UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO MESTRADO PROFISSIONAL EM ENGENHARIA DE PRODUÇÃO

Camila da Silveira Machado

APLICAÇÕES DE TÉCNICAS MULTIVARIADAS EM BANCO DE DADOS DE EVENTOS DE INCÊNDIO

Porto Alegre

2019

Camila da Silveira Machado

Aplicações de técnicas multivariadas em banco de dados de eventos de incêndio

Dissertação submetida ao Programa de Pós-Graduação Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Profissional, na área de concentração em Sistemas de Produção.

Orientador: Prof. Michel José Anzanello, *Ph D*.

Porto Alegre

2019

Camila da Silveira Machado

Aplicações de técnicas multivariadas em banco de dados de eventos de incêndio

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Profissional e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel José Anzanello, *Ph.D*.

Orientador PMPEP/UFRGS

_____

Prof. Ricardo Augusto Cassel, Dr.

Coordenador PMPEP/UFRGS

Banca Examinadora:

Prof. Alessandro Kahmann, Dr. (IMEF/FURG)

Prof. Flávio Sanson Fogliatto, PhD (PPGEP/UFRGS)

Prof. Tarcísio Saurin, Dr. (PPGEP/UFRGS)

DEDICATÓRIA

Aos meus amados pais Jair e Sandra (*in memorian*), à minha irmã e, em especial, ao meu marido Alexandre.

AGRADECIMENTOS

Ao meu marido Alexandre por seu meu melhor amigo e meu amor verdadeiro, pelo companheirismo, encorajamento e dedicação.

Aos meus pais Jair e Sandra (*in memorian*) pelo amor incondicional e, sobretudo, pelo investimento na minha educação, a maior e mais valiosa herança que poderiam deixar.

À minha irmã Jaqueline pelo carinho e incentivo.

À minha amiga, também colega de trabalho e de pós-graduação, Eveline pela parceria que nasceu durante as aulas do Programa e que se tornou uma próspera amizade.

Ao meu orientador Professor Michel José Anzanello, *Ph.D.*, pela oportunidade, assistência e confiança.

Aos professores, técnicos administrativos e demais envolvidos que compõem a estrutura do Programa de Pós-Graduação em Engenharia de Produção na UFRGS pelo ensino de referência e organização.

RESUMO

Incêndios em ambientes residenciais respondem por um quarto das ocorrências anuais totais de incidentes com fogo nos EUA, sendo responsáveis, nos últimos anos, por aproximadamente oitenta por cento das mortes e das lesões em civis nos EUA. Estatísticas de incêndio têm sido estudadas para o entendimento do comportamento e tendências de tais eventos, com vistas a auxiliar no desenvolvimento de diretivas para prevenção, proteção e combate ao fogo e alocação de recursos de segurança. Nesse sentido, percebe-se o sucessivo emprego de ferramentas multivariadas para o reconhecimento de padrões nas ocorrências de incêndios domiciliares, em particular a análise de cluster. O primeiro artigo objetiva a formação de grupos homogêneos de ocorrências de incêndios residenciais, a partir do uso de ferramentas multivariadas. O método proposto combina procedimentos hierárquicos e não hierárquicos de clusterização, Análise de Componentes Principais (ACP) e Silhouette Index (SI) para avaliar a qualidade dos agrupamentos gerados. O segundo artigo propõe o emprego de técnicas de mineração de dados para selecionar as variáveis mais relevantes para classificação das ocorrências de incêndio em classes. A sistemática apresentada combina a técnica "omita uma variável de cada vez" e a ferramenta de classificação K-Nearest Neighbor (KNN), objetivando selecionar o melhor subconjunto de variáveis independentes que descrevem as características dos eventos de incêndio para a predição da variável de resposta (classe de causa do incêndio), avaliando-se também o grau de relevância das variáveis para o modelo de classificação. Em ambos artigos, as metodologias propostas foram aplicadas ao conjunto de observações de incêndios residenciais reportados nos estados americanos da Flórida e do Texas, no período de 2010 a 2014, e associados a cinco principais causas.

Palavras-chave: Classificação. Análise de clusterização. Análise de Componentes Principais. *K-Means*. *Silhouette Index*. Ferramentas de mineração de dados. Seleção de variáveis. *K-Nearest Neighbor*. Incêndio.

ABSTRACT

Residential fires answer for one quarter of the total annual occurrences of fire incidents in the US, accounting in recent years for approximately eighty percent of deaths and injuries in civilians in the United States. Fire statistics were studied to understand the behavior and the trends of such events, aiming to assist in the development of directives for prevention, protection, firefighting and allocation of safety resources. In this regard, it is possible to identify the successive use of multivariate analysis techniques for the recognition of residential fire patterns, in particular, the cluster analysis. The first article aims to establish homogeneous groups of residential fires occurrences, based on the use of multivariate tools. The proposed method combines hierarchical and non-hierarchical procedures of clustering, Principal Component Analysis (PCA) and Silhouette Index (SI) to evaluate the quality of the generated clusters. The second article proposes the use of data mining techniques to select the most relevant variables to classifying fire occurrences. The presented system combines the "leave one variable out at a time" and the classification tool K-Nearest Neighbor (KNN), aiming to select the best independent variables that describe the attributes of fire events available to predict the response variable (the fire cause class), also evaluating the degree of relevance of these variables for the classification model. In both articles, the proposed methodologies were applied to the group of residential fires incidents reported in the states of Florida and Texas between 2010 and 2014, associating these occurrences to five main causes.

Keywords: Classification. Cluster analysis. Principal component analysis. K-Means. Silhouette Index. Data mining tools. Selection of variables. K-Nearest Neighbor. Fire.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações Iniciais

Incêndios em ambientes domésticos são responsáveis por significativas fatalidades e perdas sócio-econômicas irreparáveis. Nos EUA, as estatísticas apontam que, embora somente 26% das ocorrências anuais totais de incidentes com fogo ocorram em ambientes residenciais, incêndios domésticos são responsáveis, em média, por 81% das mortes e 77% das lesões em civis no ano (UFSA, 2017). Objetivando reduzir essas ocorrências, nota-se o avanço de tecnologias em prevenção, proteção e combate contra incêndio e o desenvolvimento de legislações e normas de segurança contra incêndios. Contudo, percebe-se que as regulamentações focam-se em sistemas de prevenção e combate ao fogo em edifícios, construções multifamiliares, industriais e comerciais, reservando menor atenção para edificações residenciais (JENNINGS, 2013; TROITZSCH, 2016).

A fim de compreender as tendências das ocorrências, desenvolver estratégias de segurança contra o fogo e fundamentar a alocação de recursos, a análise estatística dos registros de incêndios se tornaram um importante meio para prevenção, proteção e combate a incêndios (CEYHAN et al., 2013; LIZHI; AIZHU, 2008). Nesse sentido, ferramentas multivariadas têm sido aplicadas a bancos de dados que descrevem ocorrências de incêndio objetivando diminuir a complexidade do problema e compreender as tendências das ocorrências de incêndio. Neste contexto, nota-se grande número de pesquisas abordando análise de clusterização para agrupar os registros de incêndio de acordo com suas similaridades, possibilitando a identificação de perfis característicos nas ocorrências. Conjuntamente, percebe-se incremento na utilização de técnicas de seleção de variáveis em dados de incidentes de incêndio a fim de aprimorar o desempenho dos modelos de classificação dos eventos.

Diante do exposto, essa dissertação está apoiada em dois artigos. O primeiro artigo utiliza ferramentas multivariadas para gerar grupos homogêneos de ocorrências de incêndios domésticos combinando procedimentos hierárquicos e não hierárquicos de clusterização, bem como Análise de Componentes Principais (ACP) e *Silhouette Index* (SI) para avaliação da qualidade dos *clusters* gerados. O segundo artigo emprega a técnica “omita uma variável de cada vez” e a ferramenta de classificação *K-Nearest Neighbor* (KNN) com vistas a selecionar o melhor subconjunto de variáveis descritivas para classificação das ocorrências de incêndio em classes.

1.2 Objetivos

O objetivo geral dessa dissertação é aplicar abordagens multivariadas a dados de ocorrências de incêndios com vistas a caracterizar padrões de similaridade entre tais ocorrências.

Como objetivos específicos, listam-se:

a) Gerar agrupamentos homogêneos de ocorrências de incêndios residenciais através da técnica de clusterização *K-Means*, e avaliar a qualidade do procedimento via *Silhouette Index* (SI);
b) Interpretar os agrupamentos gerados, comparando os resultados obtidos pela clusterização via variáveis latentes geradas pela ACP;
c) Selecionar as variáveis mais relevantes para inserção de eventos de incêndio em classes associadas a causas específicas dos eventos;
d) Analisar qualitativamente o impacto das variáveis retidas para formação dos agrupamentos.

1.3 Justificativa do Tema

A literatura traz vasto relato acerca de catástrofes causadas por incêndios que destruíram grandes centro urbanos nos Estados Unidos, como Pittsburgh em 1845, Chicago em 1871, Boston em 1872 e São Francisco em 1906; ver Chen et al. (2015). As estatísticas de eventos de incêndio alertam para a necessidade de ações de prevenção, proteção e combate contra o fogo com vistas a reduzir mortes, lesões, danos à propriedade e demais perdas financeiras (TROITZSCH, 2016). Apesar dos avanços em regulamentações e tecnologias de segurança contra o fogo, estudos sobre incêndio doméstico ainda não são consenso. Percebem-se maiores progressos na redução das perdas por incêndio em propriedades não residenciais, visto à legislação governamental e às políticas mercadológicas em que estão inseridas (JENNINGS, 2013; SPINARDI et al., 2016). De tal forma, por focar-se em dados de incêndios em contexto residencial, tal pesquisa encontra justificativa no âmbito prático.

