, como na classificação onde $y = \{c_1, c_2, ..., c_J\}$, outra maneira de olhar para um problema de aprendizagem supervisionada é notar que Y define uma partição sobre o universo Ω, isto é:

$$\Omega = \Omega_{C_1} \cup \Omega_{C_2} \cup ... \cup \Omega_{C_J}, \quad (40)$$

Em que: Ω_{C_K} é o conjunto de objetos para os quais Y tem valor c_K. Da mesma forma, um classificador φ pode ser também considerado como uma partição do universo Ω, uma vez que define uma aproximação $\hat{Y}$ de Y. Esta partição, entretanto, é definida no espaço de entrada X, diretamente em Ω, isto é:

$$X = X^{\varphi}{}_{C_1} \cup X^{\varphi}{}_{C_2} \cup ... \cup X^{\varphi}{}_{C_J}, \quad (41)$$

Em que: $X^{\varphi}{}_{C_K}$ é o conjunto de vetores de descrição $x \in X$ tal que $\varphi(x) = c_K$. Consequentemente, o aprendizado do classificador pode ser reformulado como uma partição de X, combinando o mais próximo da melhor partição dado pelo modelo de *Bayes* φ_B sobre X:

$$X = X^{\varphi_B}{}_{C_1} \cup X^{\varphi_B}{}_{C_2} \cup ... \cup X^{\varphi_B}{}_{C_J} \quad (42)$$

Do ponto de vista geométrico, o princípio dos modelos estruturados em árvores é simples. Consiste na aproximação da partição do modelo Bayes, dividindo recursivamente o espaço de entrada X em subespaços e atribuir valores de previsão constante $\hat{y} \in y$ a todos objetos x dentro de cada subespaço terminal.

Nesses termos, um modelo estruturado em árvore (ou árvore de decisão) pode ser definido como um modelo $\varphi : X \to y$ representado por uma árvore enraizada, onde qualquer nó t representa um subespaço $X_t \subseteq X$ do espaço de

entrada, com o nó raiz t_0 correspondente ao próprio X. Os nós internos t são rotulados com uma divisão s_t tirado de um conjunto de questões Q. Ele divide o espaço X_t que o nó t representa em subespaços disjuntos, correspondendo respectivamente a cada um de seus filhos.

Algoritmo 1. Predição do valor de saída $\hat{y} = \varphi(x)$ na árvore de decisão.

1: function Predict (φ, x)

2: $t = t_0$

3: while t is not a terminal node do

4: t= the child node t' *of t such* $x \in X_t$,

5: and while

6: return $\hat{y}_t$

7: end function

2.4.2. Indução da árvore de decisão

O aprendizado de uma árvore de decisão ideal equivale a determinar a estrutura da árvore que produz a partição mais próxima da partidão delineada por Y sobre X. A construção de uma árvore de decisão é conduzida com o objetivo de encontrar um modelo que divide o conjunto de aprendizado L. Todavia, entre todas as árvores de decisão $\varphi \in H$, podem existir várias delas que explicam L igualmente melhor. Blumer et al. (1987) destaca que a solução mais simples que se adapta aos dados no conjunto de aprendizado L é usada para encontrar a menor árvore φ^* (em termos de nós internos) minimizando sua estimativa de resubstituição $\overline{E}(\varphi^*, L)$.

Como mostrado por Hyafil e Rivest (1976) encontrar a menor árvore φ^* que minimiza a estimativa de resubstituição é um problema NP-completo. Como consequência, sob o pressuposto de que $P \neq NP$, não existe um algoritmo eficiente para encontrar φ^*, sugerindo assim que encontrar heurísticas eficientes para a construção de árvore de decisão quase ótimas é a melhor solução para manter os requisitos de computação dentro de limites realistas.

Seguindo a abordagem de Breiman et al. (1984), considere uma medida de impureza $i(t)$ como uma função que avalia a bondade de qualquer nó t, sendo o menor $i(t)$, o nó mais puro e a melhor predição $\hat{y}_t(x)$ para todo $x \in L_t$, onde L_t é o subconjunto de amostras de aprendizado em t, para todo $(x,y) \in L$ tal que $x \in X_t$. Partindo de um único nó que representa todo o conjunto de aprendizado L, o algoritmo ganancioso cresce iterativamente dividindo os nós em nós mais puros, ou seja, o algoritmo divide iterativamente em subconjunto menores, os subconjuntos de L representados pelos nós, até que todos os nós terminais não possam ser mais puros, garantindo assim previsões quase ótimas sobre L. A hipótese gananciosa busca uma boa generalização para dividir cada nó t usando a divisão $s*$ que maximiza localmente a diminuição da impureza dos nós filhos resultantes.

Formalmente, a diminuição da impureza de uma divisão binária s é definida como segue:

Definição 1. A diminuição da impureza de uma divisão binária $s \in Q$ dividindo o nó t em um nó esquerdo t_E e um nó direito t_D é:

$$\Delta i(s,t) = i(t) - p_E i(t_E) - p_D i(t_D) \tag{43}$$

Em que: p_E e p_D é a proporção $\frac{N_{t_E}}{N_t}$ e $\frac{N_{t_D}}{N_t}$ da amostra L_t ir para o nó filho esquerdo t_E e nó filho direito t_D, respectivamente e N_t é o tamanho do subconjunto L_t.

Com base neste conceito, o processo geral para a indução da árvore é descrito no algoritmo 2.

Algoritmo 2. Indução gananciosa de uma árvore de decisão binária.

1: function BuildDecisionTree (L)

2: Create a decision tree φ with root node t_0

3: Create an empty stack S of open nodes (t, L_t)

4: S.PUSH (t_o, L)

5: while S is not empty do

6: $t, L_t = S.POP(\ )$

7: *if the stopping criterion is met for* t *then*

8: $\hat{y}_t$ = *some constante value*

9: *else*

10: *Find the split* L_t *that maximizes impurity decrease*

$$s^* = \arg\max_{s \in Q} \Delta i(s,t)$$

11: *Partition* L_t *into* $L_{t_E} \cup L_{t_D}$ *according to* s^*

12: *Create the left child node* t_E of t

13: *Create the right child node* t_D of t

14: $S.PUSH\ ((t_D, L_{t_D}))$

15: $S.PUSH\ ((t_E, L_{t_E}))$

16: *end if*

17: *end while*

18: *return* φ

19: *end function*

2.4.3. Regras de atribuição do nó terminal

Quando um nó t é declarado terminal dado algum critério de parada o próximo passo (linha 8 do algoritmo 2) no procedimento de indução é rotular t com um valor constante $\hat{y}_t$ para ser usado como uma predição da variável de saída *Y*. Para uma árvore de estrutura fixa, minimizando o erro global de generalização é estritamente equivalente a minimizar o erro de generalização local de cada simples modelo no nó terminal. De fato,

$$Err(\varphi) = E_{X,Y}\{L(Y, \varphi(X))\}$$

$$= \sum_{t \in \tilde{\varphi}} P(X \in X_t) E_{X,Y|t}\{L(Y, \hat{y}_t)\} \qquad (44)$$

Em que: $\tilde{\varphi}$ indica o conjunto de nós terminais em φ e onde a expectativa interna[4] é o erro de generalização local do modelo no nó t. Assim, um modelo que minimiza $Err(\varphi)$ é um modelo que minimiza a expectativa interna na folha. O melhor aprendizado possível da árvore de decisão (de estrutura fixa), portanto, equivale a encontrar a melhor constante $\hat{y}_t$ em cada nó terminal.

2.4.4. Regras de divisão

Supondo que o critério de parada não seja cumprido (LOUPPE, 2014),deve-se concentrar no problema de encontrar $s^* \in Q$ de t que maximiza a diminuição da impureza $\Delta i(s^*,t)$ (Linha 10 do algoritmo 2).

Supondo valores de entrada distintos para todas as amostras N_t, o número de partições de L_t em subconjuntos k não vazios é apresentado em Knuth (1992):

$$S(N_t,k) = \frac{1}{K!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^{N_t}\ , \tag{45}$$

que se reduz a $2^{N_t-1}-1$ para partições binárias. Dado o crescimento exponencial no número de partições, a estratégia de enumerar todas as partições e escolher o melhor delas é, muitas vezes, computacionalmente intratável. Por esta razão, os pressupostos simplificadores deve ser feito na melhor divisão s^* Mais especificamente, o algoritmo de indução geralmente assume que s^*- ou pelo menos uma aproximação suficientemente boa – está em uma família $Q \subseteq S$ das divisões candidatas de estrutura restrita.

A família usual Q de divisões é o conjunto de divisões binárias definidas em uma variável única e que resulta em subconjuntos não vazios de L_t:

$$Q = \{s | s \in \bigcup_{j=1}^{p} Q(X_j), L_{t_E} \neq \phi, L_{t_D} \neq \phi\} \tag{46}$$

[4] A expectativa conjunta de X e Y é assumida em todos os objetos $i \in Q$ tal que $x_i \in X_t$.

Do ponto de vista geométrico, as divisões desta forma dividem o espaço de entrada X com hiperplanos, como anteriormente ilustrado na Figura 2. Consequentemente, o limiar de decisão corresponde assim à distância do hiperplano separador da origem.

Dado a decomposição 46 nos subconjuntos $Q(X_j)$, a divisão $s^* \in Q$ é a melhor divisão definida em cada variável de entrada. Isto é:

$$s^* = \underset{s_j^*}{arg\,max}\, \Delta i(s_j^*, t), \qquad j = 1,...,p \tag{47}$$

$$s_j^* = \underset{\substack{s \in Q(X_j) \\ L_{T_E}, L_{T_D} \neq \phi}}{arg\,max}\, \Delta i(s_j^*, t) \tag{48}$$

Nessa estrutura, o cerne do problema, portanto, se reduz à implementação da Equação 9.

Em CART para avaliar a medida de impureza $i(t)$ é comumente utilizado o critério de Gini e a entropia de Shannon para as árvores de classificação. Na regressão (quando a variável de saída Y é quantitativa), a função de impureza $i_R(t)$ com base na estimativa de resubstituição local é definida pelo menor desvio quadrado:

$$i_R(t) = \frac{1}{N_t} \sum_{x,y \in L_T} (y - \hat{y}_t)^2 \tag{49}$$

É possivel notar que a Equação 49 corresponde à variância interna do nó do valor de saída em t. Consequentemente, s^* é a divisão que maximiza a redução da variância $\Delta i(s,t)$ nos nós filhos.

2.4.5. Encontrando a melhor divisão binária

Para uma especificação completa do algoritmo 2 descreve-se a família Q de regras de divisão e critérios de impurezas como um procedimento de otimização eficiente para encontrar a melhor divisão $s^* \in Q$. Assumindo que Q seja o conjunto de divisões binárias univariadas, demonstrou-se que s^* é a melhor das melhores

divisões binárias s_j^* definida em cada variável de entrada. Isso leva ao seguinte procedimento:

Algortimo 3. Encontre a melhor divisão $s*$ que particiona L_t.

1: function FindBestSplit (L_t)

2: $\Delta = -\infty$

3: for $j = 1,...,p$ *do*

4: Find the best binary split s_j^* *defined on* X_j

5: if $\Delta i(s_j^*, t) > \Delta$ *then*

6: $\Delta = \Delta i(s_j^*, t)$

7: $s = s_j^*$

8: end if

9: end for

10: return $s*$

11: end function

Para uma discussão da linha 4 do algoritmo 3, considere X_j ser uma variável ordenada e $Q(X_j)$ um conjunto de todos as partições binárias não cruzadas de X_j. Considere também $X_{j|L_T} = \{x_j | x, j \in L_t\}$ indicar o conjunto de valores únicos de X_j dentro das amostras de nó em t. A melhor divisão $s_j^v \in Q(X_j)$ em X_j é a melhor partição de L_t em dois subconjuntos não vazios:

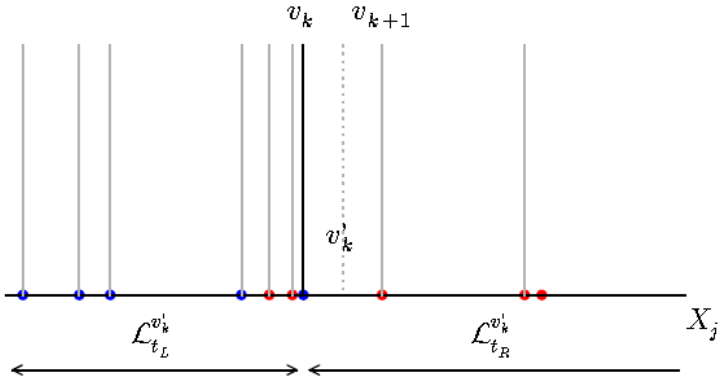
$$L_{t_E}^v = \{x, y) | (x, y) \in L_t, x_j \leq v\} \quad (50)$$

$$L_{t_D}^v = \{x, y) | (x, y) \in L_t, x_j \leq v\} \quad (51)$$

Em que: v é o limiar de decisão da divisão. Como ilustrado na Figura 3, existem $\left|X_{j|L_T}\right| - 1$ partições de L_t em dois desses subconjuntos não vazios, ou seja, um para cada valor $v_k \in X_{j|L_t}$, exceto para o último que leva a uma partição inválida.

Em particular, se X_j for localmente constante em t (isto é, se todos os pontos se sobrepõem na Figura 3), então $\left|X_{j|L_T}\right| = 1$ e X_j não pode ser usado para partição t. Mais importante ainda, geralmente, existem vários limiares v produzindo a mesma partição das amostras do nó. Se v_k e v_{k+1} são dois valores imediatamente consecutivos em $X_{j|L_t}$, então todas as divisões s_j^v para $v \in [v^k, v^{k+1}[$ produzem a mesma partição de L_t como $s_j^{v_k}$. Em termos de diminuição da impureza Δi, todos eles são equivalentes quando avaliados no L_t. Na generalização, no entanto, essas divisões podem não ser estritamente iguais, uma vez que não produzem a mesma partição de X_t. Como ajuste, os limiares de decisão que são assim escolhidos são os pontos intermediários $v_k^{'} = \frac{v_k + v_{k+1}}{2}$ entre valores consecutivos da variável, como mostrado pela linha pontilhada na Figura 3. Na prática, esta é uma boa heurística na maioria dos problemas (LOUPPE, 2014).

Figura 3 – Partição binária do L_t na variável ordenada X_j. Definido os limiares de decisão v para qualquer valor em $[v^k, v^{k+1}[$ que produz partições idênticas de L_t mas não de X_t.



Fonte: Louppe, 2014, p. 48.

Neste contexto, a melhor divisão s_j^* é a divisão $s_j^{v_k^{'}}$ que maximiza a diminuição da impureza. Computacionalmente, a valiação exaustiva de todas essas divisões pode ser realizada de forma eficiente ao considerar o limiar de decisão $v_k^{'}$ ordenado, observando que $\Delta i(s_j^{v_{k+1}^{'}}, t)$ pode ser calculado a partir de $\Delta i(s_j^{v_k^{'}}, t)$, em

várias operações linearmente proporcionais o número de amostras que vão do filho direito para o filho esquerdo. Como tal, a avaliação exaustiva de todas as divisões intermediárias pode ser realizada em tempo linear em relação a N_t, garantindo um bom desempenho.

A partir da partição inicial $v_0^{'} = -\infty$, onde t_E é vazio e t_D corresponde a t, e usando as equações de atualização para mudar $v_k^{'}$ para $v_{k+1}^{'}$, a busca da melhor divisão s_j^* em X_j pode finalmente ser implementado conforme descrito no algoritmo 4 e ilustrado na Figura 4.

Algortimo 4. Encontre a melhor divisão s_j^* que particiona L_t.

1: function FindBestSplit (L_t, X_j)

2: $\Delta = 0$

3: $k = 0$

4. $v_k^{'} = -\infty$

5: Compute the necessary statistics for $i(t)$

6: Initialize the statistics for t_E *to 0*

7: Initialize the statistics for t_D *to those of* $i(t)$

8: Sort the node samples L_t *such that* $x_{1,j} \leq x_{2,j} \leq ... \leq x_{N_t,j}$

9: $i = 1$

10: while $i \leq N_t$ *do*

11: while $i + 1 \leq N_t$ *and* $x_{i+1,j} = x_{i,j}$ *do*

12: $i = i + 1$

13: end while

14: $i = i + 1$

15: if $i \leq N_t$ *then*

16: $v_{k+1}^{'} = \frac{v_{i,j} + v_{i-1,j}}{2}$

17: Uptade the necessary statistics from $v_k^{'}$ *to* $v_{k+1}^{'}$

18: if $\Delta i(s_j^{v_{k+1}^{'}}, t) > \Delta$ *then*

19 $\Delta = \Delta i(s_j^{v_{k+1}^{'}}, t)$

20: $s_j^* = s_j^{v'_{k+1}}$

21: end if

22: $k = k + 1$

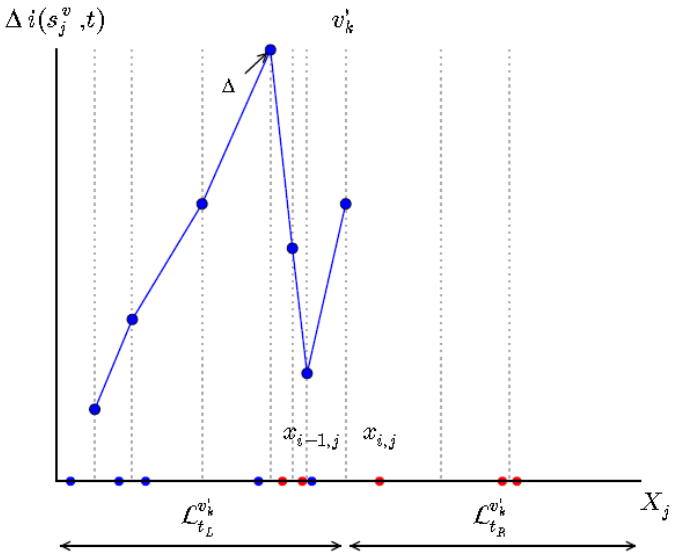
23: end if

24: end while

25: return s_j^*

26: end function

Figura 4 – Invariante do algoritmo 4. No final de cada iteração $\Delta i(s_j^{v'_k}, t)$ foi calculado a partir das estatísticas da divisão anterior em v'_{k-1} e comparado com a melhor redução na impureza Δ encontrada dentro das divisões em $v'_0 = -\infty$ para v'_{k-1}.



Fonte: Louppe, 2014, p. 51.