Em termos teóricos, percebe-se substancial utilização de técnicas multivariadas, principalmente a análise de agrupamentos (clusterização), para determinar padrões similares em dados descritivos de eventos de incêndios. Técnicas de classificação também têm sido utilizadas na análise de ocorrências de incêndios, embora não tenha se encontrado estudos integrando a técnica *K-Nearest Neighbor* (KNN) e a sistemática

"omita uma variável por vez" para seleção de variáveis em dados de incêndios. Nessa lacuna, encontra-se possibilidade para o emprego de ferramentas de mineração de dados visando identificar as variáveis mais relevantes para classificação de ocorrências de incêndio em classes associadas às causas do incêndio. Neste sentido, identifica-se oportunidade para evoluir na adaptação de métodos que visem a agrupar os incidentes de incêndio e facilitar a interpretação de grandes quantidades de observações que podem não apresentar significado perceptível quando isoladas.

1.4 Procedimentos Metodológicos

Orientada à solução de um problema específico, a pesquisa se classifica como aplicada quanto à sua natureza. Com relação à abordagem, este trabalho se caracteriza como quantitativo, visto a seleção de variáveis e técnicas multivariadas utilizadas na análise dos dados. Quanto aos objetivos, a pesquisa é descritiva aproximando-se à explicativa, isso porque objetiva, além de estabelecer associações entre variáveis, determinar a natureza dessas relações. Quanto aos procedimentos, trata-se de uma pesquisa bibliográfica e documental (GIL, 2002).

A dissertação é desdobrada em dois artigos. O primeiro artigo apresenta a sistemática apoiada em técnicas para a formação de grupos homogêneos de ocorrências de incêndios residenciais. O estudo inicialmente agrupa as observações utilizando as variáveis originais do banco de dados. Na sequência, com vistas a aprimorar a qualidade dos agrupamentos formados, os dados originais são transformados através da Análise de Componentes Principais (ACP), repetindo-se a abordagem de clusterização nas variáveis latentes oriundas da ACP. O método proposto é operacionalizado em cinco etapas: (i) geração do dendrograma para definição do número de clusters a serem formados, (ii) clusterização via K-means utilizando as variáveis originais que descrevem os eventos de incêndio, (iii) avaliação da qualidade dos agrupamentos gerados através do *Silhouette Index*, (iv) interpretação dos agrupamentos formados, e (v) repetição das etapas (i) a (iv) para dados transformados pela ACP.

O segundo artigo aborda técnicas de mineração de dados para selecionar as variáveis mais relevantes para classificação das observações (eventos de incêndio) em classes (causas específicas dos eventos). O estudo combina o método "omita uma variável de cada vez" com a ferramenta de classificação KNN. O método proposto consiste na (i) divisão do banco de dados em conjuntos de treino e teste, (ii) aplicação do KNN combinado com

técnica “omita uma variável por vez” para seleção de variáveis, (iii) geração do gráfico de frequência de retenção de variáveis, e (iv) análise qualitativa do impacto das variáveis retidas sobre os argumentos gerados. As variáveis identificadas como mais relevantes são então qualitativamente discutidas.

1.5 Estrutura da Dissertação

A dissertação está dividida em quatro capítulos. No primeiro são apresentadas as considerações iniciais, os objetivos gerais e específicos, a justificativa do tema, os procedimentos metodológicos, a estrutura da dissertação e as delimitações do estudo.

No segundo capítulo, que apresenta o primeiro artigo da dissertação, traz os *clusters* de ocorrências de incêndios residenciais gerados a partir da combinação de técnicas hierárquicas e não hierárquicas. O artigo aplica a metodologia proposta nas variáveis originais do conjunto de observações de incêndio e nas variáveis latentes oriundas da ACP, avaliando a qualidade dos agrupamentos formados através do *Silhouette Index* e interpretando os agrupamentos formados. Apresentam-se os resultados obtidos separadamente para o banco de dados de incêndio da Flórida e do Texas, comparando-os durante a discussão dos resultados.

O terceiro capítulo apresenta o segundo artigo dessa dissertação, identificando o melhor subconjunto de variáveis independentes (variáveis descritivas do tipo de residência, dentre outras) para a classificação do incidente de acordo com a sua causa. Os modelos de predição são avaliados e validados em porções de treino e teste por meio do KNN combinado com a técnica “omita uma variável por vez” para seleção das variáveis mais informativas. A frequência de retenção de variáveis é analisada separadamente para cada banco de dados (Flórida e Texas). A influência das variáveis retidas nos agrupamentos gerados e variáveis identificadas como mais relevantes são qualitativamente avaliadas.

O quarto capítulo é composto pelas considerações finais e possíveis desdobramentos deste trabalho.

1.6 Delimitações da Pesquisa

O estudo de caso dessa dissertação está delimitado aos dados provenientes de dois dos onze relatórios disponibilizados pela *National Fire Incident Reporting System* (NFIRS): o Módulo de Incêndio, que descreve cada evento de incêndio, e o Módulo de Incêndio de

Estrutura, que descreve a estrutura que sofreu essa ocorrência. Ressalta-se que os incidentes de incêndio notificados pelos departamentos de bombeiros são de preenchimento voluntário e, portanto, a NFIRS não consiste em um recenseamento de incidentes de incêndio ou acidentes (USFA, 2015; 2017).

No pré-processamento dos bancos de dados analisados, foram selecionados apenas os eventos de incêndio em edificações residenciais identificadas como construção finalizada e fechada e de pavimento único que ocorreram de forma isolada, envolvendo apenas essa edificação na ocorrência.

1.7 Referências

CEYHAN, E.; ERTUGAY, K.; DÜZGÜN, S. Exploratory and inferential methods for spatio-temporal analysis of residential fire clustering in urban areas. **Fire Safety Journal**, v. 58, p. 226-239, maio 2013.

CHEN, H; PITTMAN, W. C.; HATANAKA, L. C.; HARDING, B. Z.; BOUSSOUF, A.; MOORE, D. A.; MILKE, J. A.; MANNAN, M., A. Integration of process safety engineering and fire protection engineering for better safety performance. **Journal of Loss Prevention in the Process Industries**, v. 37, p. 74-81, set. 2015.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002. 176 p.

JENNINGS, C. R. Social and economic characteristics as determinants of residential fire risk in urban neighborhoods: A review of the literature. **Fire Safety Journal**, v. 62, parte A, p.13–19, nov. 2013.

LIZHI, W.; AIZHU, R. Urban Fire Risk Clustering Method Based on Fire Statistics. **Tsinghua science and technology**, v. 13, suplemento 1, p. 418-422, out. 2008.

TROITZSCH, J. H. Fires, statistics, ignition sources, and passive fire protection measures. **Journal of Fire Sciences**, v. 34 (3), p. 171-198, fev. 2016.

SPINARDI, G.; BISBY, L.; TORERO, J. A Review of Sociological Issues in Fire Safety Regulation. **Fire Technology**, p. 1-27, 26 jul. 2016.

UNITED STATES FIRE ADMINISTRATION. Federal Emergency Management Agency. **Fire in the United States**, 2005-2014. 18 ed., jan. 2017.

UNITED STATES FIRE ADMINISTRATION. National Fire Data Center. **National Fire Incident Reporting System 5.0** - Complete Reference Guide, 515 p., jan. 2015.

2 PRIMEIRO ARTIGO

TÉCNICAS DE CLUSTERIZAÇÃO PARA AGRUPAMENTO DE EVENTOS DE INCÊNDIOS

Resumo: *Incêndios em ambientes domésticos respondem por um quarto das ocorrências anuais totais de incidentes com fogo nos EUA, sendo responsáveis por 81% das mortes e 77% das lesões em civis ao ano. O estudo dessas estatísticas tem possibilitado o desenvolvimento de diretrizes para prevenção, proteção e combate ao fogo. Para tanto, percebe-se crescente emprego de ferramentas multivariadas para o reconhecimento de tendências das ocorrências de incêndios domiciliares, em particular a análise de cluster. Este artigo propõe o uso de técnicas multivariadas para a formação de grupos homogêneos de ocorrências de incêndios residenciais. A metodologia proposta combina procedimentos hierárquico e não hierárquico de clusterização, bem como a Análise de Componentes Principais (ACP) com vistas a aprimorar a qualidade da clusterização avaliada através do Silhouette Index (SI). A sistemática proposta foi aplicada ao conjunto de observações de incêndios residenciais reportados nos estados americanos da Flórida e do Texas, no período de 2010 a 2014. O SI médio obtido com as observações de incêndio na Flórida foi de 0,8481 e de 0,9141 no Texas. Os agrupamentos gerados pelos dados dos dois estados apresentaram semelhanças, sendo majoritariamente justificados pela presença e efetividade de operação de detectores automáticos de fogo.*

Palavras-chave: Classificação. Análise de clusterização. Análise de Componentes Principais. K-Means. Índice Silhouette. Incêndio.