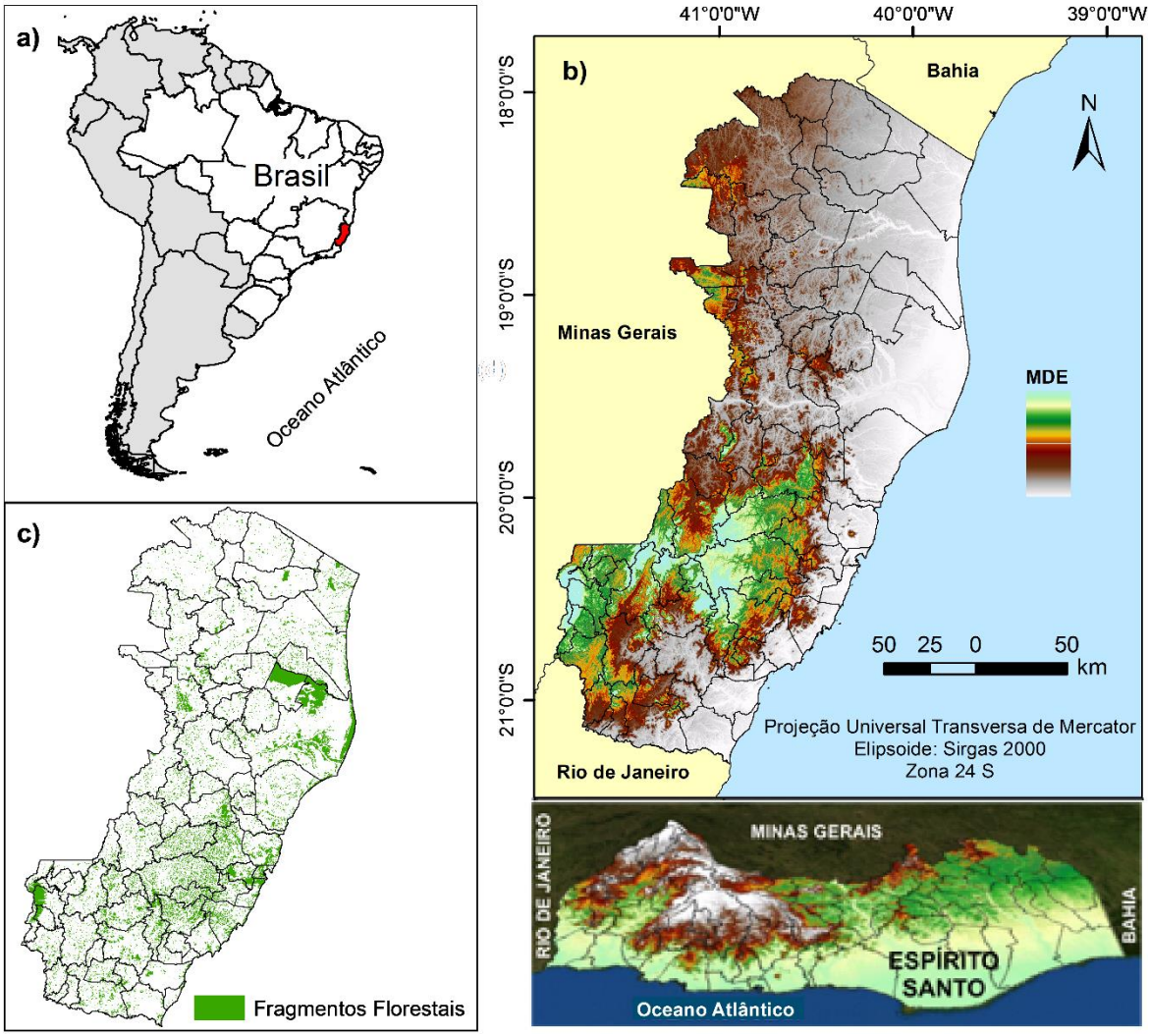
3. MATERIAIS E MÉTODOS

3.1. Área de estudo

A área de estudo é representada pelo estado do Espírito Santo, localizado na região Sudeste do Brasil (Figura 5 a), com área territorial de 46.052,64km². Está localizado entre 17°53'29" e 21°18'03" de latitude S e 39° 41'18" e 41°52'45" de longitude W. Faz fronteira com o Oceano Atlântico a Leste, estado da Bahia ao Norte, estado de Minas Gerais a Oeste e estado do Rio de Janeiro ao Sul. Em função de sua posição geográfica e geomorfologia (Figura 5 b), o estado apresenta quatro tipos de clima de acordo com a classificação de Köppen: Cwb, clima subtropical de altitude, com inverno seco e verão ameno encontrado na região montanhosa do estado; Cwa, clima subtropical de inverno seco e verão quente encontrado na região Sudoeste do estado, Am, clima tropical úmido ou sub-úmido encontrado na região Nordeste do estado, e Aw, clima tropical, com inverno seco encontrado na região Oeste do estado. Em virtude dos constantes desmatamentos e queimadas, houve a substituição de áreas de floresta natural por outras formas de uso da terra. Segundo relatório da Fundação SOS Mata Atlântica e do INPE, com os desmatamentos analisados entre 2013 e 2014, a área de estudo apresenta atualmente apenas 10,5% de seus remanescentes de floresta (Figura 5 c).

As seções a seguir descrevem a preparação do conjunto de dados e toda a metodologia utilizada. Em primeiro lugar, a densidade *kernel* e o mapa de ocorrência de incêndios são explicados. Em seguida, a teoria CART e as necessidades para a sua implementação são ilustradas. Por último, o mapa de predição de fogo é descrito prestando uma particular atenção à forma como as variáveis preditoras são implementadas. Toda a abordagem metodológica é descrita e resumida na Figura 6. O fluxograma mostra os principais procedimentos envolvidos no processo, apontando as principais etapas para determinar o mapa de densidade de fogo e predição de fogo.

Figura 5 – Localização geográfica da área de estudo (a); Modelo digital de elevação do estado do Esírito Santo (b); Remanescentes de floresta no estado (c).

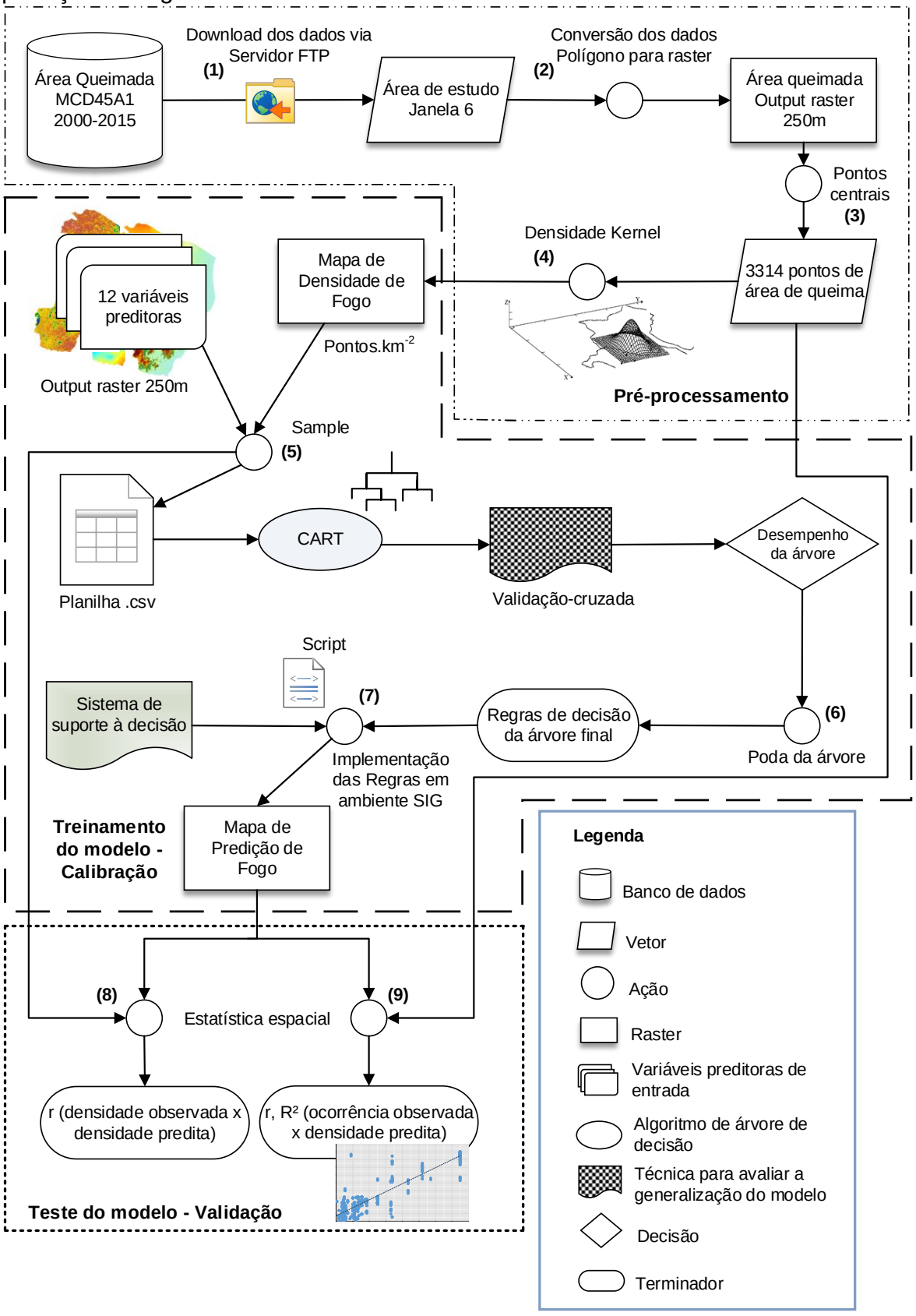


Fonte: o autor.

3.2. Conjunto de dados

Os subconjuntos mensais do produto MCD45A1 em formato *Shapefile* para o período de estudo (2000-2015), foram baixados via servidor *File Transfer Protocol* (FTP) do site *http://modis-fire.umd.edu/* pelo *software SmartFTP* (1 na Figura 6). Os arquivos estão disponíveis em projeção sinusoidal Lat-Long e extensão geográfica em janelas subcontinentais, que para a área de estudo é delimitada pela janela 6, cobrindo a América do Sul Central com latitude S entre 10° e 35° e longitude W entre 34° e 79°. Os mapas mensais de área queimada

Figura 6 – Visão geral dos principais procedimentos envolvidos no processo, apontando as principais etapas para determinar o mapa de densidade de fogo e predição de fogo.



Fonte: o autor.

foram então reprojetados para a projeção Universal Transversal de Mercador (UTM), *datum* SIRGAS 2000 e convertidos em formato raster (2 na Figura 6) com resolução espacial de 250m.

3.3. Densidade *kernel*

Para espacializar os dados de área queimada, considerou-se os pontos centrais (3 na Figura 6) de cada pixel da imagem raster, somando 3314 pontos de área de queima. As técnicas de interpolação, como um método para prever valores de atributos em locais não amostrados, a partir de observações da amostra dentro da área de estudo, pode ser usado para converter dados de observações pontuais para campos contínuos (BURROUGH; McDONNELL, 1998). Nos casos de observações pontuais finitas, os resultados das estimativas de densidade *kernel* são adequados (BOWMAN; AZZALINI, 1997). Originalmente, esta abordagem foi desenvolvida como um método alternativo para obter uma função de densidade probabilidade suave, univariada ou multivariada, a partir de uma amostra de observações (BAILEY; GATRELL, 1995; LEVINE, 2002). Como a estimativa de intensidade das observações pontuais (coordenadas dadas em *x* e *y*) é muito semelhante à densidade de probabilidade bivariada, a abordagem *kernel* pode ser adaptada para este propósito (BAILEY; GATRELL, 1995).

A estimativa de densidade kernel (4 na Figura 6) é um método estatístico não paramétrico para estimar as densidades de probabilidade. Um *kernel* (isto é, densidade de probabilidade normal bivariada) é colocado sobre cada ponto de observação e a intensidade em cada cruzamento de uma grade sobreposta é estimada (SEAMAN; POWELL, 1996). O método é semelhante ao conceito de "janela móvel", em que uma janela de tamanho específico é movida sobre os pontos de observação (GATRELL et al., 1996). Matematicamente, para um local com vetor de coordenadas x_j, a densidade $\hat{f}(x)$, pode ser expressa pela seguinte equação (PARZEN, 1962; ROSENBLATT, 1956):

$$\hat{f}(x) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left\{ \frac{(x_j - X_i)}{h} \right\} \tag{52}$$

Em que: *n* é o número de pontos de observações; x_i é o vetor de coordenadas do ponto de fogo; *K* é a função *kernel; h* é o raio de busca ou largura de banda.

As várias funções de interpolação *K,* diferem na maneira como atribui pesos para os pontos dentro do raio de busca *h,* que pode ser qualquer função de densidade probabilidade (Gaussiana, triangular, quártica, exponencial negativa ou uniforme) desde que:

$$\int_{-\infty}^{+\infty} K(h)dh = 1 \tag{53}$$

A função *kernel* quártica (SILVERMAN, 1986) foi considerada para o cálculo da função *K* no *sofware* ArcGis/ArcInfo 10.4:

$$k(h) = \frac{3}{\pi}\left(1 - h^2\right) \tag{54}$$

A função quártica pondera com maior peso, os pontos mais próximos do que pontos distantes, mais o decrescimento é gradual. O raio de busca ou largura de banda expressa o tamanho do *kernel* e controla a suavização da superfície gerada. Dois tipos de métodos alternativos podem ser aplicados na estimativa da densidade *kernel*, o método fixo e adaptativo. No método fixo, o raio de busca, que é definido em unidades de distância, é constante em toda a área de interesse. No método adaptativo, o raio de busca, que é definido pelo número mínimo de observações pontuais encontradas no *kernel*, varia de acordo com a concentração das observações pontuais. Isto significa que, em áreas de baixa concentração, o raio de busca tem valores mais elevados do que em áreas de alta concentração (WORTON, 1989).

Uma questão importante e difícil de definir ao implementar a interpolação de densidade *kernel*, é a escolha do parâmetro de suavização do *kernel*, tanto no método fixo como no método adaptativo. Um menor raio de busca permite que observações próximas dominem a estimativa de densidade, enquanto maiores raios de busca favorecem locais distantes (WORTON, 1989; SEAMAN; POWELL, 1996). Na literatura, alguns métodos diferentes foram propostos para definir parâmetros de suavização, para avaliar a ocorrência de incêndios e padrão de fogo

(AMATULLI et al., 2007; de la RIVA et al., 2004; KOUTSIAS et al., 2004; LIU et al., 2010). Entretanto, a escolha de um valor arbitrário para o parâmetro de suavização não é recomendada e ainda, deve ser realizada de maneira mais rigorosa, para que o modelo não seja penalizado. Portanto, o raio de busca *h* foi calculado usando uma variante espacial de Silverman (1986) que é robusta a *outliers* (ou seja, pontos que estão distantes dos demais pontos) e que está implementada no *software ArcGis/ArcInfo* 10.4. Assim, delineado a configuração do modelo de densidade *kernel*, o mapa de superfície de ocorrência de fogo, com resolução de célula de grade de 250m foi gerado, para então, ser utilizado na análise de regressão da árvore como variável de resposta.

3.4. Variáveis preditoras

De acordo com a literatura e considerando a relevância de cada variável em explicar a ocorrência dos incêndios florestais na área de estudo, um total de 12 variáveis foram consideradas, abrangendo aspectos topográficos, climáticos, socioeconômicos e de vegetação (Tabela 1). As variáveis determinadas de diversas bases de dados foram então transformadas em imagem raster, com resolução espacial de 250m, de acordo com os procedimentos descritos adiante. Os dados geográficos foram configurados em conformidade com o sistema de referência geocêntrico padrão (SIRGAS 2000) e integrados em ambiente SIG do *ArcGis/ArcInfo* 10.4.

3.4.1. Variáveis topográficas

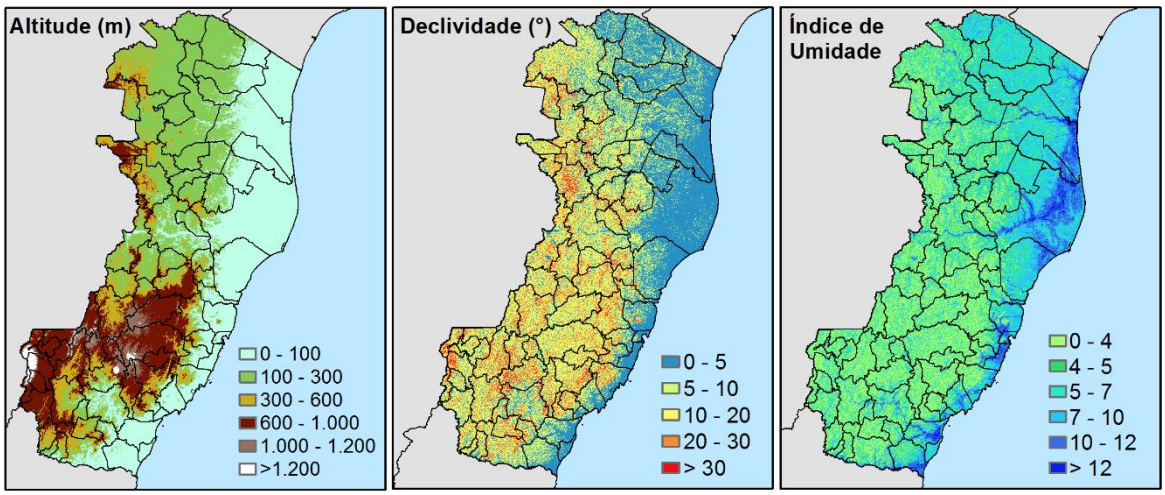
As variáveis topográficas (Figura 7) foram obtidas a partir do Modelo Digital de Elevação (MDE) de alta resolução (30m) do *Shuttle Radar Topography Mission* (SRTM), fornecido pelo *File Transfer Protocol* (FTP) do *United States Geological Survey* (USGS). O SRTM é distribuído em formato GeoTIFF, com coordenadas geográficas no sistema *WGS-84* e referenciados para ondulação do geoide *Earth Gravitational Model* 1996 (EGM96).

Tabela 1 – Variáveis preditoras consideradas na análise da árvore de decisão

Variável		Nome do arquivo da variável	Alcance	Unidade
Topográfica	Altitude	DEM	0 – 2834	m
	Declividade	DEC	0 – 85,40	º
	Índice topográfico composto	TWI	0,75 – 16,07	adimensional
Ambiental	Precipitação média anual	PREC	885,94 – 1817,98	mm
	Temperatura média anual	TEMP	6,54 – 25,79	ºC
	Radiação solar	RADSOLAR	0,23 – 1,30	MJ cm^{-2} hr^{-1}
	Deficiência hídrica média anual	DEF_HID	0 – 603	mm
Socioeconômica	Densidade demográfica	DENS_DEMOG	0 – 52397,7	hab km^{-2}
	Renda	RENDA	0 – 27017,04	R$ $mês^{-1}$
	Proximidade a estradas	PROX_ESTR	0 – 4854,12	m
Vegetação	Campo contínuo de vegetação	VCF	0 – 86	%
	Uso e cobertura da terra	1-Agricultura 2-Áreas urbanas 3-Curso d'água 4-Floresta natural 5-Manguezais 6-Pastagem 7-Silvicultura 8-Solo exposto 9-Áreas alagadas 10-Restinga		

Fonte: o autor.