2.1 Introdução

Embora o número de incêndios tenha diminuído consideravelmente em anos recentes, tal evento ainda responde por muitas fatalidades e perdas econômicas, justificando o desenvolvimento de estratégias para proteger a vida e a propriedade, bem como iniciativas para extinguir ou retardar o fogo por meio de estratégias de combate, prevenção e proteção. Regulamentos de proteção e prevenção contra o fogo remontam à Roma antiga, dada a magnitude dos prejuízos e perdas decorrentes de grandes incêndios (TROITZSCH, 2016). A literatura traz vasto relato acerca de catástrofes causadas por incêndios que destruíram grandes centro urbanos nos Estados Unidos. Em Pittsburgh (Pensilvânia), em 1845, cerca de 1.200 edifícios foram destruídos, 12.000 pessoas ficaram desabrigadas, e as perdas foram estimadas em 12 milhões de dólares na época (COOK JR., 1968). Em Chicago, em 1871, um incêndio destruiu cerca de 200 acres, desabrigando 100 mil habitantes, ocasionando uma perda de aproximadamente 200 milhões de dólares em propriedades e 300 mortos (MILLER, 1996). Em 1872, em Boston, um incêndio

extinguiu cerca de 26 hectares do centro da cidade, destruindo grande parte do distrito financeiro, causando danos de US$ 75 milhões (HORNBECK; KENISTON, 2017). Apesar de terem sido causados por um grande terremoto, em 1906, em São Francisco (Califórnia), incêndios devastadores irromperam na cidade e duraram vários dias, resultando na morte de mais de 3.000 pessoas e na destruição de mais de 80% da cidade; perdas econômicas foram estimadas em 500 milhões de dólares (SCAWTHORN et al., 2005).

Com vistas a evitar ocorrências de tamanho impacto, tecnologias avançadas em proteção contra incêndio e novas técnicas e materiais de construção têm auxiliado na prevenção contra o fogo. Além disso, países desenvolvidos e emergentes têm progressivamente adotado legislações de segurança contra incêndios. Embora as estatísticas mostrem que a maioria das perdas de incêndio ocorre em edifícios residenciais - correspondendo a 26% das ocorrências anuais totais de incidentes com fogo, incêndios domésticos contabilizam, em média, 81% das mortes e 77% das lesões em civis no ano (UFSA, 2017) - as regulamentações, na prática, são mais severas com sistemas de prevenção e combate ao fogo em edifícios altos, construções multifamiliares e comerciais do que com edificações residenciais (JENNINGS, 2013; TROITZSCH, 2016).

Amplamente utilizadas, as análises estatísticas de dados de incêndios se tornaram um importante meio para compreensão de tais eventos (LIZHI; AIZHU, 2008). A gestão de risco de incêndio objetiva compreender as tendências das ocorrências e fundamentar a seleção de medidas de prevenção e combate ao fogo, bem como visam ao desenvolvimento de estratégias para alocação eficiente dos recursos (CEYHAN et al., 2013). Essa gestão, ao envolver o estudo das estatísticas de incêndio à adaptação de regulamentos de segurança contra incêndios, tem como propósito a redução das perdas por incêndio (TROITZSCH, 2016).

No que diz respeito a abordagens quantitativas voltadas à análise de dados de incêndios, percebe-se crescente utilização de ferramentas multivariadas objetivando diminuir a complexidade do problema, como Lizhi e Aizhu (2008), Palamara et al. (2011), Higgins et al. (2013; 2014) e Chen et al. (2017) aplicaram K-means para identificar a similaridade entre as ocorrências de incêndio coletadas e Orozco et al. (2012), Wuschke et al (2013) e Xin e Huang (2013) aplicaram ferramentas de mapeamento de *clusters* para identificar a significância estatística dos agrupamentos gerados. Dentre as principais abordagens multivariadas está a análise de agrupamentos (clusterização), a qual busca padrões

similares em um conjunto de dados, agrupando observações (eventos de incêndios, no caso deste estudo) de forma que as mesmas sejam semelhantes dentro do mesmo grupo, mas mutuamente excludentes entre os grupos (HAIR JR. et al., 2009). Dada a simplicidade dos seus preceitos e flexibilidade de aplicação, a análise de clusterização tem sido vastamente aplicada na análise de dados descritivos de eventos de incêndios. Lizhi e Aizhu (2008) desenvolveram um método de classificação automática em relação ao risco de incêndio por meio de um modelo de agrupamento de *K-Means*, enquanto que Clare et al. (2012) detalharam a clusterização para avaliar o impacto do serviço de bombeiros e da campanha de educação pública de prevenção de incêndio na redução da frequência e gravidade de incêndios domésticos. Orozco et al. (2012), através do georreferenciamento, aplicaram a clusterização para detectar a distribuição espacial e temporal dos incêndios florestais e analisar suas causas de ignição. Wuschke et al. (2013) compararam as distribuições espaço-temporais de incêndios residenciais e assaltos a residências por agrupamentos, ao passo que Xin e Huang (2013) apresentaram um modelo de análise de risco baseado em *clusters* de cenários de incêndio, de supressão automática do fogo e de aspectos comportamentais.

Embora diversos estudos tenham se apoiado em técnicas de clusterização para encontrar padrões similares em dados descritivos de eventos de incêndios, percebe-se espaço para evoluir na adaptação de ferramentas com vistas a agrupar os incidentes de incêndio e permitir a interpretação de grandes quantidades de observações que podem não apresentar significado claro quando dispersas. Este artigo propõe uma sistemática apoiada em técnicas multivariadas para a formação de grupos homogêneos de ocorrências de incêndios residenciais. O método proposto é operacionalizado em cinco etapas: (i) geração do dendrograma para definição do número de clusters a serem gerados, (ii) clusterização via *K-Means* utilizando as variáveis originais que descrevem os eventos de incêndio, (iii) avaliação da qualidade dos agrupamentos gerados através do *Silhouette Index*, (iv) interpretação dos agrupamentos formados, e (v) repetição das etapas (i) a (iv) para dados transformados pela Análise de Componentes Principais (ACP). A transformação das variáveis originais via ACP visa a aprimorar a qualidade dos agrupamentos gerados através da utilização de novas variáveis não correlacionadas (scores da ACP) no procedimento de clusterização.

O artigo se divide em cinco seções. A presente introdução aponta o contexto que se insere o estudo e o objetivo proposto. A Seção 2 apresenta a fundamentação teórica das técnicas

multivariadas utilizadas no estudo. A Seção 3 apresenta a metodologia empregada, enquanto a Seção 4 descreve o conjunto de observações utilizadas na pesquisa e discorre sobre os resultados obtidos. A Seção 5 conclui o estudo e sugere possíveis desdobramentos desse trabalho.

2.2 Referencial teórico

2.2.1 Análise de agrupamentos

A análise de clusterização é método estatístico para alocar observações em grupos (*clusters*), de forma que as observações inseridas em um mesmo grupo sejam similares entre si e diferentes das inseridas em outro. Os agrupamentos gerados devem exibir elevada homogeneidade interna e externa; se representados graficamente, os objetos dentro dos *clusters* deverão estar próximos e agrupamentos distintos estarão distantes. Em termos práticos, a clusterização subsidia a interpretação de grandes quantidades de observações, que podem não apresentar significado evidente quando dispersas, com a descrição concisa das observações através da construção de agrupamentos (HAIR et al., 2009).

Existem dois procedimentos para clusterização de observações: hierárquicos e não hierárquicos. Procedimentos hierárquicos geram agrupamentos através do dendrograma, não se fazendo necessário pré-estabelecer o número de *clusters*, e avaliam progressivamente a similaridade entre os grupos e observações. Por sua vez, procedimentos não hierárquicos inserem objetos em um único movimento com base nas distâncias entre as observações, pressupondo o conhecimento prévio do número de *clusters* para alocação das observações em grupos (HAIR et al., 2009). Jain e Dubes (1988) realçam que o impacto visual que o dendrograma gera é uma característica importante dos métodos de agrupamento hierárquico, possibilitando ao pesquisador ver como os objetos estão sendo mesclados em *clusters* ou divididos em níveis sucessivos de proximidade. O dendrograma é um tipo específico de diagrama em estrutura de árvore que consiste em camadas de nós, cada um representando um *cluster*, onde as linhas conectam os nós que representam os *clusters* que estão aninhados uns aos outros e o corte horizontal do dendrograma define o agrupamento.

Embora os resultados estejam diretamente relacionados à qualidade dos dados e ao número de agrupamentos gerados, percebe-se que, no geral, métodos não hierárquicos propõem melhor desempenho para grandes conjuntos de dados, visto que dispensam o

cálculo de matrizes de semelhança entre todas as observações, requisitando apenas o cálculo da semelhança entre cada observação e o centroide do *cluster*. Além disso, são considerados menos suscetíveis a dados atípicos, à medida de distância e à inclusão de variáveis irrelevantes (HAIR et al., 2009). No entanto, conforme Hair Jr. et al (2009), a combinação de ambos procedimentos se faz conveniente para aprimorar o desempenho da clusterização: usa-se o método hierárquico para definição do número adequado de *clusters* a serem gerados, e o método não hierárquico para efetiva realização dos agrupamentos. Tal sistemática foi adotada nessa pesquisa, utilizando o método *K-Means* para clusterização das observações (RENCHER, 2002).

Operacionalmente, o método *K-Means* agrupa as observações em *K clusters*, sendo esse *K* um valor previamente conhecido (NALDI et al., 2011; KAUFMAN; ROUSSEEUW, 2005). O agrupamento é feito calculando o centroide de cada grupo e atribuindo cada observação ao grupo com o centroide mais próximo. O método fornece um algoritmo que minimiza a soma das distâncias euclidianas entre as observações e o centroide mais próximo (TABOADA; COIT, 2007). Os *K-Means* apoia-se em: (i) selecionar uma partição inicial com *clusters K*, (ii) calcular os centroides para cada *cluster* e a distância euclidiana dos centroides para cada observação do conjunto de dados, (iii) agrupar as observações aos *clusters* cujos centroides estejam mais próximos e (iv) retornar ao passo (ii) até que a associação ao *cluster* se estabilize, ou seja, até quando não ocorrer variação significativa na distância mínima entre cada observação a cada um dos centroides dos K *clusters* (JAIN; DUBES, 1988; JAIN, 2010; NALDI et al., 2011).