Figura 7 – Variáveis topográficas consideradas no estudo.



Fonte: o autor.

O Índice Topográfico Composto (ITC), também conhecido como índice de umidade; é função da área de contribuição a montante e da declividade da paisagem (MOORE et al., 1991). A distribuição espacial da água em um campo é influenciada pelo fluxo lateral e, portanto, controlada pelas diferenças de declividade. O ITC é um atributo de terreno composto, calculado a partir da área de captação específica de um ponto (A_S) e o local da inclinação gradiente, tan β (BEVEN; KIRKBY, 1979). O ITC é dado como:

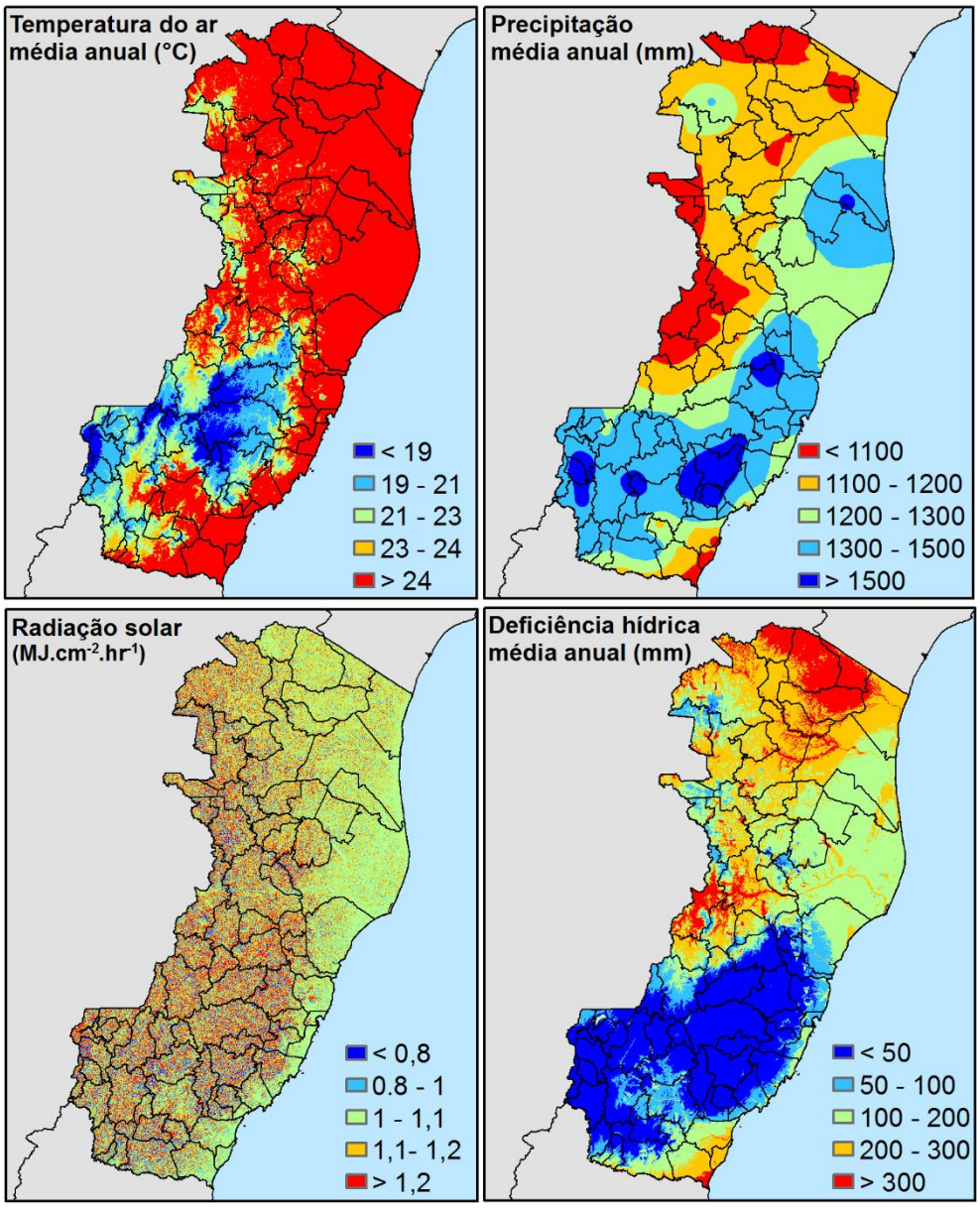
$$\mathrm{ITC} = \frac{\ln(\alpha)}{\tan\beta} \tag{55}$$

Em que: *ln* é o logaritmo natural, α é a área curva ascendente por largura da unidade de contorno e β é o ângulo de inclinação (MOORE et al., 1991). Em geral, o índice é essencialmente uma medida da tendência do acúmulo de água em qualquer ponto, em um plano inclinado. O mapa de índice de umidade indica zonas de alta umidade do solo (altos valores) e zonas potenciais que secam primeiro (valores baixos). Os valores do ITC foram calculados (BÖHNER et al., 2001) para pixels individuais usando o SRTM (MDE). Segundo Vadrevu et al. (2010), valores mais altos de ITC são um bom indicador de baixa probabilidade de fogo.

3.4.2. Variáveis climáticas

Neste estudo, a temperatura do ar média anual, precipitação média anual, radiação solar e deficiência hídrica média anual foram avaliadas como parâmetros de incêndios florestais (Figura 8). Os registros meteorológicos foram baseados em uma série histórica de 34 anos (1977 – 2011) com 110 estações no estado do Espírito Santo e áreas adjacentes. Estes dados foram utilizados para futuras interpolações e estatísticas de dados.

Figura 8 – Variáveis climáticas consideradas no estudo.



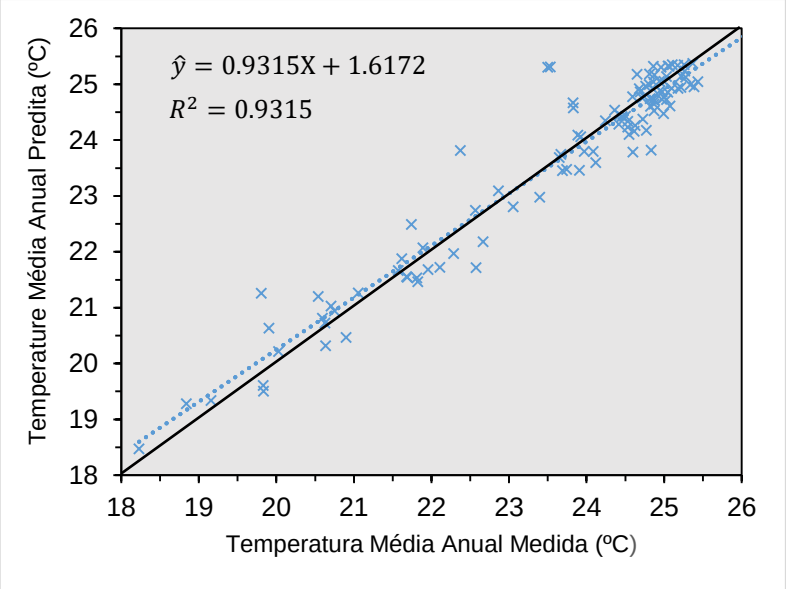
Fonte: o autor.

Os valores de temperatura do ar nas estações sem registros foram estimados pelo ajuste da Equação 56, em função da altitude, latitude e longitude (VIANELLO; ALVES, 2004), adotando-se o Mínimo Quadrado Ordinário (MQO) como técnica de análise de regressão.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \text{ Altitude} + \hat{\beta}_2 \text{ Latitude} + \hat{\beta}_3 \text{ Longitude} \qquad (56)$$

Para a espacialização da temperatura, pelo modelo de regressão linear multipla, foi utilizado o SRTM (MDE) e camadas matriciais de latitude e longitude com valores em graus decimais. O modelo apresentou boa capacidade preditiva com r² ajustado de 0,9315. A análise gráfica dos resultados medidos contra os estimados é apresentada na Figura 9.

Figura 9 – Temperatura média anual medida e estimada para o estado do Espírito Santo.



Fonte: o autor.

As médias anuais de precipitação nos pontos amostrais (estações meteorológicas), foram interpoladas para a área de estudo usando o método Krigagem Bayesiana Empírica (MARSHALL, 1991) como técnica de geoestatística e a construção do semivariograma (MATHERON, 1963) foi usado para gerar a imagem raster de precipitação média anual. O ajuste do modelo produziu por validação cruzada, uma correlação satisfatória entre os valores reais e estimados e o menor erro quadrado médio, em comparação com outros métodos de interpolação que são comumente usados.

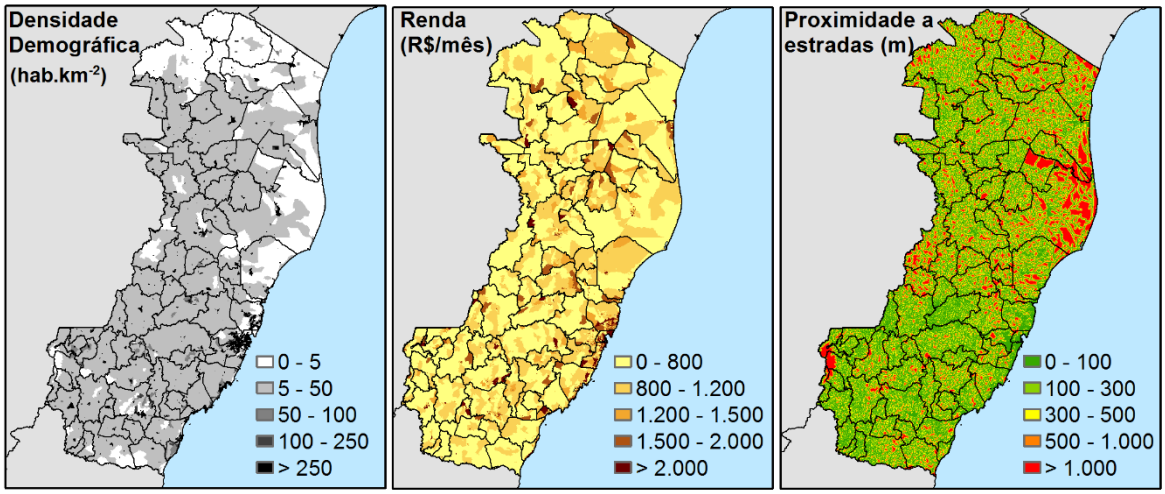
A radiação solar, descreve a incidêndia de radiação solar incidente direta sobre uma superfície local inclinada, em relação a uma superfície horizontal de uma determinada latitude e longitude. Os valores de radiação solar ($MJ.cm^{-2}.hr^{-1}$) foram calculados (McCUNE; DYLAN, 2002) para pixels individuais usando o SRTM (MDE).

A imagem matricial de deficiência hídrica média anual foi obtida por meio do balanço hídrico climatológico espacializado, pixel a pixel, considerando a Capacidade de Água Disponível (CAD) dos solos do estado do Espírito Santo, de acordo com metodologia proposta por Omena (2014), adaptada de Thornthwaite e Mather (1955).

3.4.3. Variáveis socioeconômicas

Os fatores associados com a densidade populacional, renda e proximidade a estradas foram avaliados para determinar o mapa de predição de fogo (Figura 10).

Figura 10 – Variáveis socioeconômmicas consideradas no estudo.



Fonte: o autor.

As variáveis densidade populacional e renda foram obtidas pelo Censo 2010, fornecido pelo *File Transfer Protocol* (FTP) do Instituto Brasileiro de Geografia e Estatística (IBGE). A densidade populacional (hab.km^{-2}) é o resumo mais comum da distribuição da população no espaço geográfico, sendo determinada pela Equação 57.

$$D_i = P_i / A_i \tag{57}$$

Em que: D_i é a densidade populacional em unidade de área i; P_i é a correspondente população e; A_i é a área de terra da unidade (DEICHMANN, 1996).

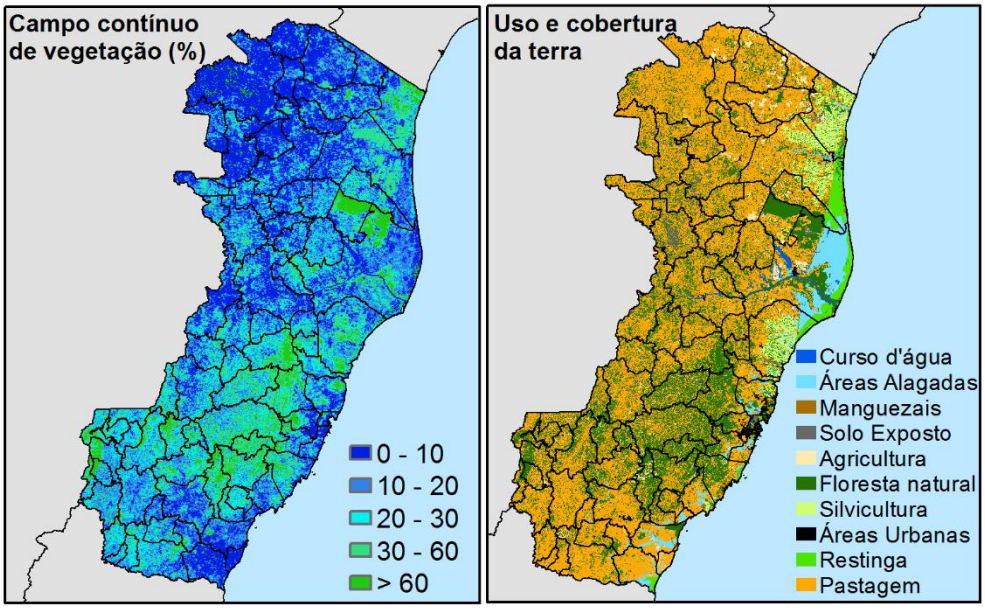
O valor do rendimento nominal mensal dos responsáveis por domicílios permanentes foi considerado como o valor de renda da população na área de estudo, sendo determinado os limites censitários como unidade de área para ambas as variáveis.

A imagem matricial de proximidade a estradas foi obtida pelo algoritmo de distância euclidiana, em função da malha rodoviária do estado, fornecida pelo Sistema Integrado de Bases Geoespaciais do Estado do Espírito Santo (GEOBASES). A distância euclidiana descreve a distância mais próxima em linha reta, entre dois pontos, a partir do centro da célula de origem da imagem matricial para o centro da célula vizinha. Em um plano, a distância entre os pontos de $D_{AB}(X_A, Y_A)$ e (X_B, Y_B) é dada pelo Teorema de Pitágoras (ROMERO-CALCERRADA et al., 2010).

3.4.4. Variáveis de vegetação

Neste estudo, as variáveis campo contínuo de vegetação e uso e cobertura da terra foram consideradas para explicar a ocorrência dos incêndios (Figura 11).

Figura 11 – Variáveis de vegetação consideradas no estudo.



Fonte: o autor.

A porcentagem de cobertura do dossel foi mapeada usando os dados do produto MOD44B do sensor MODIS pela plataforma Terra (TOWNSHEND et al., 2011). O produto *Vegetation Continuous Fiels* (VCF) da coleção 5.1 é uma representação do nível de subpixel das estimativas da cobertura da vegetação da superfície em escala global. O algoritmo VCF envolve um processo semi-automatizado para gerar árvores de regressão com *software* de aprendizado de máquinas. Gerado anualmente, o produto MOD44B é produzido usando compósitos mensais de dados de reflectância de superfície da terra, incluindo as sete bandas do sensor MODIS; e temperatura de superfície da terra. As imagens foram disponibilizadas em formato *Hierarchical Data Format* (HDF) e projeção sinusoidal Lat-Long pelo *United States Geological Survey* (USGS).

Neste estudo, foram utilizadas 2 imagens (ano base de 2015), quadrante h14v10 e h14v11, com resolução espacial de 250m e o *software Modis Reprojection Tools* (MRT) foi usado para reprojetar cada arquivo para UTM e do formato HDF para GeoTiff.

A perda de áreas de floresta está intimamente relacionada com as formas de uso da terra e com o modo de produção estabelecido nas áreas convertidas. Grandes faixas de floresta foram desmatadas para abrir novas áreas de pastagem e agricultura, devido as pressões socioeconômicas, relacionadas com o crescimento populacional e expansão da fronteira agrícola na área de estudo. A imagem matricial de uso e cobertura da terra foi fornecida pelo Geobases sendo obtida pela interpretação do ortofotomosaico. As aerofotos digitais 2007/2008, na escala 1:35.000, com resolução espacial de 1m, foram cedidas pelo Instituto Estadual de Meio Ambiente e Recursos Hídricos do Espírito Santo (IEMA).

3.5. Treinamento do modelo – calibração

Com a variável resposta de densidade de fogo e todo o conjunto de dados de variáveis preditoras foi utilizado a ferramenta *sample* (5 na Figura 6) do *ArcGis/ArcInfo* 10.4, para uma amostragem sistemática e entrada do conjunto de dados na árvore de decisão pelo algoritmo CART. A amostragem sistemática da área de estudo garante uma grande quantidade de dados para formação/calibração do algoritmo da árvore e teste/validação do mapa de saída.

O algoritmo da árvore de decisão CART (BREIMAN et al., 1984) é um procedimento de particionamento recursivo binário capaz de processar atributos contínuos e nominais como alvos e preditores. Os dados são tratados na forma bruta; nenhum *binning* é necessário ou recomendado. Começando no nó da raiz, os dados são divididos até os nós terminais, sem o uso de uma regra de parada (WU; KUMAR, 2009). O algoritmo inicia por analisar todas as variáveis de entrada e determina qual divisão binária de uma única variável de preditores, melhor reduz o desvio na variável de resposta. O processo é repetido para cada partição dos dados resultantes da primeira divisão, continuando até que os nós terminais homogêneos sejam alcançados na árvore hierárquica. A técnica normalmente causa *overfitting* do modelo, criando uma árvore que explica substancialmente todo o desvio nos dados originais. A árvore tem então de ser podada de volta para a raiz pelo método de poda de complexidade de custos. O método de poda envolve a validação cruzada (VENABLES; RIPLEY, 1997), que consiste em dividir os dados originais em conjuntos iguais; que serão usados para gerar árvores de teste, validado contra o último conjunto. As estimativas do erro médio quadrático da validação cruzada ajudam a selecionar o tamanho mais conveniente da árvore, considerando um *trade-off* entre a redução do erro esperado e a conveniência de gerar uma quantidade razoável de regras de decisão.

O mecanismo CART inclui o tratamento automático de dados faltantes, a construção de recursos dinâmicos (WU; KUMAR, 2009), são robustos a *outliers* e não requerem uma seleção *a priori* das variáveis. Além disso, pode modelar relações variantes, apesar da autocorrelação espacial significativa (CABLK et al., 2002) e; os relatórios finais incluem uma classificação de importância relativa das variáveis utilizadas no modelo (STEINBERG; COLLA, 1997).

O conjunto de dados de treinamento foi usado para implementar o algoritmo de árvore de regressão CART usando a versão de demonstração do *software Salford Predictive Modeler* (SPM) 8.0. Neste estudo, um procedimento interno de validação/calibração do desempenho da árvore foi realizado por uma metodologia de validação cruzada *10-fold* capaz de produzir erros de validação para cada árvore gerada. Primeiro, uma grande árvore foi gerada por meio dos critérios de divisão dos mínimos quadrados; sucessivamente, a grande árvore foi podada (6 na Figura 6) para obter um bom nível de erro de validação cruzada, permitindo a seleção de uma árvore menor. Amatulli et al. (2006) menciona que

uma grande árvore pode produzir um processo de regressão muito detalhado, criando regras de decisão para pequenas unidades de risco de incêndio. Como consequência, o algoritmo seria então muito complexo, reduzindo sua natureza interpretável. Além disso, pequenas unidades aumentam sem sentido a segmentação do planejamento de incêndio, perdendo assim a eficiência operacional. Em geral, um nível satisfatório de erro de validação cruzada e tamanho da unidade deve ser identificado, considerando a interpretabilidade das regras de decisão.

As regras de decisão, baseadas nos limiares de valores específicos de cada variável preditora; foram implementadas em linguagem *python,* para leitura do banco de dados do arquivo Excel, sendo possível determinar os valores de saída da árvore. Os valores de saída foram então importados para ambiente SIG (7 na Figura 6), para mapear a densidade prevista do ponto de fogo para cada célula de grade, permitindo a criação do mapa final de predição do fogo. Essa combinação, fornece uma importante ferramenta, para localizar espacialmente as ações de prevenção que devem ser tomadas, no âmbito de um sistema de gestão de fogo.

3.6. Teste do modelo-calibração

Uma validação espacial do mapa resultante foi então obtida por meio do coeficiente r, analisando a correlação entre o valor previsto de risco de incêndio e os valores observados de ocorrência de incêndios (8 na Figura 6).

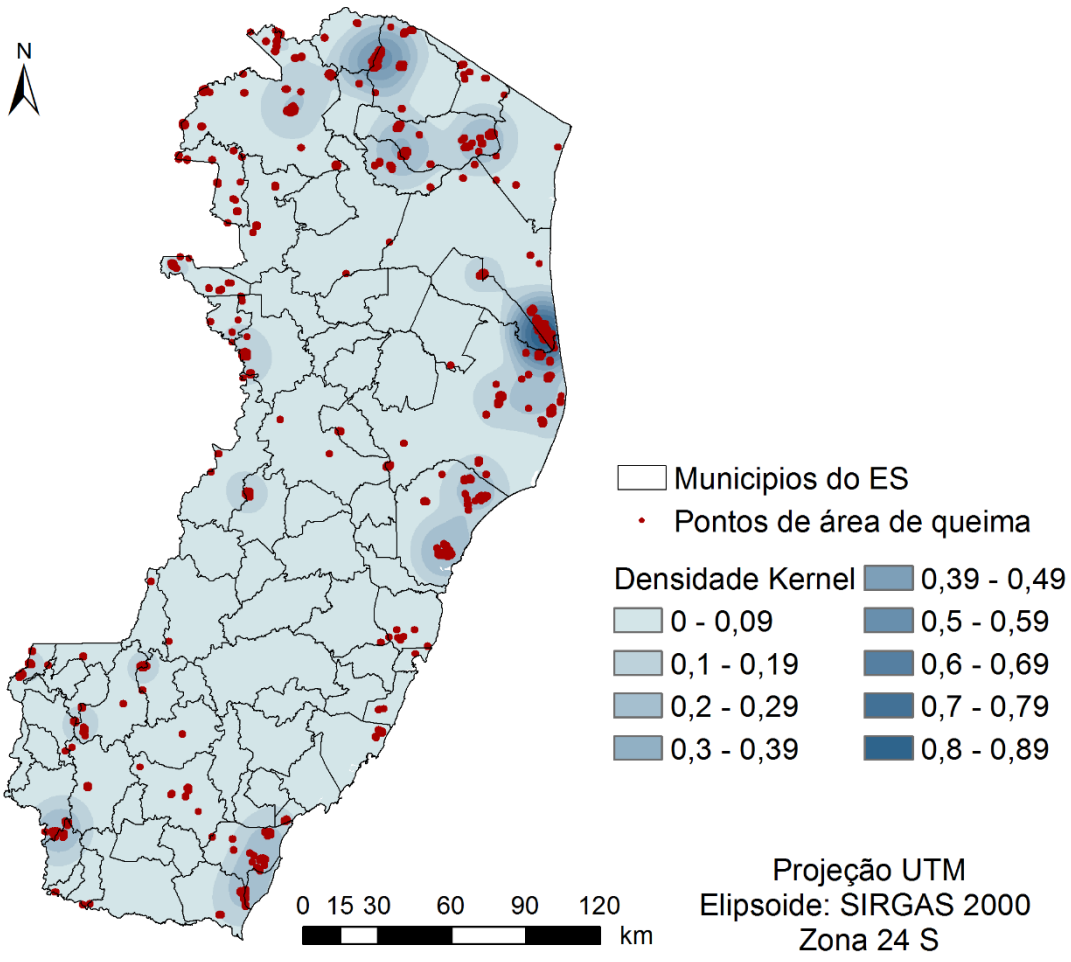
Sucessivamente, a fim de verificar a capacidade preditiva do modelo de regressão da árvore e a influência da posição do ponto de fogo na área de estudo, uma segunda análise de correlação foi realizada, no qual os pontos de área de queima em cada unidade de gestão de incêndios, expressos em densidade observada, foram representados graficamente em função dos valores de densidade predita do mapa de risco de incêndio (9 na Figura 6). A equação de regressão obtida com a sua inclinação, intercepto e coeficiente de *Pearson*, foi então calculado. Além disso, a análise CART permitiu a definição da importância relativa de cada variável no processo de regressão, revelando as capacidades preditivas de cada variável para a predição do risco de incêndio.

4. RESULTADOS E DISCUSSÃO

4.1. Mapa de densidade de fogo

O mapa *kernel* de densidade de fogo (Figura 12) destacou quatro principais áreas de maior queima. Uma na região nordeste do estado, com picos que variam de 0,2 a 0,6 pontos km^{-2}. Uma na região do rio Doce, com picos mais destacados que variam de 0,3 a 0,9 pontos km^{-2} e; duas outras áreas nas regiões costa sul e Caparaó com picos que variam de 0,2 a 0,3 pontos km^{-2}. Nas demais regiões da área de estudo, a ocorrência de fogo diminui ligeiramente para valores próximos de 0 a 0,2 pontos km^{-2}. O padrão espacial da distribuição dos incêndios tem o comportamento agrupado, característico dos incêndios florestais causados pelo ser humano.

Figura 12 – Mapa k*ernel* de densidade de fogo.



Fonte: o autor.

A tecnologia da informação tornou-se importante para monitorar a área de queima. O desenvolvimento de modelos estatísticos espaciais levou a melhorias notáveis na capacidade preditiva do fogo pela integração de um sistema de classificação de risco de incêndio com aplicação de informação e tecnologia espacial. Métodos como a análise de densidade *kernel* fornecem um instrumento para gestores florestais por desenvolver mapas de ocorrência de fogo em situações de variabilidade espacial e temporal.

A técnica de densidade *kernel* é frequentemente utilizada para diferentes aplicações ecológicas, tais como análises de área de vida (BLUNDELL et al., 2001; MILLSPAUGH et al., 2006; SEAMAN; POWELL, 1996; WORTON, 1989) e estudos epidemiológicos espaciais (GATRELL et al., 1996).

A escolha de um parâmetro de suavização (ou seja, largura de banda) apropriado é a etapa mais importante na obtenção de um estimador de densidade *kernel* (WORTON, 1989), mas não há acordo sobre como abordar esse problema (DOWNS; HORNER, 2007; FIEBERG, 2007; GITZEN et al., 2003; HORNE; GARTON, 2006).

O parâmetro de suavização (*h*) determina a propagação do *kernel* centrado em cada observação. Se o valor de *h* for pequeno, os *kernels* individuais serão estreitos e a estimativa de densidade *kernel* em um determinado ponto será baseada em apenas algumas observações. Isso pode não permitir a variação entre as amostras e pode produzir um mapa pouco suavizado (valores altos). Por outro lado, se o valor de *h* for grande, os *kernels* individuais serão amplos, o que pode esconder detalhes finos resultando em um mapa muito suavizado (valores baixos). Portanto, a escolha da abordagem de suavização a usar depende do conjunto de observações e as considerações ecológicas específicas para cada estudo devem ser levadas em conta. A distribuição espacial dos incêndios florestais na área de estudo não é aleatória e a ocorrência de incêndio em áreas específicas depende de uma série de fatores relacionados aos tipos de proteção legal dos recursos naturais, propriedade e manejo florestal.

4.2. Mapa de predição de fogo

O grande conjunto de dados e número de variáveis preditoras criou uma árvore muito complexa, com 5137 nós terminais e um erro de validação cruzada de

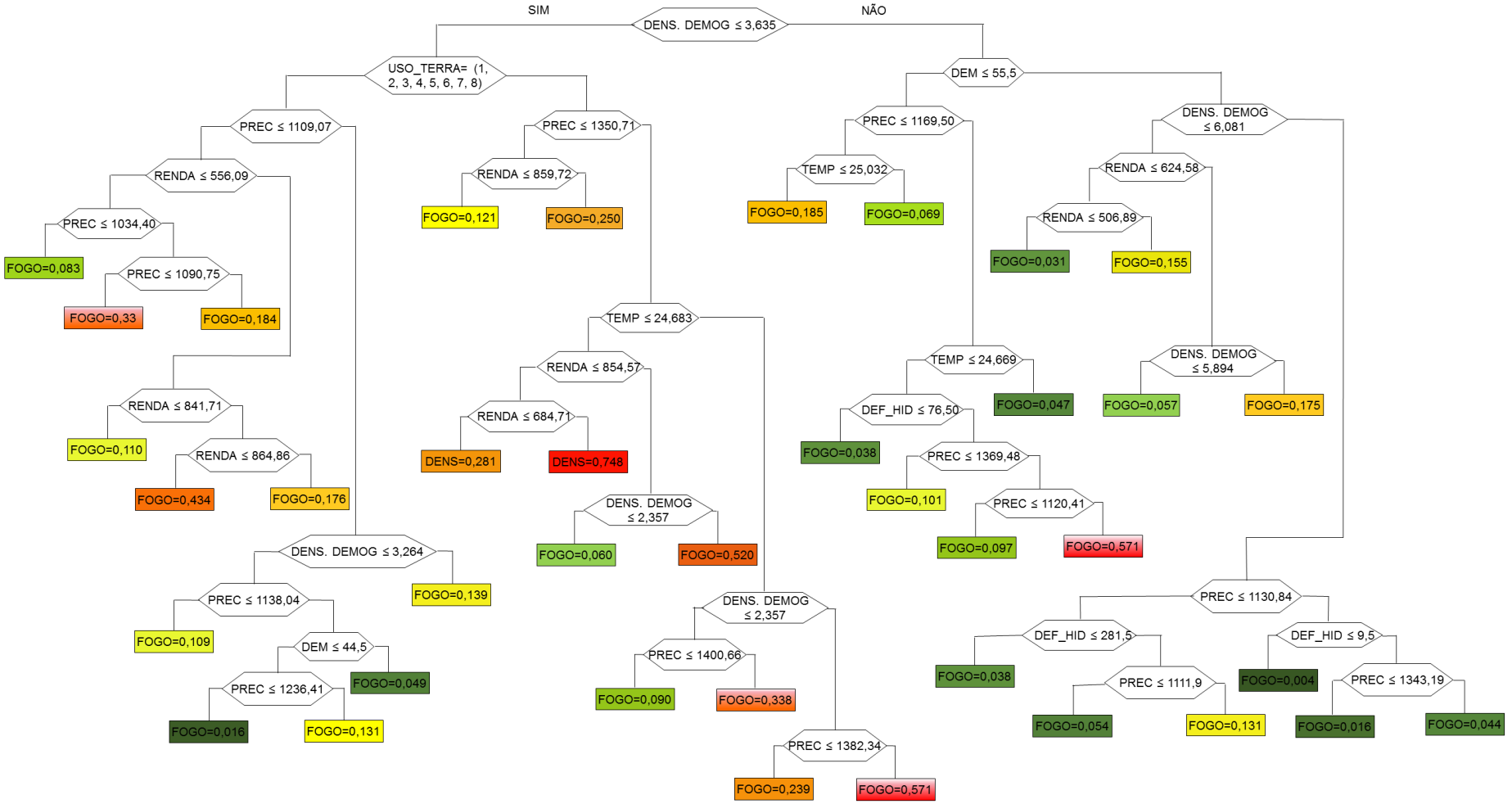
0,01. A ação de poda da árvore foi então realizada sucessivamente para obter uma árvore mais simples, com 38 nós terminais (Figura 13) e um erro de validação cruzada com valores aceitáveis (0,3). As regras de decisão da árvore identificaram vários limiares únicos para cada variável e específicos para cada unidade de gestão de incêndio, úteis para prever 38 densidades médias, variando de 0,004 a 0,748 pontos km^{-2}, suavizando os valores máximos (0,9 pontos km^{-2}) do mapa de densidade de fogo.

Conclusões mais precisas podem ser alcançadas com o tamanho e a estrutura do algoritmo da árvore. Em particular, a ação de poda permite a simplificação da estrutura da árvore, produzindo uma melhoria na computação e uma interpretabilidade fácil das regras de decisão. Zonas de alta densidade de fogo estão associadas, tanto com uma alta ou baixa densidade populacional. Áreas de vegetação de restinga e áreas alagadas, geralmente, estão associadas a zonas de alta densidade de fogo (0,748). Algumas estratégias de gestão podem ser realizadas em tais áreas como a restrição aos locais de risco e o manejo de combustível para evitar o início e propagação do fogo. Uma importante aplicação do algoritmo é a possibilidade da auto-alimentação dos dados para desenvolver automaticamente a predição do fogo na área de estudo. Ou ainda, a criação de cenários com a simulação de mudanças nos dados, sendo possível observar as novas disposições espaciais das unidades de manejo do fogo e seu valor de risco.

Um papel importante é dado pelo parâmetro climático representado pelas variáveis precipitação, temperatura e deficiência hídrica. Este parâmetro é importante indicador ecológico, não só na definição da composição das espécies, mas também na sua distribuição. Tais variáveis são capazes de discriminar as condições ecológicas e susceptibilidades de incêndio não detectadas pelos dados de cobertura da terra. A susceptibilidade das espécies aos incêndios florestais não está apenas relacionada com à inflamabilidade das espécies arbóreas e à estrutura do povoamento, mas também ao estado de estresse hídrico diretamente influenciado pelas condições meteorológicas médias (AGUADO et al., 2003; CHUVIECO; MARTIN, 1994). As previsões dos regimes de fogo assumem uma forte ligação entre o clima e o fogo, mas geralmente com menor ênfase nos efeitos de fatores locais, como a atividade humana (LIU et al., 2010; WOTTON et al., 2010).

As florestas tropicais da Mata Atlântica, apesar de sua localização em uma das áreas mais úmidas do Brasil, onde a precipitação anual média é superior a

Figura 13 – Regras de regressão descritas na forma de árvore binária.



*Os códigos das variáveis preditoras e suas classes ou intervalos estão listadas na Tabela 1.

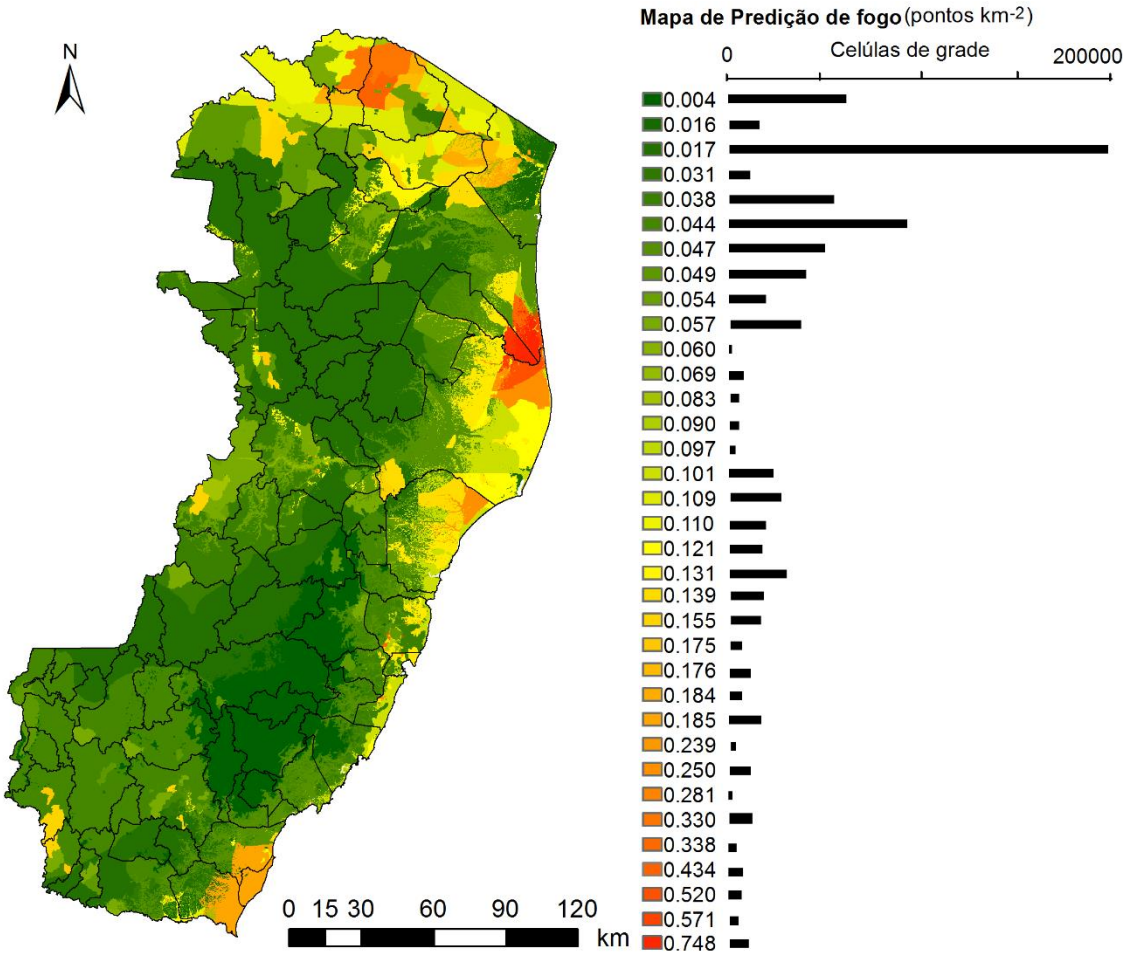
Fonte: o autor.