A qualidade dos agrupamentos gerados através de ferramentas de clusterização pode ser avaliada através do *Silhouette Index* (SI), que afere a qualidade dos agrupamentos formados por qualquer técnica de clusterização. O SI mede o grau de similaridade de uma observação em relação às demais inseridas em seu próprio *cluster*, comparando-a com as observações alocadas em outros *clusters* (KAUFMAN; ROUSSEEUW, 2005). O SI é calculado através da equação (1), onde $a(n)$ é a distância média da n-ésima observação em relação às demais observações do grupo em que foi alocada e $b(n)$ é a distância média entre a n-ésima observação em relação às observações do grupo vizinho mais próximo.

$$SI_n = \frac{b(n)-a(n)}{max\{b(n),a(n)\}} \quad (1)$$

Cada observação apresenta um SI_n no intervalo [–1; 1], onde n representa a observação avaliada, sendo $n = 1, ..., N$. Quanto mais elevado o valor do índice SI, melhor a alocação

do objeto em seu *cluster* de destino. Dessa forma, valores próximos a –1 apontam que a observação foi equivocadamente inserida no grupo; valores intermediários, próximos a zero, denotam observações que a observação poderia estar tanto em seu grupo atual quanto em algum outro grupo; e valores de SI próximos a 1 revelam que a distância, ou dissimilaridade, entre a observação e demais observações inseridas em outros grupos é pequena, considerando-se, portanto, que a observação foi adequadamente alocada em seu *cluster* de destino (ROUSSEEUW, 1987).

2.2.2 Análise de componentes principais

O objetivo central ACP é reduzir a dimensionalidade de um conjunto de observações que consiste em um grande número de variáveis inter-relacionadas, mantendo o máximo possível a variabilidade do sistema. Isso é alcançado por meio da conversão em um novo conjunto de variáveis, os componentes principais (CPs), que não são correlacionados e que são ordenados de forma que os primeiros retenham a maior parte da variação presente em todas as variáveis originais (JOLLIFFE, 2002).

Conforme Rencher (2002), o primeiro componente principal é a combinação linear com variância máxima – onde se busca uma dimensão ao longo da qual as observações sejam separadas ou dispersas ao máximo; já o segundo componente principal é a combinação linear com variância máxima em uma direção ortogonal ao primeiro componente principal e assim por diante.

A operacionalização da ACP ocorre por meio do cálculo dos autovalores e dos autovetores da matriz de covariância dos dados originais (JOLLIFFE, 2002). Considere *x* um vetor de *P* variáveis. O primeiro componente principal é definido como $\alpha_1^T x$ tal que os elementos em *x* apresentem máxima variância, onde $\alpha_1^T = [\alpha_{11}\, \alpha_{12} \ldots \alpha_{1p}]$ são atribuídos como pesos. O segundo componente é estabelecido como $\alpha_2^T x$, não correlacionado com $\alpha_1^T x$, e com os elementos de *x* tendo a máxima variância possível. Os vetores α_j são autovetores da matriz Σ, considerada como a matriz de variâncias e covariâncias de *x*. Finalmente, impõe-se à formulação de maximização de variância entre os componentes a restrição $\alpha_j^T \alpha_j = 1$, impondo o comprimento unitário nos autovetores. Nessa notação, cada autovetor α_j está relacionado com λ_j, o j-ésimo maior autovalor da matriz Σ. O problema se resume a maximizar a variância de $\alpha_1^T x = \alpha_1^T \sum \alpha_1$, sujeito à restrição $\alpha_1^T \alpha_1 = 1$ (CERVO; ANZANELLO, 2015).

Por conta de sua reconhecida habilidade na análise de dados, a ACP encontra aplicabilidade em dados de incêndios e outras naturezas de acidentes. Miguel et al. (2010) utilizaram a ACP para avaliar a influência de um período de treinamento no comportamento do trabalhador e, consequentemente, no tempo de evacuação, em caso de emergência. Hastie e Searle (2016), por ACP e regressão de mínimos quadrados ordinários, desenvolveram um modelo para explicar a variação nas taxas de incêndio acidental usando variáveis socioeconômicas. Conedera et al. (2018) utilizaram a análise de redundância (RDA), descrita como uma ACP em conjunto da análise de agrupamento hierárquico baseado no índice de dissimilaridade de Bray-Curtis, visando a identificar as características climáticas, ambientais e socioeconômicas dos *clusters* gerados para os regimes de incêndio nas Região dos Alpes. Caleffi et al. (2017) apresentaram uma abordagem multivariada para seleção de variáveis aplicando em conjunto a Distância de Bhattacharyya (BD) e a ACP a um conjunto de observações composto de variáveis que potencialmente ajudam a explicar a ocorrência de conflitos de tráfego, usando também um modelo de Análise Linear Discriminante (LDA) com as variáveis selecionadas para estimar a ocorrência de conflitos. Rundmo e Hale (2003) valeram-se da ACP para identificar as dimensões das atitudes de segurança e de prevenção de acidentes entre gestores de uma empresa. Dimitriou et al. (2018), para demonstrar uma abordagem de investigação da magnitude das inconsistências de dados estatísticos de segurança rodoviária, utilizaram em conjunto a ACP, a análise de clusterização hierárquico e a Modelagem de Equações Estruturais (SEM).

2.2.3 Técnicas multivariadas aplicadas a dados descritivos de incêndio

A literatura traz vasto número de aplicações de ferramentas de clusterização em dados associados a eventos de incêndios. Higgins et al. (2013) utilizaram *K-Means* para identificar os perfis de cidades inglesas por meio da coleta de informações demográficas e de estilo de vida dos indivíduos; a pesquisa integrou técnicas estatísticas e um modelo espacial para compreender o risco de incêndio e identificar o índice de vulnerabilidade dos indivíduos a incêndios. Higgins et al. (2014) também se valeram de *K-Means* para identificar as características socioeconômicas e níveis de risco de incêndio dos segmentos populacionais dentro da região estudada visando a examinar a evolução de um sistema de informação geográfica para apoio à prevenção de incêndios.

Guldaker e Hallin (2014), para identificar as causas da distribuição espacial e espaço-temporal de incêndios intencionais e analisar como as diferentes condições de vida podem determinar a frequência dessas ocorrências, aplicaram o método de densidade de *Kernel* (*Kernel Density Estimation*) agregado à análise de clusterização para revelar *clusters* entre as subáreas geradas. Tung e Kim (2011) propuseram um algoritmo para detecção de fumaça usando imagens de vídeo, aplicando a técnica de clusterização *Fuzzy C-Means* para extrair parâmetros de características da fumaça que foram utilizados como vetores para máquina de suporte vetorial (*Support Vector Machine - SVM*). Na mesma linha, Alamgir et al. (2018) desenvolvem um algoritmo de detecção de fumaça baseado em vídeo combinando Padrões de Co-ocorrência Binária Local com a análise de *clusters Fuzzy C-Means* e SVM para explorar as características da textura da fumaça. Palamara et al. (2011) desenvolveram uma abordagem utilizando em conjunto os algoritmos Mapa Auto-Organizacional de Kohonen (*Kohonen's Self-Organizing Map* - SOM) e clusterização *K-Means* a fim de agrupar os acidentes ocupacionais em diferentes classes, comparando os resultados da sistemática proposta com a clusterização hierárquica. Por fim, Chen et al. (2017) aplicaram *K-Means* para investigar as características do fogo florestal e seus potenciais impulsionadores.

Lizhi e Aizhu (2008) desenvolveram um método para classificação automática do risco de incêndio urbano o qual apoia-se em número de ocorrências de incêndio, perdas econômicas diretas e baixas de incêndio - ponderados conforme sua importância para avaliação do risco. O estudo baseou-se nas estatísticas de incêndio em Shenyang, na China, no período de 2000 a 2004, realizando a análise de agrupamento em relação à estatística de incêndio com o método de agrupamento *K-Means*.

Orozco et al. (2012) agregaram os incêndios florestais registrados em Cantão Ticino (Suíça) de 1969 a 2008, em três sub-períodos (*clusters*), conforme as disposições legais preventivas à época, e em duas causas de ignição. A metodologia usou permutação estatística de varredura, no espaço e do tempo, e sistema georreferenciado para visualização dos dados e resultados; testes de hipóteses de Monte Carlo foram aplicados para avaliar a significância estatística de cada *cluster*.

A relação de incêndios e de roubos em residência, de 2004 a 2006, em Surrey (Canadá), serviu para a aplicação de ferramentas de mapeamento de *clusters*, identificando as localidades estatisticamente significativas. Foram analisados os padrões espaciais e temporais de incêndios e de eventos de roubo a residências. Percebeu-se que incêndios,

como crimes, não são uniformemente distribuídos no tempo e no espaço, ocorrendo de forma previsível e não aleatória (WUSCHKE, K. et al., 2013).

Análises de risco foram realizadas nos incêndios ocorridos em edifícios residenciais na China de 2007 a 2010. As características estruturais e de ocupação, os sistemas de prevenção e proteção contra o fogo e os perigos de incêndio presentes nas edificações constituíram os cenários; o número de mortes e as perdas financeiras compuseram os índices de risco. Os agrupamentos de cenários de incêndio especificaram elementos relativos ao fogo, como desenvolvimento e propagação. O *cluster* de cenário de supressão automática de fogo descreveu o processo de extinção ou de controle desse. O *cluster* de cenários comportamentais descreveu a reação dos ocupantes em resposta ao início do fogo e a intervenção dos serviços de combate e resgate (XIN; HUANG, 2013).