1500mm, sofrem esporadicamente com incêndios (OLIVEIRA; PASSACANTILI, 2010). Segundo Hammond et al. (2007) florestas úmidas sofrem frequentemente com incêndio nas Guianas, devido a impactos antrópicos. Carcaillet et al. (2002) relatam que nos últimos 2000 anos, na Amazônia, os incêndios também estão presentes no registro paleoambiental, igualmente associados à influência antrópica. Estudos mais recentes em florestas temperadas encontraram resultados semelhantes na transformação de grandes áreas de floresta, em vegetação aberta ou baixa, em regimes de fogo alterados, em face da mudança climática e de uso da terra (PARITSIS et al., 2015; TEPLEY et al., 2016).

A importância dos fatores sociais também pode ser apontada pela variável renda, posto que os incêndios florestais se associam a locais de menor renda. Estudos na Argentina, também relatam que os bairros carentes, os que têm alto desemprego e poucas crianças que frequentam a escola são atingidos com incêndios florestais mais frequentes (CURTH et al., 2012). Nos EUA, os incêndios florestais que começam em comunidades pobres são menos propensos a se extinguir rapidamente por falta de recursos (MERCER; PRESTEMON, 2005). Os danos ecológicos resultantes dos incêndios florestais podem prejudicar a base dos recursos naturais, a partir da qual as comunidades derivam sua atividade econômica e emprego (BUTRY et al., 2001). Tais danos nas comunidades extrativistas dependentes da indústria podem ter efeitos econômicos duradouros (NIEMI; LEE, 2001). Em conjunto, essas descobertas sugerem que as condições sociais podem ser determinantes fundamentais da vulnerabilidade social e dos riscos de incêndios. Compreender como esses componentes de vulnerabilidade variam pode ajudar os gestores a desenvolver estratégias de proteção e mitigação adequadas a locais e populações específicas.

As unidades de gestão variam de grandes (199101 células de grade) a pequenas (179 células de grade). O mapa de predição de fogo com o histograma é representado na Figura 14. De modo efetivo, a localização espacial das unidades de gestão de incêndios e um conhecimento da área de estudo, em termos de ecologia e fatores socioeconômicos, podem ser indicadores importantes das causas de incêndio.

Figura 14 – Mapa de predição de incêndio obtido pela aplicação das regras de decisão.



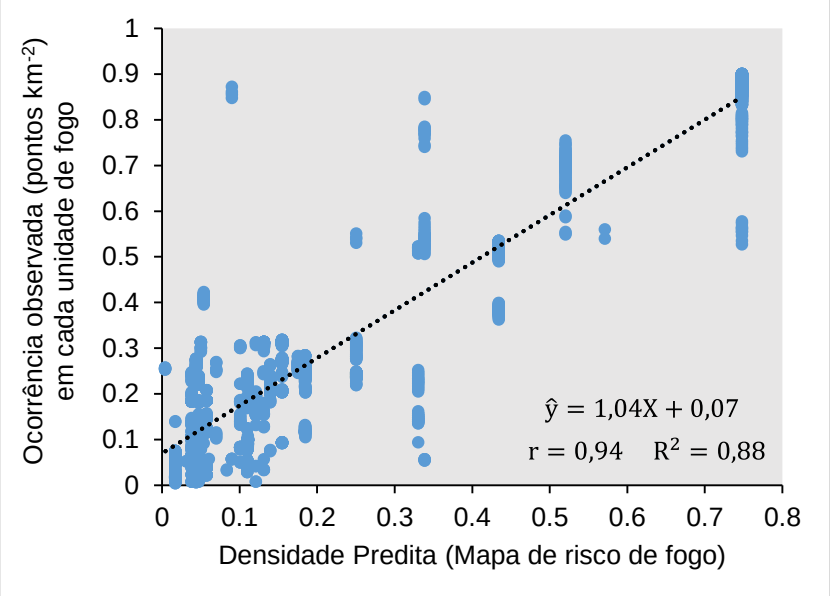
Fonte: o autor.

Na região de maior predição de incêndios florestais, o risco está relacionado a duas causas principais. A primeira é devido ao uso do fogo como recurso para o manejo de áreas para cultivo nas propriedades rurais. Estas zonas são caracterizadas por um elevado nível de fragmentação (JUVANHOL et al., 2017). Estes resultados são consistentes, de acordo, com trabalhos realizados em florestas tropicais (COCHRANE; LAURANCE, 2002; HOLDSWORTH; UHL, 1997), onde as bordas entre a interface meio urbano e floresta são consideradas as mais vulneráveis a incêndios florestais, como também ocorre em florestas temperadas (GANTEAUME et al., 2013; MASELLI et al., 2003; YANG et al., 2007). O nível de fragmentação da paisagem amplia fortemente as fronteiras das bordas, aumentando a probabilidade de que as atividades humanas afetem os processos ecológicos nas áreas naturais; neste caso específico aumentando a vulnerabilidade

ao fogo (JUVANHOL et al., 2017; LEONE; LOVREGLIO, 2003). A segunda causa é mais relacionada com a vegetação, pela faixa de transição entre a floresta e restinga. O predomínio de herbáceas estabelecido em solo arenoso e com grande concentração de matéria orgânica facilita a ignição do fogo (JUVANHOL et al., 2017).

A validação espacial do mapa indica uma satisfatória correlação (0,82), confirmando que a árvore selecionada foi capaz de fornecer um mapa de predição de fogo confiável, seguindo a tendência do mapa de densidade de fogo. Além disso, o algoritmo CART gerou um modelo de regressão aceitável, com um coeficiente de determinação ajustado de 0,88 e correlação de 0,94 entre a densidade predita do mapa de risco de incêndio e a densidade observada nos pontos de área de queima. Em geral, os menores valores de densidade no mapa de risco apresentam melhor ajuste no modelo (Figura 15).

Figura 15 – Equação de regressão obtida entre a densidade predita e a ocorrência observada de incêndios em cada unidade de manejo de fogo.



Fonte: o autor.

A teoria CART pode ser aplicada para prever valores de dados, ao mesmo tempo em que revela quaisquer interações hierárquicas e não-lineares das variáveis. Assim, a técnica pode ser usada não somente para prever os dados, mas também para fornecer informações sobre a estrutura preditiva das variáveis

independentes. Esta característica é importante tanto no sentido exploratório como teórico, pois permite a eventual descoberta de relações desconhecidas entre preditores e variáveis de resposta.

A capacidade de estimar a variável de resposta contínua (característica da regressão) produz uma tendência de risco de incêndio mais realista, em termos de densidade ou probabilidade preditiva do mapa de fogo. Além disso, os problemas associados com a análise de regressão clássica surgem quando são utilizados muitos preditores, o que leva a um aumento maior do que exponencial no número de estruturas possíveis de regressão (BLACKBURN; STEELE, 1999).

Os resultados apresentados pelo modelo da árvore de regressão são melhores quando se compara com outros trabalhos realizados a nível local e nacional (ARPACI, 2014; OLIVEIRA et al., 2012; ZHANG et al., 2016). Particularmente, dentro dos projetos europeus de grandes incêndios, a precisão global alcançada por meio de regressão logística e redes neurais foi de 60% e 69%, respectivamente (BART, 1998; CHUVIECO, 1999). No entanto, algumas observações podem ser extraídas do modelo. Em primeiro lugar, o método de máquina de vetor de suporte pode ser recomendado para o particionamento do conjunto de exemplos da variável a ser utilizada em cada nó, devido sua capacidade de generalização e independência da distribuição dos dados (BRUZZONE; PERSELLO, 2009). Segundo, pelo fato do algoritmo CART operar sem considerar a relação espacial, entre cada célula da grade e nenhuma informação sobre a localização espacial é fornecida no modelo, preditores de localização espacial, como coordenadas x e y, podem ser usados como entrada no modelo. Desta forma, as áreas com ocorrência de incêndio semelhante e próximas as outras podem ser agrupadas na mesma unidade de gestão de incêndio. Em seguida, um tamanho mínimo ou máximo da unidade pode ser definido com base nas ações de prevenção prescritas e nas diretrizes de planejamento do incêndio. Os tamanhos da unidade de gestão de incêndio predefinidos devem ser usados na fase de crescimento da árvore, de modo a reforçar as regras de regressão para agrupar pixels homogêneos em áreas maiores ou menores do que a dimensão definida. Este processo também permitiria uma redução evidente no tamanho da árvore, melhorando a compreensão de todo o processo de regressão.

Para a maior compreensão do método proposto, mais pesquisas são necessárias para ampliar ainda mais a implementação e validação do modelo,

usando outros locais de estudo e um conjunto de dados mais extenso de variáveis preditoras. A mesma metodologia pode até mesmo ser aplicada no quadro de avaliação de risco de incêndio de curto prazo.

A Tabela 2 mostra a pontuação de cada variável no processo da árvore de decisão. A variável com maior capacidade preditiva é a densidade demográfica (100,0), seguida da variável precipitação (78,4) e uso e cobertura da terra (75,07). Renda (36,05) e altitude (33,51) também tiveram boa capacidade preditiva no modelo, enquanto, a variável campo contínuo (1,07) e radiação solar (0,01) não são significativas no desenvolvimento do modelo. Em geral, o fator socioeconômico, ambiental e vegetação, são mais importantes na predição de fogo em escala regional, pois apresentam maiores pontuações e também são confirmadas pela sua presença abundante na estrutura da árvore, para todas as escalas de densidade de fogo (Figura 13). As variáveis topográficas, pelo contrário, apresentam menor importância no modelo preditivo. Entre estas, a altitude é a variável mais relevante. Por último, a variável proximidade a estradas mostra menos influência no processo de regressão, quando comparada com as variáveis acima mencionadas.

Tabela 2 – Pontuação de cada variável no processo da árvore de decisão

Variável	Pontuação (%)
Densidade Demográfica	100
Precipitação média anual	78,4
Uso e cobertura da terra	75,07
Renda	36,05
Altitude	33,51
Índice topográfico composto	24,06
Deficiência hídrica média anual	22,89
Temperatura média anual	22,13
Proximidade a estradas	9,92
Declividade	9,14
Campo contínuo de vegetação	1,04
Radiação solar	0,01

Fonte: o autor.

A pouca influência no processo de regressão da variável distância a estradas é principalmente devido a dois motivos. Em primeiro lugar, a área de estudo é caracterizada por uma rede de estradas prolongada, que produz uma variável contínua heterogênea, sem padrões espaciais evidentes de qualquer dica para o

processo de regressão. O segundo motivo é devido à natureza incompatível entre as duas variáveis (distância a estradas e densidade *kernel*). A distância a estrada perde o seu significado quando correlacionado com a saída de densidade kernel, em vez da localização do ponto de fogo (AMATULLI et al., 2006). A variável pode ser melhor analisada, por exemplo, considerando a distância absoluta dos pontos de fogo ou em termos de densidade da rede de estradas (CHUVIECO et al., 1999). Estas abordagens alternativas podem destacar a importância da variável como fator de risco de incêndio, com seu papel nas atividades operacionais de pré-supressão de incêndio.

Em relação as variáveis topográficas, geralmente apresentam menor importância nos modelos preditivos de ocorrência de incêndios, em comparação com outros fatores. Entre estas variáveis, a altitude é a mais relevante (ARGAÑARAZ et al., 2015; ARPACI et al., 2014; DLAMINI, 2010; OLIVEIRA et al., 2012; WU et al., 2014;) e; a contribuição da declividade é mais explicada pela sua maior importância para os modelos de comportamento do fogo.

Compreender a confluência da vulnerabilidade social e biofísica é especialmente relevante para os incêndios florestais. A frequência, a gravidade e o padrão dos incêndios florestais estão significativamente relacionados às atividades humanas, incluindo o uso da terra, os padrões de estabelecimento da população e o manejo da vegetação (HAWBAKER et al., 2013; SYPHARD et al., 2007, 2013). Por exemplo, a ocorrência de incêndios florestais está positivamente associada à densidade populacional e habitacional (SYPHARD et al., 2007; HAWBAKER et al., 2013), pois as pessoas causam a maioria das ignições de incêndios e os usos da terra influenciam muito os padrões de vegetação e, portanto, o comportamento do fogo (PRESTEMON et al., 2013).

O desenvolvimento em paisagens propensas a incêndios florestais nos EUA é facilitado por condições e processos políticos e econômicos (STETLER; VENN; CALKIN, 2010). Embora algumas políticas, como aquelas de apoio a pesquisas sobre formas sustentáveis de agricultura, venham sendo introduzidas nos últimos anos, áreas florestadas continuam recebendo tratamento menos favorável do regime de taxação das áreas rurais. Acima de tudo, a estrutura política e institucional, ainda favorece sobremaneira, as práticas mais extensivas de cultivo e a conversão da terra.

Devemos dar maior atenção ao desenvolvimento de escalas apropriadas para as ações de conservação. Ressalta-se, a necessidade específica de fortalecimento dos esforços em áreas onde a proteção das terras já esteja estabelecida, bem como naquelas que contêm serviços ambientais vitais para as comunidades locais e naquelas em que os impactos e as ameaças estão particularmente concentrados.

Claramente, é necessário compreender melhor o efeito das ações de gerenciamento, sobre a ocorrência e comportamento do fogo. Isto implicaria necessariamente, estudos de campo em vários locais e em várias escalas, avaliando uma série de questões de pesquisa que vão desde o desenvolvimento de técnicas de prevenção e ações de pré-supressão do fogo. Os governos estaduais e locais também estão cada vez mais envolvidos na gestão dos incêndios florestais (DAVIS, 2001), e existem alguns programas governamentais de gestão de incêndios florestais e programas diretos de assistência ao proprietário (REAMS et al., 2005).

Algumas políticas federais indiretamente apoiam os indivíduos na redução de sua vulnerabilidade ao incêndio. A Lei de Restauração das Florestas Saudáveis de 2003 nos EUA oferece às comunidades oportunidades para desenvolver planos comunitários de proteção contra incêndios florestais, para melhorar sua capacidade de adaptação aos incêndios florestais (GRAYZECK-SOUTER et al., 2009; JAKES et al., 2011; WILLIAMS et al., 2012), e estes planos demonstraram melhorar a resiliência da comunidade (JAKES; STURTEVANT, 2013). Embora os planos identifiquem e priorizem as terras para a redução de combustíveis, eles geralmente incluem recomendações para reduzir a inflamabilidade das estruturas construídas (JAKES et al., 2011). Facilitar o desenvolvimento de planos comunitários de proteção contra incêndios florestais em comunidades vulneráveis poderia ajudar a reduzir sua susceptibilidade a impactos de incêndios, a depender de comunidades que tenham acesso a recursos adequados (JAKES et al., 2011). As comunidades socialmente vulneráveis, geralmente são menos envolvidas nestes e outros programas de mitigação de incêndios florestais (GAITHER et al., 2011), mesmo quando estão expostos a altos níveis de risco de incêndio (OJERIO et al., 2011).

Os esforços federais de planejamento de gestão de incêndios florestais procuram incorporar avaliações de condições sociais selecionadas. A base e implementação de políticas devem ser flexíveis e conscientes das diferenças de

nível comunitário, para incentivar e facilitar a adoção e implementação de estratégias e planos sustentáveis (CHAMP et al., 2012; GRAYZECK-SOUTER et al., 2009; OLSEN; SHARP, 2013; WILLIAMS et al., 2012). Um processo de identificação dos lugares mais vulneráveis aos incêndios florestais apoiaria o desenvolvimento de políticas específicas em diferentes níveis jurisdicionais.

Embora, com foco na avaliação de risco de incêndio e implementação em SIG, este método de análise de dados avançada, poder ser estendida a outros campos da ciência. Em especial, ao avaliar os riscos de desastres naturais, vários fatores são geralmente envolvidos. Sua natureza e comportamento muitas vezes não é muito conhecido e a interação multidisciplinar é necessária, a fim de salientar o complexo mecanismo de suas possíveis relações.

O Sistema de Apoio à Decisão (DSS) proposto constitui uma base sólida no contexto geral de análise de risco. Técnicas de geoprocessamento são geralmente aplicadas na gestão do risco de desastres naturais, devido à sua capacidade de integrar e visualizar os diferentes conjuntos de dados geográficos. No entanto, mais ênfase deve ser posta em aplicação de técnicas não paramétricas. Este é um fator chave para abordar corretamente uma ampla gama de questões relacionadas com o ambiente, a fim de explorar a distribuição de dados e as relações intrinsecamente variáveis.

5. CONCLUSÕES E IMPLICAÇÕES

Este estudo demonstrou que as duas técnicas não paramétricas, combinadas com SIG, pode fornecer um modelo significativo para predizer unidades de risco de incêndios florestais no estado do Espírito Santo.

O modelo de árvore de decisão resultante mostra um bom desempenho entre sua dimensão e o erro de validação cruzada.

As áreas de maiores riscos de incêndios no estado são representadas pela região do vale do rio doce, nordeste e sudeste (costa sul).

Os limiares de decisão de cada variável preditora, pode apoiar os gestores florestais em importante tomada de decisão para as ações de planejamento do fogo em cada unidade de manejo.

O fator socioeconômico, ambiental e vegetação, são mais importantes na predição de fogo em escala regional, pois apresentam maiores pontuações e pela sua maior representatividade na estrutura da árvore, para todas as escalas de densidade de fogo.

A técnica proposta pode ser considerada como uma alternativa a outras técnicas utilizadas, por lidar com estruturas de dados heterogêneas e complexas e pelos melhores resultados apresentados.

Este estudo apresenta um método de análise de dados avançada, cuja aplicação pode ser estendida a outros campos da ciência.

6. REFERÊNCIAS

AGER, A. A.; VAILLANT, N. M.; FINNEY, M. A. A comparison of landscape fuel treatment strategies to mitigate wildland fire risk in the urban interface and preserve old forest structure. **Forest Ecology and Management**, v. 259, n. 8, p. 1556–1570, 2010.

AGUADO, I, CHUVIECO E, MARTÍN, P, SALAS, J. Assessment of forest fire danger conditions in southern Spain from NOAA images and meteorological indices. **International Journal of Remote Sensing**, v. 24, n. 8, p. 1653–1668, 2003.

ALBERTSON, K.; AYLEN, J.; CAVAN, G.; McMORROW, J. Climate change and the future occurrence of moorland wildfires in the Peak District of the UK. **Climate Research**, n. 2007, p. 1–14, 2010.

AMATULLI, G.; RODRIGUES, M. J.; TROMBETTI, M.; LOVREGLIO, R. Assessing long-term fire risk at local scale by means of decision tree technique. **Journal of Geophysical Research: Biogeosciences**, v. 111, n. 4, p. 1–15, 2006.

AMATULLI, G.; PERÉZ-CABELLO, F.; DE LA RIVA, J. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. **Ecological Modelling**, v. 200, n. 3–4, p. 321–333, 2007.

ANDERSON. A model to predict lightning-caused fire occurences. **International Journal of Wildland Fire**, n. 11, p. 163–172, 2002.

ANDREAE, M O; MERLET, P. Emission of trace gases and aerosols from biomass burning. **Global Biogeochemical Cycles**, v. 15, n. 4, p. 955–966, 2001.

ANDREWS, P. L.; LOFTSGAARDEN, D. O.; BRADSHAW, L. S. Evaluation of fire danger rating indexes using logistic regression and percentile analysis. **International Journal of Wildland Fire,** v. 12, p. 213 – 226, 2003.

ARGAÑARAZ, J. P.; PIZARRO, G. G.; ZAK, M.; LANDI, M. A.; BELLIS, L. M. Human and biophysical drivers of fires in Semiarid Chaco mountains of Central Argentina. **Science of the Total Environment**, v. 520, p. 1–12, 2015.

ARPACI, A.; MALOWERSCHNIG, B.; SASS, O. VACIK, H. Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. **Applied Geography**, v. 53, p. 258-270, 2014.

ASCOUGH, J. C.; MAIER, H. R.; RAVALICO, J. K.; STRUDLEY, M. W. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. **Ecological Modelling**, v. 219, p. 383–399, 2008.

ATKINSON, D.; CHLADIL, M.; JANSSEN, V.; LUCIEER, A. Implementation of quantitative bushfire analysis in a GIS environment. **International Journal of Wildland Fire**, p. 1-15, 2010.

BACHMANN, A.; ALLGÖWER, B.; Uncertainty propagation in wildland fire behaviour modelling. **International Journal of Geographical Information Science**, v. 16, n. 2, p. 115-127, 2002.

BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**. Harlow: Longman, 1995.

BARANAUSKAS, J. A.; MONARD, M. C. **Reviewing some machine learning concepts and methods**. Instituto de Ciências Matemáticas e de Computação. Relatórios Técnicos do ICMC, São Carlos. 2000.

BAR-MASSADA, A.; RADELOFF, V. C; STEWART, S. I.; HAWBAKER, T. J. Wildfire risk in the wildland – urban interface: A simulation study in northwestern Wisconsin. **Forest Ecology and Management**, v. 258, p. 1990–1999, 2009.

BART, L. C. Evaluación de la estimación de grandes incendios forestales en la cuenca mediterránea Europea por redes neuronales y regresión logística. **Série Geográfica**, v. 7, p. 73–85, 1998.

BEVEN, K. J.; KIRKBY, M. J. A physically based, variable contributing area model of basin hydrology. **Hydrological Sciences**, v. 24, n. 1, p. 43–69, 1979.

BLACKBURN, G. A.; STEELE, C. M. Towards the Remote Sensing of Matorral Vegetation Physiology: Relationships between Spectral Reflectance , Pigment , and Biophysical Characteristics of Semiarid Bushland Canopies. **Remote Sensing of Environment**, v. 70, p. 278–292, 1999.

BLUMER, A.; EHRENFEUCHT, A.;HAUSSLER, D.; WARMUTH, M. K. Occam's Razor. **Information Processing Letters**, v. 24, n. April, p. 377–380, 1987.

BLUNDELL, G M, MAIER, J A K, DEBEVEC, E. M. Linear Home Ranges: Effects of Smoothing, Sample Size, and Autocorrelation on Kernel Estimates. **Ecological Monographs**, v. 71, n. 3, p. 469–489, 2001.

BÖHNER, J.; KÖHTHE, R.; CONRAD, O.; GROSS, J.; RINGELER, A.; SELIGE, T. Soil regionalisation by means of terrain analysis and process parameterisation. **European Soil Bureau**. n. 7, p. 213-222, 2002.

BONAZOUNTAS, M.; KALLIDROMITOU, D.; KASSOMENOS, P. A.; PASSAS, N. Forest fire risk analysis. **Human and Ecological Risk Assessment**, v. 11, n. 3, p. 617-626, 2005.

BORCHERS, J. G. Accepting uncertainty, assessing risk: Decision quality in managing wildfire , forest resource values , and new technology. **Forest Ecology and Management**, v. 211, p. 36–46, 2005.

BOWMAN, A. W.; AZZALINI, A. **Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrattions**. New York: Publications, Oxford Science, 1997.

BRAGA, J.; STARMER, C. Preference Anomalies, Preference Elicitation and the Discovered Preference Hypothesis. **Environmental & Resource Economics**, v. 32, p. 55–89, 2005.

BRAVO, S.; KUNST, C.; GRAU, R.; ARÁOZ, E. Fire-rainfall relationships in Argentine Chaco savannas. **Journal of Arid Environments**, v. 74, p. 1319–1323, 2010.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**. Chapman & Hall, 1984.

BRESLOW, L. A.; AHA, D. W. **Simplifying Decision Trees**: A Survey.Technical Report Nº AIC-96-014, 1997.

BROWN, A. A.; DAVIS, K.P. **Forest Fire – Control and Use**. Second edition, New York: McGraw Hill, 1973. 686p.

BROWN, T. C.; KINGSLEY, D.; PETERSON, G. L.; FLORES, N. E.; CLARKE, A.; BIRJULIN, A. Reliability of individual valuations of public and private goods: Choice consistency , response time , and preference refinement. **Journal of Public Economics**, v. 92, p. 1595–1606, 2008.

BRUGNACH, M., DEWULF, A.; HENRIKSEN, H. J.; der KEUR, V. V. More is not always better: Coping with ambiguity in natural resources management. **Journal of Environmental Management**, v. 92, n. 1, p. 78–84, 2011.

BRUZZONE, L.; PERSELLO, C. A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples. **IEEE Transactions on Geoscience and Remote Sensing**, v. 47, n. 7, p. 2142–2154, 2009.

BURROUGH, P. A; MCDONNELL, R. A. **Principles of Geographical Information Systems**. Oxford University Press, 1998.

BUTRY, D. T.; MERCER, D. E.; PRESTEMON, J. P.; PYE, J. M.; HOLMES, T. P. What is the price of catastrophic wildfire?. **Journal o Forestry**, p. 9–17, 2001.

CAETANO, M. R.; FREIRE, S.; CARRÃO, H. Fire risk mapping by integration of dynamic and structural variables. **Remote Sensing in Trasition**, p. 319-326, 2004.

CARCAILLET, C. et al. Holocene biomass burning and global dynamics of the carbon cycle. **Chemosphere**, v. 49, p. 845–863, 2002.

CARDILLE, J. A.; VENTURA, S. Occurrence of wildfire in the northern Great Lakes Region: Effects of land cover and land ownership assessed at multiple scales. **International Journal of Wildland Fire**, n. 10, p. 145–154, 2001.

CARDILLE, J. A; VENTURA, S. J.; TURNER, M. G. Environmental and Social Factors Influencing Wildfires in the Upper Midwest, United States. **Ecological Applications**, v. 11, n. 1, p. 111–127, 2001.

CARMEL, Y.; PAZ, S.; JAHASHAN, F.; SHOSHANY, M. Assessing fire risk using Monte Carlo simulations of fire spread. **Forest Ecology and Management**, v. 257, p. 370–377, 2009.

CATRY., F. X.; REGO, F. C.; BACAO, F. L.; MOREIRA, F. Modeling and mapping wildfire ignition risk in Portugal. **International Journal of Wildland Fire**, v.18, p. 1–11, 2009.

CESTNIK, B.; BRATKO, I. On Estimating Probabilities in Tree Pruning. In: KODRATOFF, I. (Ed.). **Machine Learning: EWSL-91**. Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag, v. 482, 1991. p. 138-150.

CHAMP, J. G.; BROOKS, J. J.; WILLIAMS, D. R. Stakeholder Understandings of Wildfire Mitigation: A Case of Shared and Contested Meanings. **Environmental Management**, v. 50, p. 581–597, 2012.

CHOU, Y. H.; MINNICH, R. A.; CHASE, R. A. Mapping Probability of Fire Occurrence in San Jacinto Mountains , California. **Environmental Management**, USA. v. 17, n. 1, p. 129–140, 1993.

CHUVIECO, E. **Remote Sensing of Large Wildfires in the European Mediterranean Basin**. Berlim: Springer, 1999.

CHUVIECO, E.; GONZÁLEZ, I.; VERDÚ, F.; AGUADO, I.; YEBRA, M. Prediction of fire occurrence from live fuel moisture content measurements in a Mediterranean ecosystem. **International Journal of Wildland Fire**, v. 18, p. 430–441, 2009.

CHUVIECO, E.; CONGALTON, R. G. Application of Remote Sensing and Geographic Information Systems to Forest Fire Hazard Mapping. **Remote Sensing of Environment**, v. 159, p. 147–159, 1989.

CHUVIECO, E.; MARTIN, M. P. Global fire mapping and fire danger estimation using AVHRR images. **Photogrammetry and Remote Sensing**, v. 60, n. 5, p. 563–570, 1994.

COCHRANE, M. A.; LAURANCE, W. F. Fire as a large-scale edge effect in Amazonian forests. **Journal of Tropical Ecology**, v. 18, p. 311–325, 2002.

COVINGTON, W. W.; MOORE, M. M. Southwestern Ponderosa Forest Structure: Changes since Euro-American Settlement. **Journal of Forestry**, v. 92, n. 1, p. 39–47, 1994.

CRUZ, M. G.; ALEXANDER, M. E. Assessing crown fire potential in coniferous forests of western North America: A critique of current approaches and recent simulation studies. **International Journal of Wildland Fire**, v. 19, p. 377–398, 2010.

CUNNINGHAM, A. A.; MARTELL, D. L. A stochastic model for the occurrence of man-caused forest fires. **Canadian Journal of Forest Research**, v. 3, p. 282-287, 1973.

CURTH, M. D. T.; BISCAYART, C.; GHERMANDI, L.; PFISTER, G. Wildland–Urban Interface Fires and Socioeconomic Conditions: A Case Study of a Northwestern Patagonia City. **Environmental Management**, v. 49, p. 876–891, 2012.

DAVIS, C. The West in flames: The intergovernmental politics of wildfire suppression and prevention. **Publius**, v. 31, n. 3, p. 97–110, 2001.

DEEMING, J. E.; BURGAN, R. E.; COHEN, J. D. **The National Fire Danger Rating System – 1978**. USDA Forest Service, Intermountain Forest and Range Experimental Station, INT-39, Odgen, Utah, USA.

DE LA RIVA, J.; PÉREZ-CABELLO, F.; LANA-RENAULT, N.; KOUTSIAS, N. Mapping wildfire occurrence at regional scale. **Remote Sensing of Environment**, v. 92, n. 2, p. 288–294, 2004.

DEICHMANN, U. **A review of spatial population database design and modeling**. Santa Barbara: Techinal Report 96-3, 1996.

DIAZ-BALTEIRO, L.; ROMERO, C. Making forestry decisions with multiple criteria: A review and an assessment. **Forest Ecology and Management**, v. 255, n. 8–9, p. 3222–3241, 2008.

DILTS, T. E.; SIBOLD, J. S.; BIONDI, F. A Weights-of-Evidence Model for Mapping the Probability of Fire Occurrence in Lincoln County, Nevada. **Annals of the Association of American Geographers**, v. 99, n. 4, p. 712–727, 2009.

DLAMINI, W. M. A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. **Environmental Modelling and Software**, v. 25, n. 2, p. 199–208, 2010.

DONOGHUE, L. R.; MAIN, W. A. Some factors influencing wildfire occurrence and measurement of fire-prevention effectiveness. **Journal of Environmental Management**, v. 20, p. 87-96, 1985.

DOWNS, J. A.; HORNER, M. W. Effects of Point Pattern Shape on Home-Range Estimates. **Journal of Wildlife Management**, v. 72, n. 8, p. 1813–1818, 2007.

ESPOSITO, F.; MALERBA, D.; SEMERARO, G. A Comparative Analysis of Methods for Pruning Decision Trees. **IEEE Transactions on Patterns Analysis and Machine Intelligence**, v. 19, n. 5, p. 476–491, 1997.

EUGENIO, F. C.; SANTOS, A. R.; FIEDLER, N. C.; RIBEIRO, G. A.; SILVA, A. G.; SANTOS, A. B.; PANETO, G. G.; SCHETTINO, V. R. Applying GIS to develop a model for forest fire risk: A case study in Espírito Santo, Brazil. **Journal of Environmental Management**, v. 173, p. 65–71, 2016.

FAIRBROTHER, A.; TURNLEY, J. G. Predicting risks of uncharacteristic wildfires: Application of the risk assessment process. **Forest Ecology and Management**, v. 211, p. 28–35, 2005.

FIEBERG, J. Kernel density estimators of home range: smoothing and the

autocorrelation red herring. **Ecology**, v. 88, n. 4, p. 1059–1066, 2007.

FINNEY, M. A. The challenge of quantitative risk analysis for wildland fire. **Forest Ecology and Management**, v. 211, p. 97–108, 2005.

FLANNIGAN, M D; WOTTON, B. M. Climate, Weather, and Area Burned. In: JOHNSON E. A.; MIYANISKI, K. (Ed.). **Forest Fires: Behavior and Ecological Effects**. Academic Press, 2001. p. 351–373.

FONSECA, J. M. M. R. **Indução de árvores de decisão HistClass - Proposta de um algoritmo não paramétrico**. 1994. 134 f. Dissertação (Mestrado em Informática). Universidade Nova de Lisboa, Lisboa.1994.

FRANK, E. **Pruning Decision Trees and Lists**. 2000. 204 f. Tese (Doctor of Philosophy). University of Waikato, New Zealand. 2000.

GAITHER, C. J.; NEELAM, C. P.; GOODRICK, S.; BOWKER, J. M.; MALONE, S.; GAN, J. Wildland fire risk and social vulnerability in the Southeastern United States: An exploratory spatial data analysis approach. **Forest Policy and Economics**, v. 13, n. 1, p. 24–36, 2011.

GAMA, J. M. P. **Combining classification algoritms**. 1999. 190 f. Tese (Doutorado em Ciências de Computadores). Universidade do Porto, Porto. 1999.

GANTEAUME, A.; CAMIA, A.; SAN-MIGUEL-AYANZ, M. J. J. A Review of the Main Driving Factors of Forest Fire Ignition Over Europe. **Environmental Management**, v. 51, p. 651–662, 2013.

GARCIA-DIEZ, E. L.; RIVAS, L. S.; PABLO, F.; GARCIA-DIEZ, A. An objetive forecasting model for the daily outbreak of forest fires based on meteorological considerations. **Journal of Applied Meteorology**, v. 33, p. 519-526, 1994.

GARCIA-DIEZ, E. L.; RIVAS, L. S.; PABLO, F.; GARCIA-DIEZ, A. Prediction of the daily number of forest fires. **International Journal of Wildland Fire**, v. 9, p. 207-2011, 1999.

GATRELL, A. C., BAYLEY, T. C.; DIGGLE, P. J., ROWLINGSON, B. S. Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology. **Transactions of the Institute of British Geographers**, v. 21, n. 1, p. 256–274, 1996.

GENTON, M. G.; BUTRY, D. T.; GUMPERTZ, M. L.; PRESTEMON, J. P. Spatio-temporal analysis of wildfire ignitions in the St Johns River Water Management District , Florida. **International Journal of Wildland Fire**, n. 15, p. 87–97, 2006.

GITZEN, ROBERT A, MILLSPAUGH, JOSHUA J, KERNOHAN, B. J. Bandwidth Selection for Fixed-Kernel Analysis of Animal Utilization Distributions. **Journal of Wildlife Managementife**, v. 70, n. 5, p. 1334–1344, 2003.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

GONZÁLEZ, J. R., PALAHÍ, M.; TRASOBARES, A.; PUKKALA, T. A fire probability model for forest stands in Catalonia (north-east Spain). **Annals of Forest Science**, v. 63, p. 169-176, 2006.

GONZÁLEZ, J. R.; KOLEHMAINEN, O.; PUKKALA, T. Using expert knowledge to model forest stand vulnerability to fire. **Computers and Electronics in Agriculture**, v. 55, p. 107–114, 2007.

GRAYZECK-SOUTER, S.; NELSOM, K. C.; BRUMMEL, R. F.; JAKES, P.; WILLIAMS, D. R. Interpreting federal policy at the local level: the wildland – urban interface concept in wildfire protection planning in the eastern United States. **International Journal of Wildland Fire**, v. 18, p. 278–289, 2009.

HAINES, D.A.; MAIN, W. A.; FROST, J. S.; SIMARD, A. J. Fire-danger rating and wildfire occurrence in the northeastern United States. **Forest Science,** v. 29, p. 679-696, 1983.

HAMMOND, D. S.; TER STEEGE, H.; VAN DER BORG, K. Upland Soil Charcoal in the Wet Tropical Forests of Central Guyana. Biotropica, v. 39, n. 2, p. 153-160, 2007.

HARDY, C.; OTTMAR, R. D.; PETERSON, J. L.; CORE, J. E.; SEAMON, P. **Smoke management guide for prescribed and wildland fire.** National Wildfire Coordinating Group, Boise, Idaho. 2001.

HARDY, C. C. Wildland fire hazard and risk: Problems, definitions, and context. **Forest Ecology and Management**, v. 211, p. 73–82, 2005.

HAWBAKER, T. J.; RADELOFF, V. C.; STEWART, S.; HAMMER, R. B.; KEULER, N. S.; CLAYTON, M. K. Human and biophysical influences on fire occurrence in the United States. **Ecological Applications**, v. 23, n. 3, p. 565–582, 2013.

HERING, A. S.; BELL, C. L.; GENTON, M. G. Modeling spatio-temporal wildfire ignition point patterns. **Environmental Ecology Statistics**, n. 16, p. 225–250, 2009.