2.3 Método

O método proposto é dividido em cinco etapas: (i) geração do dendrograma para definição do número de clusters a serem formados, (ii) clusterização via *K-Means* utilizando as variáveis originais que descrevem os eventos de incêndio, (iii) avaliação da qualidade dos agrupamentos gerados através do *Silhouette Index*, (iv) interpretação dos agrupamentos formados, e (v) repetição das etapas (i) a (iv) para dados transformados pela ACP.

Etapa 1: Geração do dendrograma para definição do número de clusters a serem formados

Consideram-se as x variáveis originais do conjunto de observações. Aplica-se a elas o método hierárquico de dendrograma para definir o número K de *clusters* recomendados. O dendrograma apresenta visualmente as características das observações avaliadas, representando, por meio do gráfico tipo árvore, a similaridade entre os grupos e observações.

Etapa 2: Clusterização via K-Means utilizando as variáveis originais que descrevem os eventos de incêndio

Conhecendo o número adequado de *clusters* de acordo com o dendrograma resultante na etapa anterior, aplica-se a ferramenta não hierárquica *K-Means* às variáveis originais para agrupar as observações em *K clusters*.

Etapa 3: Avaliação da qualidade dos agrupamentos gerados através do Silhouette Index

A qualidade dos grupos gerados é avaliada através da média do SI, fazendo uso da distância Euclidiana para medir o grau de similaridade de uma observação em relação às demais alocadas no mesmo *cluster* e comparando-a com as observações inseridas em outros clusters.

Etapa 4: Interpretação dos agrupamentos formados

A quarta etapa consiste em interpretar qualitativamente os resultados da clusterização, identificando particularidades que definem cada grupo, apontando os fatores mais relevantes, similaridades e dissimilaridades entre as observações de cada grupo. Para tanto, essa etapa apoia-se na literatura específica da área, buscando compreender os padrões de comportamento das ocorrências de incêndio residencial dos bancos de dados estudados.

Etapa 5: Repetição das etapas 1 a 4 para dados transformados pela ACP

A fim de avaliar a influência que os níveis de correlação das variáveis exercem sobre a formação dos agrupamentos, converte-se o conjunto original de variáveis em um novo conjunto de variáveis latentes *t* (PCA scores). Passa-se então a utilizar tais variáveis nas 4 etapas descritas acima.

2.4 Estudo de caso

2.4.1 Área de estudo

O *National Fire Incident Reporting System* (NFIRS) é um relatório utilizado por bombeiros americanos para reportar atividades e eventos de incêndios. Anualmente, mais de 30.000 departamentos de bombeiros reportam-se ao NFIRS, relatando quase 1 milhão de incidentes de incêndio. Ressalta-se, no entanto, que por se tratar de um sistema de preenchimento voluntário, inclui apenas os incidentes de incêndio notificados pelos departamentos de bombeiros que se reportam ao NFIRS, não se constituindo em um relato de todas as ocorrências de incêndio. Assim, a NFIRS não é representativa da totalidade de incidentes de incêndio nos Estados Unidos, não servindo como um recenseamento de incidentes de incêndio ou acidentes (USFA, 2015; 2017).

Os dados utilizados nesse estudo são provenientes de dois dos onze relatórios disponibilizados pela NFIRS: o Módulo de Incêndio (NFIRS-2), que descreve cada evento de incêndio, e o Módulo de Incêndio de Estrutura (NFIRS-3), que especifica a estrutura que sofreu essa ocorrência. Os dados coletados referem-se aos incêndios

residenciais ocorridos nos estados da Flórida e do Texas no período de 2010 a 2014 associados a cinco causas: incêndio intencional, descuido ao cozinhar, mau funcionamento na rede elétrica, falhas em eletrodomésticos e incêndio a partir de chamas abertas. Optou-se por realizar o estudo sobre os dados desses estados devido à representatividade de seus registros. As cinco causas de incêndio foram selecionadas para a investigação por representarem grande parte da ocorrência de incêndio doméstico no contexto analisado.

As 14 variáveis que compõem os bancos de dados e explicam os eventos são: (i) *área de origem do fogo*, (ii) *fonte de calor*, (iii) *primeiro item a sofrer ignição*, (iv) *causa da ignição*, (v) *fator humano*, (vi) *propagação do fogo*, (vii) *condições de uso e ocupação da edificação*, (viii) *grau do dano causado pelo fogo ao pavimento*, (ix) *presença de sistema de detecção automática do fogo*, (x) *tipo de detector automático do fogo*, (xi) *sistema de alimentação do detector automático do fogo*, (xii) *operação do detector automático do fogo*, (xiii) *efetividade do detector automático do fogo* e (xiv) *presença de sistemas automáticos de extinção do fogo*.

Em termos da composição e pré-processamento dos bancos de dados analisados, foram selecionados exclusivamente os eventos de incêndio em edificações residenciais de um pavimento que ocorreram de forma isolada, envolvendo apenas essa edificação no incidente. As observações contam somente com as edificações identificadas como construção finalizada e fechada. Considera-se edificação fechada aquela que não cumpre os requisitos para ser tratada como aberta ou parcialmente aberta; edificações abertas são aquelas que cada parede contém, pelo menos, 80% da área aberta. Classificam-se parcialmente abertas as edificações em que a área total de aberturas em uma parede, que recebe pressão externa positiva, excede em mais de 10% a soma das áreas de aberturas no balanço do envoltório do edifício (paredes e telhado) e que a área total de aberturas em uma parede, que recebe uma pressão externa positiva, excede 0,37 m^2 ou 1% da área dessa parede, o que for menor, e a porcentagem de aberturas no balanço do envoltório do edifício não excede 20% (ASCE, 2010).

No tratamento das observações, também foram eliminadas aquelas identificadas como espúrias, ora porque continham registros indeterminados nas variáveis, ora porque os registros eram conflitantes entre as variáveis, restando então 782 observações para compor o banco de dados da Flórida e 1596 para o banco do Texas. Além disso, as

ocorrências registradas de forma incompleta ou inconsistente foram excluídas do conjunto de dados. Tashakkori e Teddlie (2002) e Sandelowski (2000; 2001) sugerem que temas qualitativos sejam representados numericamente em escores ou escalas para facilitar a interpretação do fenômeno. Dessa forma, as observações qualitativas da amostra foram quantificadas através no intervalo 0 a 1, pontuados através da lógica "maior-é-melhor", conforme Anexo A.

2.4.2 Resultados

2.4.2.1 Banco de dados da Flórida

As Figuras 2.1, 2.2 e 2.3 apresentam, respectivamente, os dendrogramas gerados pelo método hierárquico nas 14 variáveis originais e com 2 e 3 CPs obtidos pela ACP, respectivamente. Percebe-se, nos 3 casos, consistente sugestão para formação de 2 ou 3 *clusters*. Por se tratar de uma análise exploratória, optou-se pela formação de 3 agrupamentos (K=3).

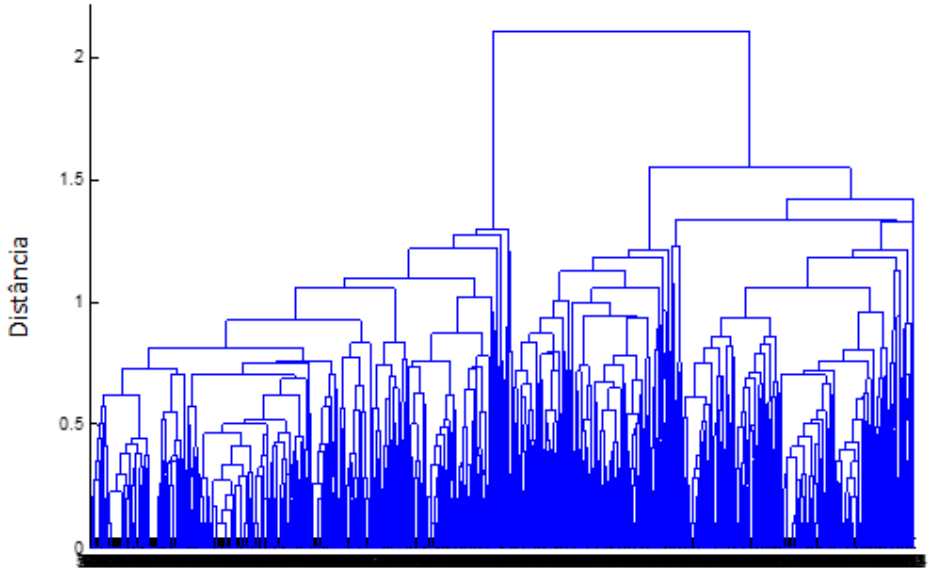

Figura 2.1*:* Dendrograma gerado a partir das 14 variáveis originais do banco de dados da Flórida

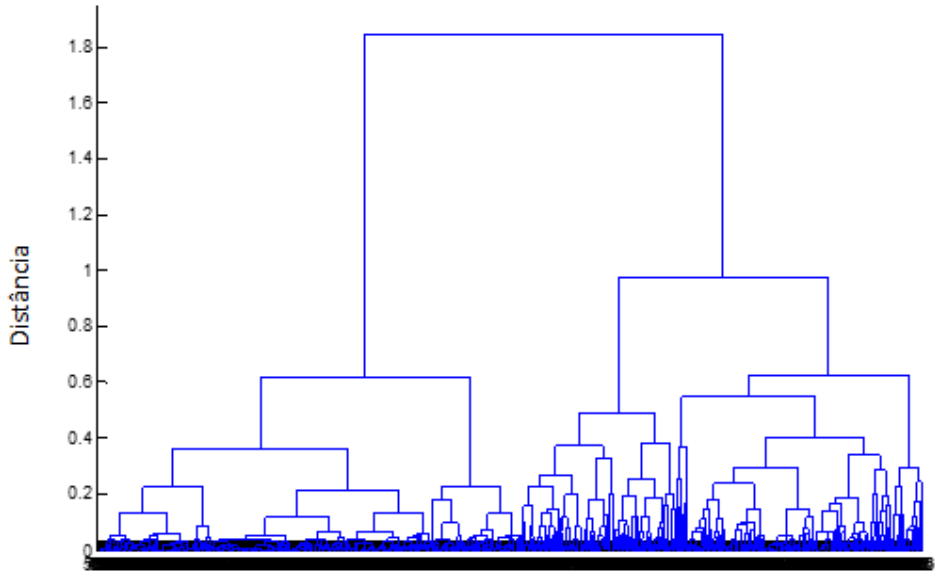

Figura 2.2: Dendrograma gerado a partir dos 2 CPs obtidos pela ACP do banco de dados da Flórida

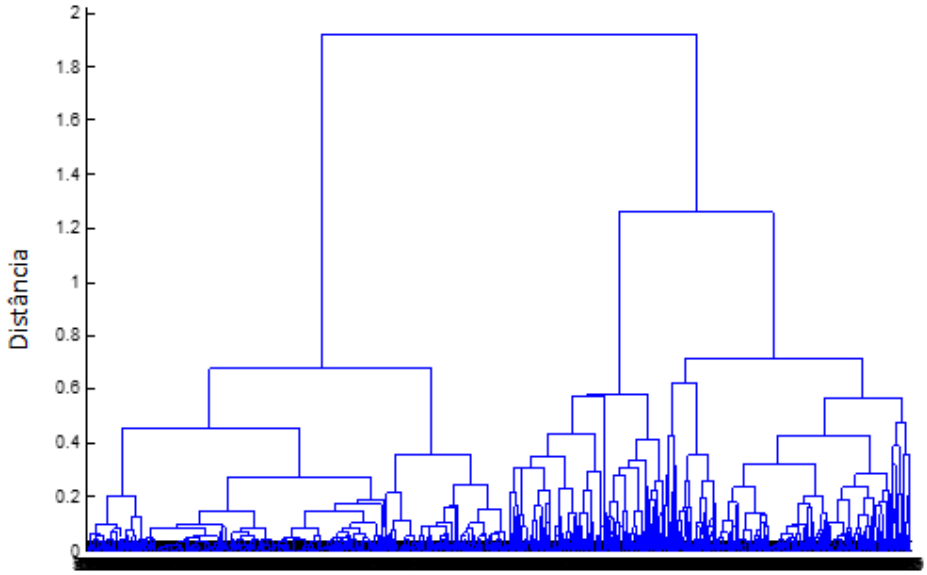

Figura 2.3: Dendrograma gerado a partir dos 3 CPs obtidos pela ACP do banco de dados da Flórida

Na sequência, os eventos descritos pelas 14 variáveis originais foram clusterizadas através do método *K-Means*, assim como os conjuntos de 2 e 3 CPs gerados pela ACP. A qualidade da clusterização foi avaliada através do SI médio, conforme apresentado na Tabela 2.1; o melhor valor foi 0,8481, obtido tanto para o agrupamento com as 14 variáveis originais como pela clusterização realizada com os 3 CPs. De tal forma, percebe-se que as correlações tipicamente verificadas nas variáveis originais não comprometem a qualidade da clusterização, tendo em vista que a qualidade do

procedimento é equivalente ao agrupamento gerado com base em 3 CPs, o que é amparado pelo fato da análise de clusters estar baseada em distâncias e não em correlações.

Tabela 2.1: Índices Silhouette médios obtidos para banco de dados da Flórida com a aplicação do método k-means com 14 variáveis originais e para 2 e 3 CPs obtidos pela ACP

	Clusterização utilizando método *K-means* – banco de dados Flórida		
		Com o uso da ACP	
	Com o uso das 14 variáveis originais	2 CPs	3 CPs
k	3	3	3
SI médio	0,8481	0,7660	0,8481

A análise dos agrupamentos formados pelas 14 variáveis e pelas 3 CPs mostrou formação de agrupamentos idênticos, sendo a proporção de eventos em cada cluster igual a 20%, 29% e 51%. O maior dos *clusters* gerados, com 51% das observações agrupadas, contém todos os incidentes de incêndio que não contava com sistema de detecção automática do fogo, ou seja, a variável *presença de sistema de detecção automática do fogo* é nula e, por conseguinte, as variáveis *tipo de detector automático do fogo*, *sistema de alimentação do detector automático do fogo*, *operação do detector automático do fogo* e *efetividade do detector automático do fogo* também não possuem valor nesse grupo. O *cluster* que abrange 29% dos eventos concentra eventos nos quais o sistema de detecção automática do fogo operou plenamente. O *cluster* com 20% das observações é composto pelas demais observações que possuem sistema de detecção automática do fogo. No entanto, percebe-se que, na maioria dos eventos, esse sistema não operou plenamente (97%) e, portanto, a variável *efetividade do detector automático do fogo* tenha sido registrada apenas para 3% dos incidentes.

2.4.2.2 Banco de dados do Texas

A Figura 2.4 apresenta o dendrograma gerado com aplicação do método sobre as 14 variáveis originais, enquanto que as Figuras 2.5 e 2.6 mostram os dendrogramas aplicados sobre sobre 2 CPs e 3 CPs. Novamente, entende-se que a formação de 3 clusters é satisfatória para evidenciar os diferentes padrões dos eventos de incêndio.

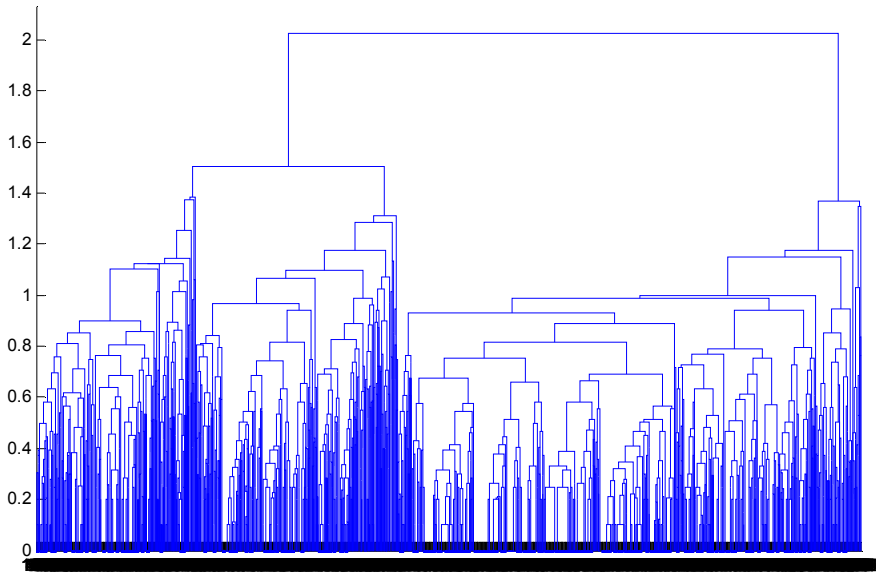

Figura 2.4: Dendrograma gerado a partir das 14 variáveis originais do banco de dados da Texas

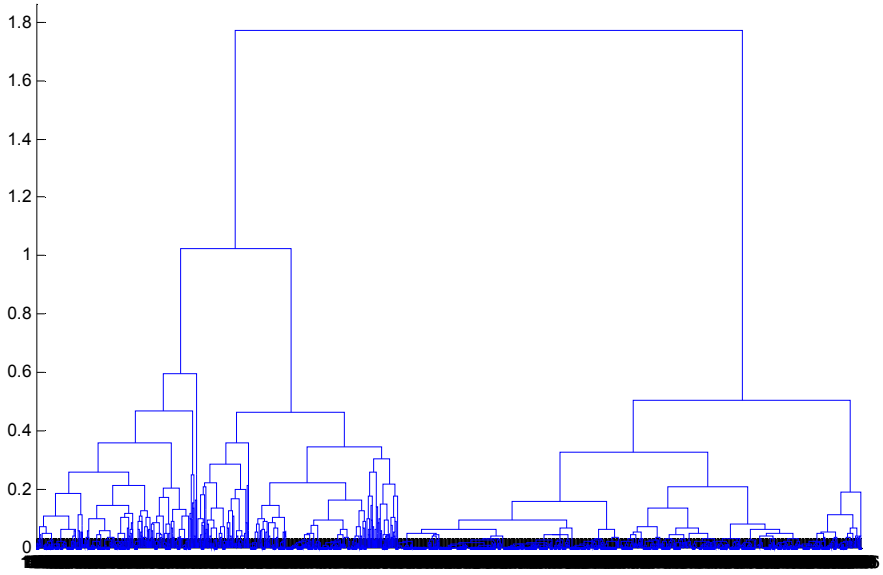

Figura 2.5: Dendrograma gerado a partir dos 2 CPs obtidos pela ACP do banco de dados da Texas