HESSBURG, P. F.; REYNOLDS, K. M.; KEANE, R. E.; JAMES, K. M.; SALTER, R. B. Evaluating Wildland Fire Danger and Prioritizing Vegetation and Fuels Treatments. **Forest Ecology and Management**, v. 247, p. 1–17, 2007.

HOLDSWORTH, A R; UHL, C. Fire in amazonian selectively logged rain forest and the potential for fire reduction. **Ecological Applications**, v. 7, n. 2, p. 713–725, 1997.

HORNE, JON S, GARTON, E. O. Likelihood Cross-Validation Versus Least Squares Cross- Validation for Choosing the Smoothing Parameter in Kernel Home-Range Analysis. **Journal of Wildlife Management**, v. 70, n. 3, p. 641–648, 2006.
HUNT, E.; MARIN, J.; STONE, P. **Experiments in induction**. Academic Press. 1966.

HYAFIL, L.; RIVEST, R. L. Constructing optimal binary decision trees is np-

complete. **Information Processing Letters**, v. 5, n. 1, p. 15–17, 1976.

JAKES, P. J. et al. Community wildfire protection planning: is the Healthy Forests Restoration Act's vagueness genius? **International Journal of Wildland Fire**, v. 20, p. 350–363, 2011.

JAKES, P. J.; STURTEVANT, V. Trial by fire: Community wildfire protection plans put to the test. **International Journal of Wildland Fire**, v. 22, p. 1134–1143, 2013.

JONES, S. D.; GARVEY, M. F.; HUNTER, G. J.; Where's the fire? Quantifying uncertainty in a wildfire threat model. **International Journal of Wildland Fire**, v. 13, p. 17-25, 2004.

JUVANHOL, R. S.; FIEDLER, N. C.; ROSA, A. R.; SILVA, G. F.; OMENA, M. S.; PINHEIRO, C. J. G.; EUGÊNIO, F. C. Gis and Fuzzy logic applied to modelling forest fire risk. **Environmental Development**, 2017. No prelo.

KALABOKIDIS, K. D.; KOUTSIAS, N.; KONSTANTINIDIS, P.; VASILAKOS, C. Multivariate analysis of landscape wildfire dynamics in a Mediterranean ecosystem of Greece. **Area**, v. 39, n. 3, p. 392–402, 2007.

KALKANIS, G. The application of confidence interval error analysis to the design of decision tree classifiers. **Patterns Recognition Letters**, v. 14, p. 355–361, 1993.

KALOUDIS, S.; TOCATLIDOU, A.; LORENTZOS, N. A.; SIDERIDIS, A. B.; KARTERIS, M. Assessing Wildfire Destruction Danger: a Decision Support System Incorporating Uncertainty. **Ecological Modelling**, v. 181, p. 25–38, 2005.

KALOUDIS, S. T.; YIALOURIS, C. P.; LORENTZOS, N. A.; KARTERIS, M.; SIDERIDIS, A. B. Forest management planning expert system for wildfire damage reduction. **Computers and Electronics in Agriculture**, v. 70, p. 285–291, 2010.

KANGAS, A. S.; KANGAS, J. Probability, possibility and evidence: approaches to consider risk and uncertainty in forestry decision analysis. **Forest Policy and Economics**, v. 6, p. 169–188, 2004.

KENNEDY, R. E., TOWNSEND, P. A.; GROSS, J. E.; COHEN, W. B.; BOLSTAD, P.; WANG, Y. Q.; ADAMS, P. Remote sensing change detection tools for natural resource managers : Understanding concepts and tradeoffs in the design of landscape monitoring projects. **Remote Sensing of Environment**, v. 113, n. 7, p. 1382–1396, 2009.

KILINK, M.; BERINGER, K. E.The Spatial and Temporal Distribution of Lightning Strikes and Their Relationship with Vegetation Type , Elevation , and Fire Scars in the Northern Territory. **American Meteorological Society**, v. 20, p. 1161–1173, 2007.

KNUTH, D. E. Two Notes on Notation. **The American Mathematical Monthly**, v. 99, n. 5, p. 403–422, 1992.

KOUTSIAS, N.; KALABOKIDIS, K. D.; ALLGÖWER, B. Fire occurrence patterns at the landscape level: beyond positional accuracy of ignition points with kernel density methods. **Natural Resource Modeling**, v. 17, n. 4, p. 359–375, 2004.

KROUGLY, Z. L.; CREED, I. F.; STANFORD, D. A. A stochastic model for generating disturbance patterns within landscapes. **Computers and Geosciences**, v. 35, n. 7, p. 1451–1459, 2009.

KRUSEL, N.; PACKHAM, D.R.; TAPPER, N. Wildfire activity in mallee shrubland of Victoria, Australia. **International Journal of Wildland Fire**, v. 3, p. 277-288, 1993.

KUSHLA, J. D.; RIPPLE, W. J. The role of terrain in a fire mosaic of a temperate coniferous forest. **Forest Ecology and Management**, v. 95, n. 2, p. 97–107, jul. 1997.

LENTILE, L. B.; HOLDEN, Z. A.; SMITH, A. M. S.; FALKOWSKI, M. J.; HUDAK, A. T.; MORGAN, P.; LEWIS, S. A.; GESSLER, P. E.; BENSON, N. C. Remote sensing techniques to assess active fire characteristics and post-fire effects. **International Journal of Wildland Fire**, v. 15, p. 319–345, 2006.

LEONE, V.; LOVREGLIO, R. Human fire causes: A challenge for modelling. In: CHUVIECO, E.; MARTIN, P.; JUSTICE, C. **Annals of the 4th Intern. Workshop on Remote Sensing and GIS Applications to Forest Fire Management:** Innovative Concepts and Methods in Fire Danger Estimation, 2003. p. 89–94.

LEVINE, N. **CrimeStat II: A Spatial Statistics Program for the Analysis of Crime Incident Locations**. The National Institute of Justice, Washington, DC. 2002.

LIU, W.; WANG, S.; ZHOU, Y.; WANG, L.; ZHANG, S. Spatian Distribution Patterns of Historical Forest Fires in DaXingAn Mountains of China. International Conference on Computer Application and System Modeling. **Anais do ICCASM**, p. 634-638, 2010.

LOUPPE. G. **Understanding Random Forests from theory to practice**. 2014. 213 f. PhD dissertation (Faculty of Applied Scienes). University of Liège. 2014.

MAGUIRE, L. A.; ALBRIGHT, E. A. Can behavioral decision theory explain risk-averse fire management decisions? **Forest Ecology and Management**, v. 211, p. 47–58, 2005.

MAINGI, J. K.; HENRY, M. C. Factors influencing wildfire occurrence and distribution in eastern Kentucky , USA. **International Journal of Wildland Fire**, n. 16, p. 23–33, 2007.

MANDALLAZ, D.; YE, R. Prediction of forest fires with Poisson models. **Canadian Journal of Forest Research**, v. 27, p. 1685-1694, 1997.

MARSHALL, R. J. Mapping Disease and Mortality Rates Using Empirical Bayes Estimators. **Applied Statistics**, v. 40, n. 2, p. 283–294, 1991.
MARTELL, D. L.; BEVILACQUA, E.; STOCKS, B. J. Modelling seasonal variation in

daily people- caused fire occurrence. **Canadian Journal of Forest Research**, v. 19, p. 1555–1563, 1989.

MARTELL, D. L. Forest fire management. In: WEINTRAUB, A.; ROMERO, C.; BJØRNDAL, T., EPSTEIN, R., MIRANDA, J. (Eds.), **Handbook of Operations Research in Natural Resources**. Springer, US. 2007, p. 488-509.

MARTÍNEZ, J.; VEGA-GARCIA, C.; CHUVIECO, E. Human-caused wildfire risk rating for prevention planning in Spain. **Journal of Environmental Management**, v. 90, p. 1241–1252, 2009.

MARTÍNEZ-FERNÁNDEZ, J.; CHUVIECO, E.; KOUTSIAS, N. Modelling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression. **Natural Hazards and Earth System Science**, v. 13, n. 2, p. 311–327, 2013.

MASELLI, F.; ROMANELLI, S.; BOTTAI, L.; ZIPOLI, G. Use of NOAA-AVHRR NDVI images for the estimation of dynamic fire risk in Mediterranean areas. **Remote Sensing of Environment**, v. 86, p. 187–197, 2003.

MATHERON, G. Principles of Geostatistics. **Economic Geology**, v. 58, p. 1246–1266, 1963.

McCUNE, B.; DYLAN, K. Equations for potential annual direct incident radiation and heat load. **Journal of Vegetation Science**, v. 13, n. 4, p. 603–606, 2002.

McRAE, R. H. D. Prediction of areas prone to lightning ignition. **International Journal of Wildland Fire**, v. 2, n. 3, p. 123–130, 1992.

MENDOZA, G. A.; MARTINS, H. Multi-criteria decision analysis in natural resource management : A critical review of methods and new modelling paradigms. **Forest Ecology and Management**, v. 230, p. 1–22, 2006.

MERCER, D. E.; PRESTEMON, J. P. Comparing production function models for wildfire risk analysis in the wildland – urban interface. **Forest Ecology and Management**, v. 7, p. 782–795, 2005.

MICHAELSEN, J.; SCHIMEL, D. S.; FRIEDL, M. A.; DAVIS, F. W.; DUBAYAH, R. C. Regression Tree Analysis of satellite and terrain data to guide vegetation sampling and surveys. **Journal of Vegetation Sciences**, v. 5, p. 673–686, 1994.

MILLSPAUGH, J. J.; NIELSON, R. M.; McDONALD, L.; MARZLUFF, J. M.; GITZEN, R. A.; RITTENHOUSE, C. D.; HUBBARD, M. W.; SHERIFF, S. L. Analysis of Resource Selection Using Utilization Distributions. **Journal of Wildlife Management**, v. 70, n. 2, p. 384–395, 2006.

MINGERS, J. Expert Systems-Rule Induction With Statistical Data. **Journal Operational Research Society**, v. 38, p. 39-47, 1987.

MINGERS, J. An Empirical Comparison of Pruning Methods for Decision Tree Induction. Acade. **Machine Learning**, v. 4, p. 227-243, 1989.

MITCHELL, T. M. **Machine Learning**. WCB/McGraw-Hill, 1997.

MOORE, I. D.; GRAYSON, R. B.; LADSON, A. R. Terrain based catchment partitioning and runoff prediction using vector elevation data. **Water Resources Research**, v. 27, p. 1177–1191, 1991.

MORENO, J. M.; ZUAZUA, E.; PÉREZ, B.; VELASCO, A.; de DIOS.; V. R. Rainfall patterns after fire differentially affect the recruitment of three Mediterranean shrubs. **Biogeosciences**, v. 8, p. 3721–3732, 2011.

MORGAN, P.; HARDY, C. C.; SWETNAM, T. W.; ROLLINS, M. G.; LONG, D. G. Mapping fire regimes across time and space: Understanding coarse and fine-scale fire patterns. **International Journal of Wildland Fire**, v. 10, n. 4, p. 329–342, 2001.

MURTHY, S. K. **On Growing Better Decision Trees from Data**. 1997. 206 f. Thesis (Doctor of Philosophy). Johns Hopkins University, Baltimore, Maryland.1997.

MURTHY, S. K. **Automatic Construction of Decision Trees from Data : A Multi-Disciplinary Survey**. Kluwer Academic Publichers, Boston. 1998.

NIBLETT, T.; BRATKO, I. **Learning Decision Rules in Noisy Domains**. Proc. Expert Systems 86, Cambridge: Cambridge University Press, 1986.

NIEMI, E.; LEE, K. **Wildfire and Poverty An overview of the interactions among wildfires, fire-related programs, and poverty in the Western States**. The Center for Watershed and Community Health. Portland, Oregon. 2001.

OJERIO, R.; MODELEY, C.; LYNN, K.; BANIA, N. Limited involvement of socially vulnerable populations in Federal Programs to mitigate wildfire risk in Arizona. **Natural Hazards Review**, v. 12, n. 1, p. 28–36, 2011.

OLIVEIRA, P. E.; PASSACANTILI, M. G. S. B. Influência antrópica em três ecótonos floresta/campo da Floresta Atlântica do Sudeste do Brasil: análise de micropartículas carbonizadas em solos superficiais. **Hoehnea**, v. 37, n. 4, p. 777–789, 2010.

OLIVEIRA, S.; OEHLER, F.; SAN-MIGUEL-AYANZ, J.; CAMIA, A.; PEREIRA, J. M. C. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. **Forest Ecology and Management**, v. 275, p. 117–129, 2012.

OLSEN, C. S.; SHARP, E. Building community-agency trust in fire-affected communities in Australia and the United States. **International Journal of Wildland Fire**, v. 22, p. 822–831, 2013.

OMENA, M. S. **Conjunto de ferramentas computacionais para análises agroclimáticas**. 2014. 106 f. Dissertação (Mestrado em Produção Vegetal).

Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes - RJ. 2014.

ONODA, M. Estudo sobre um algoritmo de árvore de decisão acoplado a um sistema de banco de dados relacional. 2001. 110 f. Dissertação (Mestrado) – Universidade Federal do Rio de Janeiro, Rio de Janeiro. 2001.

PADILLA, M.; VEGA-GARCIA, C. On the comparative importance of fire danger rating indices and their integration with spatial and temporal variables for predicting daily human-caused fire occurences in Spain. **International Journal of Wildland Fire**, v. 20, p. 46–58, 2011.

PARITSIS, J.; VEBLEN, T. T.; HOLZ, A. Positive fire feedbacks contribute to shifts from Nothofagus pumilio forests to fire-prone shrublands in Patagonia. **Journal of Vegetation Science**, v. 26, p. 89–101, 2015.

PARZEN, E. On Estimation of a Probability Density Function and Mode. **Annals of Mathematical Statistics**, v. 33, n. 3, p. 1065–1076, 1962.

PAUSAS, J. G. Changes in fire and climate in the Eastern Iberian Peninsula (Mediterranean Basin). **Climatic Change**, v. 63, p. 337–350, 2004.

PEREIRA, M. G.; TRIGO, R. M.; CAMARA, C. C.; PEREIRA, J. M. C.; LEITE, S. M. Synoptic patterns associated with large summer forest fires in Portugal. **Agricultural and Forest Meteorology**, v. 129, p. 11–25, 2005.

PERRY, D. A. The scientific basis of forestry. **Annual Review of Ecology and Systematics,** v. 29, p. 435-466, 1998.

PEW, K. L.; LARSEN, C. P. S. GIS analysis of spatial and temporal patterns of human-caused wildfires in the temperate rain forest of Vancouver Island, Canada. **Forest Ecology and Management**, v. 140, n. 1, p. 1–18, 2001.

PLUCINSKI, M. P. **A review of wildfire occurrence research**. Bushfire Cooperative Research Centre. 2012.

PODUR, J.; MARTELL, D. L.; CSILLAG, F. Spatial patterns of lightning-caused forest fires in Ontario , 1976 – 1998. **Ecological Applications**, v. 164, p. 1–20, 2003.

PODUR, J.; WOTTON, M. Will climate change overwhelm fire management capacity ? **Ecological Modelling**, v. 221, p. 1301–1309, 2010.

PRASAD, V. K.; BADARINATH, K. V. S.; EATURU, A. Biophysical and anthropogenic controls of forest fires in the Deccan Plateau , India. **Journal of Environmental Management**, v. 86, p. 1–13, 2008.

PREISLER, H. K.; BRILLINGER, D. R.; BURGAN, R. E.; BENOIT, J. W. Probability based models for estimation of wildfire risk. **International Journal of Wildland Fire**, v. 13, p. 133–142, 2004.

PRESTEMON, J. P. et al. **Wildfire Ignitions: A review of the science and recommendations for empirical modeling**. Asheville: [s.n.].

PYNE, S. J.; ANDREWS, P. L., LAVEN, R. D. **Introduction to Wildland Fire**. Second Edition. John Wiley & Sons, Inc., 1996.

QUINLAN, J. R. Induction of Decision Trees. **Machine Learning**, v, 1, p. 81-106, 1986.

QUINLAN, J. R. Inductive knowledge acquisition: A case study. In: QUINLAN, J. R. (Ed.). **Aplications of expert systems**. Addison-Wesley, 1987. p. 157-173.

QUINLAN, J. R. **C4.5: Programs for machine learning**. Morgan Kaufmann Publishers Inc. San Mateo, California. 1993.

REAMS, M. A.; HAINES, T. K.; RENNER, C. R.; WASCOM, M. W.; KINGRE, H. Goals , obstacles and effective strategies of wildfire mitigation programs in the Wildland – Urban Interface. **Forest Policy and Economics**, v. 7, p. 818–826, 2005.

RIESKAMP, J.; BUSEMEYER, J. R.; MELLERS, B. A. Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. **Journal of Economic Literature**, v. XLIV, p. 631–661, 2006.

RODRÍGUEZ-SILVA, F.; JULIO, G.; CASTILLO, M.; MOLINA, J. R.; HERRERA, M. A.; TORAL, M.; CERDA, C.; GONZÁLEZ, L. **Aplicación y adaptación del modelo SEVEIF para la evaluación socioeconómica del impacto de incêndios forestales em la Provincia de Valparaíso**. Chile: Agencia Española de Cooperación Internacional para el Desarrollo. 2010.

ROMERO-CALCERRADA, R.; NOVILLO, C. J.; MILLINGTON, J. D. A.; GOMEZ-JIMENEZ, I. GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain). **Landscape Ecology**, n. 23, p. 341–354, 2008.

ROMERO-CALCERRADA, R.; BARRIO-PARRA, F.; MILLINGTON, J. D. A.; NOVILLO, C. J. Spatial modelling of socioeconomic data to understand patterns of human-caused wildfire ignition risk in the SW of Madrid (central Spain). **Ecological Modelling**, v. 221, p. 34–45, jan. 2010.

ROSENBLATT, M. Remarks on Some Nonparametric Estimates of a Density Function. **Annals of Mathematical Statistics**, v. 27, n. 3, p. 832–837, 1956.

ROTHERMEL, R. C. **How to predict the spread and intensity of forest and range fires.** General Technical Report INT-143, USDA Forest Service. 1983.
ROY, D. P.; JIN, Y.; LEWIS, P. E.; JUSTICE, C. O. Prototyping a global algorithm for systematic fire-affected area mapping using MODIS time series data. **Remote Sensing of Environment**, v. 97, p. 137–162, 2005.

ROY, D. P.; BOSCHETTI, L.; JUSTICE, C. O.; JU, J. The collection 5 MODIS burned area product — Global evaluation by comparison with the MODIS active fi re product. **Remote Sensing of Environment**, v. 112, p. 3690–3707, 2008.

SAFAVIAN, S. R.; LANDGREBE, D. A Survey of Decision Tree Classifier Methodology. **IEEE Transactions on systems**, v. 21, n. 3, p. 660–674, 1991.