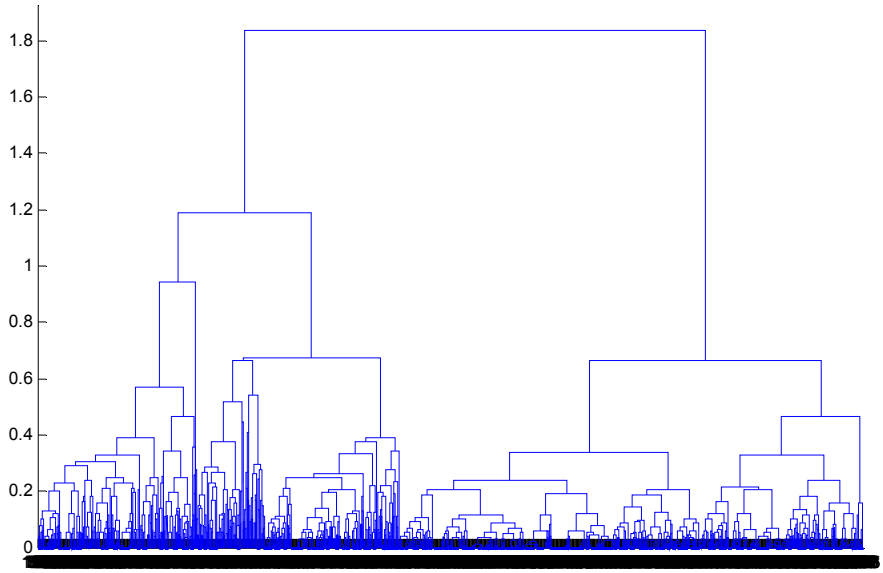

Figura 2.6: Dendrograma gerado a partir dos 3 CPs obtidos pela ACP do banco de dados da Texas

A Tabela 2.2 mostra os valores médios de SI obtido pela clusterização via *K-Means* aplicada ao conjunto de variáveis originais do banco de dados do Texas e aos conjuntos de 2 e 3 CPs gerados pela ACP nesse mesmo banco. O melhor valor de SI médio encontrado foi de 0,9141 para a clusterização aplicada aos conjuntos de 2 e 3 CPs obtidos pela ACP, sugerindo satisfatória estratificação das observações.

Tabela 2.2: Índices Silhouette médios obtidos para banco de dados do Texas com a aplicação do método k-means com 14 variáveis originais e para 2 e 3 CPs obtidos pela ACP

	Clusterização utilizando método *K-means* – banco de dados Texas		
	Com o uso das 14 variáveis originais	Com o uso da ACP	
		2 CPs	3 CPs
k	3	3	3
SI médio	0,7875	0,9141	0,9141

Os clusters gerados a partir de 2 e 3 CPs conduziram a alocações idênticas dos eventos nos clusters de destino. A proporção do agrupamento das observações para o banco de dados do Texas foi de 19%, 24% e 56%. O *cluster* com 19% dos incidentes é composto por observações que apresentavam sistema de detecção automática do fogo embora, em 98% dos casos, esse sistema não tenha operado plenamente e, por isso, a variável *efetividade do detector automático do fogo* foi registrada em apenas 2% das observações

(7 eventos). O *cluster* constituído por 24% dos eventos detém quase todas observações onde o detector automático do fogo operou plenamente; a exceção é encontrada em apenas uma única observação em que o detector não operou porque o fogo foi considerado pequeno para ativação. O maior dos *clusters*, com 56% das observações, traz as observações que não apresentam sistema de detecção automática do fogo, ou seja, a variável *presença de sistema de detecção automática do fogo* é nula e, em consequência, as variáveis *tipo de detector automático do fogo*, *sistema de alimentação do detector automático do fogo*, *operação do detector automático do fogo* e *efetividade do detector automático do fogo* também são nulas nesses registros.

2.5 Discussão dos resultados

A aplicação da abordagem proposta conduziu a resultados alinhados com as pesquisas e estudos relacionados ao gerenciamento de risco de incêndio trazidos pela literatura. Amplamente tratados em relatórios oficiais de incêndios, os sistemas de detecção do fogo fazem parte das exigências entre os sistemas de segurança contra o fogo nos EUA, sendo foco de diversas campanhas de prevenção contra incêndio e testados exaustivamente em estudos, tanto científicos (acerca de sua eficiência), como também comportamentais (acerca da conduta dos ocupantes da edificação em caso de incêndio). Os agrupamentos das observações foram gerados com relação às informações acerca do sistema de detecção automática do fogo: sua presença e sua efetividade de operação. Os agrupamentos resultantes em ambos bancos de dados encontraram padrões em torno das variáveis *presença de sistema de detecção automática do fogo* e *efetividade do detector automático do fogo* - se a detecção automática operou plenamente ou não.

Nos EUA, cada Estado tem seu próprio código de edificações, que na sua maioria adota os códigos da NFPA como base para a legislação local de segurança contra incêndio. A *National Fire Protection Association* (NFPA), uma organização global sem fins lucrativos, dedica-se à redução de mortes, ferimentos, propriedades e perdas econômicas decorrentes de eventos de incêndio. Entre suas diversas publicações, o NFPA 72, *National Fire Alarm and Signaling Code*, e o NFPA 101, *Life Safety Code*, exigem que residências novas e existentes de uma e duas famílias tenham detectores de fumaça dentro de todos os dormitórios e fora deles, estando presentes em todos os pavimentos. Além disso, exige-se que os detectores de incêndio estejam ligados diretamente à rede elétrica, bem como estarem interconectados e possuírem bateria auxiliar, de forma que todos soem simultaneamente alertando todos os ocupantes da edificação (AHRENS, 2015).

Os agrupamentos gerados em ambos conjuntos de dados apresentaram semelhança entre si, tanto pelas características das observações em relação às variáveis como pela proporção de distribuição das observações. Pode-se intitular os *clusters* gerados como: *cluster* de ocorrências de incêndio que não possuíam sistema de detecção automática do fogo, *cluster* de ocorrências de incêndio em que o sistema de detecção automática do fogo operou plenamente e *cluster* de ocorrências de incêndio em que a operação do sistema de detecção automática do fogo falhou.

Embora as observações coletadas para esse estudo sejam referentes aos anos de 2010 a 2014, a proporção dos *clusters* gerados diverge bastante das estatísticas registradas no período de 2009 a 2013, quando os detectores de fumaça - incluindo os eventos em que o fogo foi considerado muito pequeno para ativá-los, as situações em que operaram plenamente e os casos em que falharam-, estavam presentes em 73% dos incêndios domésticos registrados à NFIRS, tendo sido acionados em 53% dos incêndios domiciliares reportados. O relatório oficial da NFIRS também traz que, quando presentes em incêndios domésticos considerados grandes para ativá-los, os detectores de fumaça operaram em 87% das ocorrências (AHRENS, 2015).

Para a Flórida, as observações nos clusters gerados agruparam-se de forma que 51% das ocorrências de incêndio não possuíam sistema de detecção automática do fogo, 29% das ocorrências de incêndio contaram com o sistema de detecção automática do fogo operando plenamente e 20% das ocorrências de incêndio apresentaram falha na operação do sistema de detecção automática do fogo. Para o Texas, essa distribuição de agrupamento das observações foi, respectivamente, 56%, 24% e 19%. Uma das explicações para essa discrepância no tamanho dos agrupamentos com as estatísticas registradas é o fato que detectores de fumaça ligados diretamente à rede elétrica oferecem maior efetividade de operação quando comparados a alarmes alimentados por bateria. Isso se evidencia nas características do *cluster* de ocorrências de incêndio que apresentaram falha na operação do sistema de detecção automática do fogo: na Flórida, percebe-se que 27% das observações contavam com o sistema ligado à rede elétrica com ou sem bateria, enquanto 36% possuíam apenas alimentação à bateria; no Texas, 61% das ocorrências possuíam sistema de detecção alimentado somente por bateria e apenas 18% contavam com a interligação à rede elétrica, em conformidade às normas de segurança.

As pesquisas publicadas corroboram esse quadro. Em 2011, a *American Housing Survey* (AHS) realizou uma pesquisa onde três a cada cinco entrevistados relataram possuir

alarmes de fumaça alimentados apenas por baterias, enquanto um terço disse que seus alarmes eram alimentados por eletricidade e baterias e 7% tinham alarmes alimentados apenas por eletricidade. A pesquisa revelou também que, apesar das recomendações dos códigos da NFPA, 30% das residências com menos de cinco anos de idade que possuíam alarmes de fumaça contavam apenas com o sistema de alimentação por bateria. Outra pesquisa, realizada em 2010 pela *Harris Interactive* e NFPA, revelou que aproximadamente dois em cada cinco domicílios possuíam alarmes de fumaça em todos os dormitórios (como recomendado pelo NFPA 101), mas apenas um quarto das residências contavam com o sistema de detecção interconectado à rede elétrica.

Com relação ao *cluster* de ocorrências de incêndio que não possuíam de sistema de detecção automática do fogo, esperava-se que a variável *presença de sistema de detecção automática do fogo* fosse nula e, por conseguinte, as variáveis *tipo de detector automático do fogo*, *sistema de alimentação do detector automático do fogo*, *operação do detector automático do fogo* e *efetividade do detector automático do fogo* também não possuíssem valor nesse grupo, o que ocorreu em ambos conjuntos de dados analisados.

Para os *clusters* de ocorrências de incêndio em que o sistema de detecção automática do fogo operou plenamente e de ocorrências de incêndio em que a operação do sistema de detecção automática do fogo falhou, percebe-se que a maioria das observações contavam com sistemas de alarmes de fumaça, em comparação a sistemas de detecção de calor ou sistemas combinados de detecção de fumaça e calor. Esse comportamento vai de encontro aos números consolidados por Ahrens (2015) de 2009 a 2013, quando 88% dos dispositivos de detecção automático do fogo nos incêndios domiciliares relatados foram projetados para detectar apenas a fumaça.