SAN-MIGUEL-AYANZ, J.; CARLSON, J. D.; ALEXANDER, M.; TOLHURST, K.; MORGAN, G.; SNEEUWJAGT, R. Current methods to assess fire danger potencial. In: CHUVIECO, E. (Ed.). **Wildland Fire Danger Estimation and Mapping** - The role of Remote Sensing Data. World Scientific Publishing, 2003. p. 21-61.

SANTOS, R. D. C. 2011. Disponível em: <http: //www.lac.inpe.br /~rafael.santos /Docs/ELAC/2010/Elac01_DM_Dia2.pdf>. Acesso: Agosto de 2017.

SEAMAN, D. E; POWELL, R. A. An Evaluation of the Accuracy of Kernel Density Estimators for Home Range Analysis. **Ecology**, v. 77, n. 7, p. 2075–2085, 1996.

SEMERARO, T.; MASTROLEO, G.; ARETANO, R.; FACCHINETTI, G.; ZURLINI, G.; PETROSILLO, I. GIS Fuzzy Expert System for the assessment of ecosystems vulnerability to fire in managing Mediterranean natural protected areas. **Journal of Environmental Management**, v. 168, p. 94–103, 2016.

SIKDER, I. U.; MAL-SARKAR, S.; MAL, T. K. Knowledge-Based Risk Assessment Under Uncertainty for Species Invasion. **Risk Analysis**, v. 26, n. 1, p. 239-252, 2006.

SILVERMAN, B. W. **Density estimation for statistics and data analysis**. Monographs on Statistics and Applied Probabilit, London: Chapman and Hall, 1986.

SOTO, M. E. C. The identification and assessment of areas at risk of forest fire using fuzzy methodology. **Applied Geography**, v. 35, n. 1–2, p. 199–207, 2012.

STEFANIDOU, M.; ATHANASELIS, S.; SPILIOPOULOU, C. Health impacts of fire smoke inhalation. **Inhalation Toxicology**, v. 20, p. 761-766, 2008.

STEINBERG, D.; COLLA, P**. CART: Classification and regression trees**. Software, Salford-System, San Diego, California. 1997.

STETLER, K. M.; VENN, T. J.; CALKIN, D. E. The effects of wildfire and environmental amenities on property values in northwest Montana, EUA. **Ecological Economics**, v. 69, n. 11, p. 2233–2243, 2010.

STOLE, F.; LAMBIN, E. F. Interprovincial and interannual differences in the causes of land-use fires in Sumatra , Indonesia. **Environmental Conservation**, v. 30, n. 4, p. 375–387, 2003.
SYPHARD, A.; RADELOFF, V. C.; KEELEY, J. E.; HAWKAKER, T. J.; CLAYTON, M. K.; STEWART, S. I.; HAMMER, R. B. Human influences on California fire regimes. **Ecological Applications**, v. 17, n. 5, p. 1388–1402, 2007.

SYPHARD, A. D.; RADELOFT, V.; KEULER, N. S.; TAYLOR, R. S.; HAWBAKER, T. J.; STEWART, S. I; CLAYTON, M. K. Predicting spatial patterns of fire on a southern California landscape. **International Journal of Wildland Fire**, n. 17, p. 602–613, 2008.

SYPHARD, A. D.; MASSADA, A. B.; BUTSIC, V.; KEELEY, J. E. Land use planning and wildfire: Development policies influence future probability of housing loss. **Plos One**, v. 8, n. 8, p. 1–12, 2013.

TEPLEY, A. J.; VEBLEN, T. T.; PERRY, G. L. W.; STEWART, G. H.; NAFICY, C. E. Positive Feedbacks to Fire-Driven Deforestation Following Human Colonization of the South Island of New Zealand. **Ecosystems**, v. 19, n. 8, p. 1325–1344, 2016.

THOMPSON, M. E.; CALKIN, D. E. Uncertainty and risk in wildland fire management: A review. **Journal of Environmental Management**, v. 92, p. 1895-1909, 2011.

THORNTHWAITE, C. W.; MATHER, J. R. The water balance. **Publications in climatology**. Drexel Institute of Technology, 1955.

TODD, B.; KOURTZ, P. H. **Predicting the daily occurrence of people-caused forest fires**. Petwawa National Forestry Institute. Information Report PI-X-103, Canadá. 1991.

TOWNSHEND, J.; HANSEN, M.;CARROL., M; DIMICELI, C.; SOHLBERG, R.; HUANG, C. **User Guide for the MODIS Vegetation Continuous Fields product Collection 5 version 1**. University of Maryland, 2011.

TROLLOPE, W. S. W.; TROLLOPE, L.; HARTNETT, D. C. Fire behaviour a key factor in the fire ecology of African grasslands and savannas. In: VIEGAS, D. X (Ed.). **Forest Fire Research & Wildland Fire Safety**. Rotterdam: Millpress. p. 1–14. 2002.

TURNER, R. Point patterns of forest fire locations. **Environmental Ecology Statistics**, v. 16, p. 197–223, 2009.

VADREVU, K. P.; EATURU, A.; BADARINATH, K. V. S. Fire risk evaluation using multicriteria analysis-a case study. **Environmental Monitoring and Assessment**, v. 166, p. 223–239, 2010.

VASCONCELOS, M. J. P.; SILVA, S.; TOMÉ, M.; ALVIM, M.; PEREIRA, J. M. C. Spatial Prediction of Fire Ignition Probabilities: Comparing Logistic Regression and Neural Networks. **Photogrammetric Engineering & Remote Sensing**, v. 67, n. 1, p. 73–81, 2001.

VASILAKOS, C.; KALABOKIDIS, K.; HATZOPOULOS, J.; MATSINOS, I. Identifying wildland fire ignition factors through sensitivity analysis of a neural network. **Natural Hazards**, n. 50, p. 125–143, 2009.

VAYSSIÈRES, M. P.; PLANT, R. E.; ALLEN-DIAZ, B. H. Classification trees: An alternative non-parametric approach for predicting species distributions. **Journal of Vegetation Science**, v. 11, p. 679–694, 2000.

VÁZQUEZ, A.; MORENO, J. M. Patterns of Lightning- , and People-Caused Fires in Peninsular Spain Patterns of Lightning- , and People-Caused Fires in Peninsular

Spain. **International Journal of Wildland Fire**, v. 8, n. 2, p. 103–115, 1998.

VEGA-GARCIA, C.; WOODARD, P. M.; TITUS, S. J.; ADAMOWICZ, W. L.; LEE, B. S. A Logit Model for Predicting the Daily Occurrence of Human Caused Forest Fires. **International Journal of Wildland Fire**, v. 5, n. 2, p. 101-111, 1995.

VEGA-GARCIA, C.; LEE, B. S.; WOODARD, P. M.; TITUS, S. J. Applying neural network technology to human-caused wildfire occurrence prediction. **AI Applications**, v. 10, n. 3, p. 9-18, 1996.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S-Plus**. Second Edition. New York: Springer, 1997.

VENN, T. J.; CALKIN, D. E. Challenges of Socio-economically Evaluating Wildfire Management on Non-industrial Private and Public Forestland in the Western United States. **Small-scale Forestry**, v. 8, p. 43–61, 2009.

VIANELLO, R. L.; ALVES, A. R. **Meteorologia básica e aplicações**.1.ed. Viçosa: UFV, 2004. 449p.

VIEGAS, D. X.; VIEGAS, M. T.; FERREIRA, A. D. Moisture content of fine forest fuels and fire occurrence in central Portugal. **International Journal of Wildland Fire**, v. 2, p. 69-86, 1992.

VIEGAS, D. X.; BOVIO, G.; FERREIRA, A.; NOSENZO, A.; SOL, B. Comparative Study of Various Methods of Fire Danger Evaluation in Southern Europe. **International Journal of Wildland Fire**, v. 9, n. 4, p. 235–246, 1999.

VILAR, L.; WOOLFORD, D. G.; MARTELL, D. L.; MARTÍN, M. P. A model for predicting human-caused wildfire occurrence in the region of Madrid, Spain. **International Journal of Wildland Fire**, v. 19, p. 325–337, 2010.

WAGNER, V. **Development and Structure of the Canadian Forest Fire Weather Index System**. Canadian Forestry Service. Forestry Technical Report 35. Ottawa. 1987.

WATKINS, C. J. C. H. Combining Cross-Validation and Search. In: BRATKO, I.; LAVRAC, N. (Eds.). **Progress in Machine Learning, Proc. EWSL 87**. Wilmslow: Sigma Press, 1987. p. 79-87.
WEINTRAUB, A.; ROMERO, C. Operations Research Models and the Management of Agricultural and Forestry Resources: A Review and Comparison. **Interfaces**, v. 36, n. 5, p. 446–457, 2006.

WEISS, L. A. Bankruptcy resolution: Direct costs and violation of priority of claims. **Journal Financial Economics**, v. 27, n. 2, p. 285-314, 1990.

WHELAN, R. J. **The ecology of fire**. Cambridge Univerity Press, 1995.

WILLIAMS, D. R. et al. Community wildfire protection planning: The importance of

framing, scale, and building sustainable capacity. **Journal o Forestry**, v. 110, n. 8, p. 415–420, 2012.

WORTON, B. J. Kernel Methods for Estimating the Utilization Distribution in Home-Range Studies. **Ec**, v. 70, n. 1, p. 164–168, 1989.

WOTTON, B. M.; MARTELL, D. L.; LOGAN, K. A. Climate change and people-caused forest fire occurence in Ontario. **Climatic Change**, v. 60, p. 275–295, 2003.

WOTTON, M.; NOCK, C. A.; FLANNIGAN, M. Forest fire occurrence and climate change in Canada. **International Journal of Wildland Fire**, v. 19, p. 253–271, 2010.

WU, Z.; HE, H. S.; YANG, J.; LIU, Z.; LIANG, Y. Relative effects of climatic and local factors on fi re occurrence in boreal forest landscapes of northeastern China. **Science of the Total Environment**, v. 493, p. 472–480, 2014.

WU, X.; KUMAR, V. **The Top Ten Algorithms in Data Mining**. Chapman & Hall Book, 2009.

YAKUBU, I.; MIREKU-GYIMAH, D.; DUKER, A. A. Review of methods for modelling forest fire risk and hazard. **African Journal of Environmental Science and Technology**, v. 9, n. 3, p. 155–165, 2015.

YANG, J.; HE, H. S.; SHIFLEY, S. R.; GUSTAFSON, E. J. Spatial patterns of modern period human-caused fire occurrence in the Missouri Ozark Highlands. **Forest Science**, v. 53, n. 1, p. 1–15, 2007.

ZHANG, Y.; LIM, S.; SHARPLES, J. J. Modelling spatial patterns of wildfire occurrence in South-Eastern Australia. **Geomatics, Natural Hazards and Risk**, v. 7, n. 6, p. 1800–1815, 2016.

ZHIJUN, T.; JIQUAN, Z.; XINGPENG, L. GIS-based risk assessment of grassland fire disaster in western Jilin province , China. **Stochastic Environmental Research and Risk Assessment**, v. 23, p. 463–471, 2009.

ZIGHED, D. A.; RAKOTOMALALA, R. **Graphes d'Induction** – **Apprentissage et Data Mining**. Paris: Hermes Science Publications, 2000.

APÊNDICE A – CÓDIGO DE PROGRAMAÇÃO DAS REGRAS DA ÁRVORE DE DECISÃO.

```
# -*- coding: utf-8 -*-
""
    Identificacao: Unidades de manejo do fogo no estado do Espírito Santo
    Autor: Ronie Silva Juvanhol
           Thiago Tuler
    Utilidades: Dado os preditores de entrada, esse algoritmo fornece zonas de
risco de incêndio por meio de um conjunto de regras gerados a partir de uma
árvore de decisão.
""
import os
import openpyxl

tree = {
    'index' : 0,
    'right': {
        'index': 5,
        'right': {
            'index': 0,
            'right': {
                'index': 2,
                'right':{
                    'index' : 3,
                    'right' :{
                        'index' :2,
                        'right': 0.044, # terminal node 38
                        'value': 1343.19,
                        'left': 0.017 # terminal node 37
                    },
                    'value': 9.5,
                    'left': 0.004 # terminal node 36
                },
                'value': 1130.84399,
                'left': {
                    'index': 3,
                    'right':{
                        'index': 2,
                        'right': 0.131, # terminal node 35
                        'value': 1111.9,
                        'left': 0.054 # terminal node 34
                    },
                    'value': 281.5,
                    'left': 0.038 # terminal node 33
                }
            },
            'value': 6.08065,
```

```
        'left': {
            'index': 1,
            'right': {
                'index': 0,
                'right': 0.175359, # terminal node 32
                'value': 5.89440,
                'left': 0.0575384 # terminal node 31
            },
            'value': 624.57501,
            'left': {
                'index': 1,
                'right': 0.155, # terminal node 30
                'value': 506.88501,
                'left': 0.0315917 # terminal node 29
            }
        }
    },
    'value': 55.50000,
    'left': {
        'index': 2,
        'right': {
            'index': 6,
            'right': 0.0473021, # terminal node 28
            'value': 24.66975,
            'left': {
                'index': 3,
                'right': {
                    'index': 2,
                    'right': {
                        'index': 1,
                        'right': 0.571, # terminal node 27
                        'value': 1120.41,
                        'left': 0.097 # terminal node 26
                    },
                    'value': 1369.48,
                    'left': 0.101 # terminal node 25
                },
                'value': 76.5,
                'left': 0.038 # terminal node 24
            }
        },
        'value': 1169.50452,
        'left': {
            'index': 6,
            'right': 0.0699455, # terminal node 23
            'value': 25.03165,
            'left': 0.185039 # terminal node 22
        }
    }
},
```

```
'value': 3.63463,
'left': {
    'index': 4,
    'right': {
        'index': 2,
        'right': {
            'index': 6,
            'right': {
                'index': 0,
                'right': {
                    'index': 2,
                    'right': 0.571113, # terminal node 21
                    'value': 1382.34058,
                    'left': 0.239003 # terminal node 20
                },
                'value': 2.35704,
                'left': {
                    'index': 2,
                    'right': 0.338749, # terminal node 19
                    'value': 1400.65942,
                    'left': 0.0901171 # terminal node 18
                }
            },
            'value': 24.68275,
            'left': {
                'index': 1,
                'right': {
                    'index': 0,
                    'right': 0.52072, # terminal node 17
                    'value': 2.35704,
                    'left': 0.0606344 # terminal node 16
                },
                'value': 854.57001,
                'left': {
                    'index': 1,
                    'right': 0.748276, # terminal node 15
                    'value': 684.70502,
                    'left': 0.281141 # terminal node 14
                }
            }
        },
        'value': 1350.50757,
        'left': {
            'index': 1,
            'right': 0.250526, # terminal node 13
            'value': 859.72498,
            'left': 0.121 # terminal node 12
        }
    },
    'value': 8,
```

```
'left': {
    'index': 2,
    'right': {
        'index': 0,
        'right': 0.139411, # terminal node 11
        'value': 3.26368,
        'left': {
            'index': 2,
            'right': {
                'index': 5,
                'right': 0.0499697, # terminal node 10
                'value': 44.50000,
                'left': {
                    'index': 2,
                    'right': 0.131192, # terminal node 09
                    'value': 1236.40845,
                    'left': 0.0168376 # terminal node 08
                }
            },
            'value': 1138.03552,
            'left': 0.109868 # terminal node 07
        }
    },
    'value': 1109.07446,
    'left': {
        'index': 1,
        'right': {
            'index': 1,
            'right': {
                'index': 1,
                'right': 0.175612, # terminal node 06
                'value': 864.86499,
                'left': 0.434419 # terminal node 05
            },
            'value': 841.71002,
            'left': 0.110923 # terminal node 04
        },
        'value': 556.09497,
        'left': {
            'index': 2,
            'right': {
                'index': 2,
                'right': 0.18456, # terminal node 03
                'value': 1090.74597,
                'left': 0.330305 # terminal node 02
            },
            'value': 1034.39551,
            'left': 0.0835317 # terminal node 01
        }
    }
```

```
        }
    }
}

class FireAlert (object):
    #Inicia a variável relativa ao caminho dos arquivos utilizados
    def __init__(self):
        self.__location__ = os.path.realpath(
            os.path.join(os.getcwd(), os.path.dirname(__file__)))

    #Converte o nome da classe de uso da terra para um coeficiente numérico
    def str_to_int(self, uso_terra):
        if uso_terra == "Agricultura":
            return 1
        elif uso_terra == "Areas urbanas":
            return 2
        elif uso_terra == "Curso d'agua":
            return 3
        elif uso_terra == "Floresta natural":
            return 4
        elif uso_terra == "Solo Exposto":
            return 5
        elif uso_terra == "Pastagem":
            return 6
        elif uso_terra == "Silvicultura":
            return 7
        elif uso_terra =="Manguezais":
            return 8
        elif uso_terra == "Areas alagadas":
            return 9
        elif uso_terra == "Restinga":
            return 10

    # Faz uma predição com uma árvore de decisão
    def predict(self, node, row):
        if row[node['index']] <= node['value']:
            if isinstance(node['left'], dict):
                return self.predict(node['left'], row)
            else:
                return node['left']
        else:
            if isinstance(node['right'], dict):
                return self.predict(node['right'], row)
            else:
                return node['right']

    #Utiliza os dados da tabela para localizar o coeficiente de incêndio
    def process(self):
        print "Abrindo o arquivo..."
```

```
        workbook1 = openpyxl.load_workbook(os.path.join(self.__location__,
'dados_entrada.xlsx'), read_only=True)
        workbook2 = openpyxl.Workbook(write_only=True)
        print "Lendo a planilha..."
        worksheet1 = workbook1.active
        worksheet2 = workbook2.create_sheet()
        for index, cells in enumerate(worksheet1.iter_rows()):
            if index != 0:
                row = [float(cells[2].value), float(cells[3].value), float(cells[4].value),
float(cells[5].value), FireAlert().str_to_int(cells[6].value), float(cells[7].value),
float(cells[8].value)]
                score = FireAlert().predict(tree, row)
                #print 'Calculando {0} de {1}\r'.format(index+1, worksheet1.max_row),
                print index+1
                worksheet2.append([float(cells[0].value), float(cells[1].value),
float(cells[2].value), float(cells[3].value), float(cells[4].value), float(cells[5].value),
cells[6].value, int(cells[7].value), float(cells[8].value), float(cells[9].value), score])
            else:
                worksheet2.append([cells[0].value, cells[1].value, cells[2].value,
cells[3].value, cells[4].value, cells[5].value, cells[6].value, cells[7].value,
cells[8].value, cells[9].value, cells[10].value])
        workbook2.save(os.path.join(self.__location__, 'dados_saida.xlsx'))

if __name__ == '__main__':
    FireAlert().process()
```