As demais variáveis que compõem os bancos analisados, como *área de origem do fogo*, *fonte de calor*, *primeiro item a sofrer ignição*, *causa da ignição* e outras, embora expressem características das observações, têm seus valores extremamente pulverizados nos *clusters*, não apresentando relevância que auxilie na caracterização dos agrupamentos gerados. Esse comportamento era esperado para algumas variáveis, a exemplo da variável *presença de sistemas automáticos de extinção do fogo*: no período de 2010 a 2014, apenas 8% dos incêndios residenciais relatados à NFIRS contavam com a presença de *sprinklers*. Além disso, os Estados da Flórida e do Texas não possuem regulamentação estadual que exija o uso de *sprinklers* em ambientes residências, embora a Flórida permita que

jurisdições locais prescrevam *sprinklers*, enquanto o Texas não autoriza legislações locais a disporem sobre esse assunto (AHRENS, 2017; NFPA, 2017).

2.6 Conclusão

A análise de *clusters* tem sido reiteradamente aplicada em diversas pesquisas que buscam a melhor compreensão das tendências dos incidentes de incêndio, fundamentando a escolha desse artigo em agrupar as observações para identificar as particularidades predominantes que definem cada grupo.

A abordagem proposta associou técnicas hierárquica e não-hierárquica de clusterização, aplicando o método *K-Means* para aprimoramento das técnicas hierárquicas e a ACP para melhorar a qualidade da análise de clusterização, avaliada do através do SI. As etapas dessa pesquisa compreenderam em (i) geração do dendrograma para definição do número de clusters a serem formados, (ii) clusterização via *K-Means* utilizando as variáveis originais que descrevem os eventos de incêndio, (iii) avaliação da qualidade dos agrupamentos gerados através do *Silhouette Index*, (iv) interpretação dos agrupamentos formados; e (v) repetição das etapas (i) a (iv) para variáveis transformadas pela ACP. A utilização dos CPs como parâmetros de entrada da clusterização (substituindo as variáveis originais) objetiva avaliar um eventual aprimoramento da qualidade dos grupos gerados com base na utilização de variáveis não correlacionadas (caso dos CPs).

A abordagem foi aplicada a dois conjuntos de eventos de incêndios descritos pelas mesmas 14 variáveis. Para a Flórida, o melhor valor de SI médio foi de 0,8481, obtido tanto a partir do agrupamento gerado pelo *K-Means* com as 14 variáveis originais como pela clusterização realizada com os 3 CPs. Para o Texas, o SI médio encontrado foi de 0,914, resultante da clusterização aplicada aos conjuntos de 2 e 3 CPs obtidos pela ACP. Os dendrogramas de ambos os bancos de dados recomendaram o uso de 3 *clusters* para a aplicação do método *K-Means*. A distribuição das observações nesses *clusters* foi similar para Flórida e Texas, sendo os 3 grupos rotulados como (i) ocorrências de incêndio que não possuíam de sistema de detecção automática do fogo, (ii) ocorrências de incêndio em que o sistema de detecção automática do fogo operou plenamente, e (iii) ocorrências de incêndio em que a operação do sistema de detecção automática do fogo falhou. Os agrupamentos gerados pela aplicação do método atendem a expectativa, haja vista a vasta recorrência do assunto na literatura especializada e nos relatórios oficiais acerca de incêndio.

Para trabalhos futuros, sugere-se o emprego de diferentes ferramentas multivariadas para a formação dos agrupamentos, comparando os resultados obtidos com a metodologia proposta por esse artigo. Recomenda-se ainda o uso de outras métricas para aferir a qualidade dos clusters gerados.

2.7 Referências

AHRENS, M. Smoke Alarms in U.S. Home Fires. **NFPA Research, Data and Analytics Division**, set. 2015.

AHRENS, M. U.S. Experience with Sprinklers. **NFPA Research, Data and Analytics Division**, jul. 2017.

ALAMGIR, N. Combining multi-channel color space with local binary Co-occurrence feature descriptors for accurate smoke detection from surveillance videos. **Fire Safety Journal**, disponível on line em 24. set. 2018.

AMERICAN SOCIETY OF CIVIL ENGINEERS. **Minimum Design Loads for Buildings and Other Structures**, ASCE/SEI Standard 7-10, 2010.

CALEFFI, F.; ANZANELLO, M. J.; CYBIS, H. B. B. A multivariate-based conflict prediction model for a Brazilian freeway. **Accident Analysis and Prevention**, v. 98, p. 295-302, jan. 2017.

CERVO, V. L.; ANZANELLO, M. J. Seleção de variáveis para clusterização de bateladas produtivas através de ACP e remapeamento kernel. **Production**, v. 25, n. 4, p. 826-833, out./dez. 2015.

CEYHAN, E.; ERTUGAY, K.; DÜZGÜN, S. Exploratory and inferential methods for spatio-temporal analysis of residential fire clustering in urban areas. **Fire Safety Journal**, v. 58, p. 226-239, maio 2013.

CLARE, J.; GARIS, L.; PLECAS, D.; JENNINGS, C. Reduced frequency and severity of residential fires following delivery of fire prevention education by on-duty fire fighters: Cluster randomized controlled study. **Journal of Safety Research**, v. 43, p. 123-128, abr. 2012.

CHEN, D.; PEREIRA, J. M. C.; MASIEIRO, A.; PIROTTI, F. et al. Mapping fire regimes in China using MODIS active fire and burned area data**. Applied Geography**, v. 85, p. 14-26, ago. 2017.

CONEDERA, M.; KREBS, P.; VALESE, E.; COCCA, G.; SCHUNKD, C.; MENZEL, A.; VACIK, H.; CANEGH, D.; JAPELJ, A.; MURI, B.; RICOTTA, C.; OLIVERI, S.; PEZZATTI, G. B. Characterizing Alpine pyrogeography from fire statistics. **Applied Geography**, v. 98, p. 87-99, set. 2018.

COOK JR, D. E. The Great Fire of Pittsburgh in 1845 or How a Great American City Turned Disaster into Victory, **The Western Pennsylvania Historical**, v. 51, n. 2, p. 127-153, abril 1968.

DIMITRIOU, L.; NIKOLAOU, P.; ANTONIOU, C. Exploring the temporal stability of global road safety statistics. **Accident Analysis and Prevention**, disponível on line em 21 fev. 2018.

GULDAKER, N.; HALLIN, P. O. Spatio-temporal patterns of intentional fires, social stress and socio-economic determinants: A case study of Malmö, Sweden. **Fire Safety Journal**, v. 70, p. 71-80, nov. 2014.

HAIR JR., J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. Recurso eletrônico. Tradução Adonai Schlup Sant'Anna. 6. ed. Porto Alegre: Bookman, 2009.

HASTIE, C.; SEARLE, R. Socio-economic and demographic predictors of accidental

dwelling fire rates. **Fire Safety Journal**, v. 84, p. 50-56, ago. 2016.

HIGGINS, E.; TAYLOR, M.; FRANCIS, H.; JONES, M.; APPLETON, D. The evolution of geographical information systems for fire prevention support. **Fire Safety Journal**, v. 69, p. 117-125, out. 2014.

HIGGINS, E.; TAYLOR, M.; JONES, M.; LISBOA, P. J. G. Understanding community fire risk - A spatial model for targeting fire prevention activities. **Fire Safety Journal**, v. 62, parte A, p. 20-29, nov. 2013.

HORNBECK, R.; KENISTON, D. Creative Destruction: Barriers to Urban Growth and the Great Boston Fire of 1872, **American Economic Review**, v. 107, n. 6, p. 1365–1398, jun. 2017.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, ed. 8, p. 651-666, jun 2010.

JAIN, A. K; DUBES, R. C. **Algorithms for clustering data**. Englewood Cliffs: Prentice Hall; 1988.

JENNINGS, C. R. Social and economic characteristics as determinants of residential fire risk in urban neighborhoods: A review of the literature. **Fire Safety Journal**, v. 62, parte A, p.13–19, nov. 2013.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer-Verlag, 2002.

KAUFMAN, L.; ROUSSEEUW, P. **Finding groups in data: an introduction to cluster analysis**. New Jersey: Wiley Interscience; 2005.

LIZHI, W.; AIZHU, R. Urban Fire Risk Clustering Method Based on Fire Statistics. **Tsinghua science and technology**, v. 13, suplemento 1, p. 418-422, out. 2008.

MIGUEL, A. S.; GÓIS, J.; SILVA, J. Study on workers' evacuation in an industrial company. **Safety Science**, v. 48, p. 1050-1053, out. 2010.

MILLER, D. City of the Century: The Epic of Chicago and the Making of America. In: _____. **My Lost City, The Great Fire**. New York: Simon & Schuster, 1996. cap. 6. p. 143-157.

NALDI, M. C.; CAMPELLO, R.J.G.B.; HRUSCHKA, E., R.; CARVALHO, A.C.P.L.F. Efficiency issues of evolutionary k-means. **Applied Soft Computing**, v. 11, ed. 2, p. 1938-1952, mar. 2011.

NATIONAL FIRE PROTECTION ASSOCIATION. Fire Sprinkler Initiative. **Home fire sprinkler requirements at a glance**. Disponível em: <http://www.firesprinklerinitiative.org/legislation/sprinkler-requirements-by-state.aspx>. Acesso em 30 set. 2017.

OROZCO, C. V.; TONINI, M.; CONEDERA, M.; KANVESKI, M. Cluster recognition in spatial-temporal sequences: the case of forest fires. **Geoinformatica**, v. 16, ed. 4, p. 653–673, out. 2012.

PALAMARA, F.; PIGLIONE, F.; PICCININI, N. Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. **Safety Science**, v. 49, p. 1215-1230, out. 2011.

RENCHER, A. C. **Methods of Multivariate Analysis**. 2. ed. New York: John Wiley & Sons, 2002.

ROUSSEEUW, P. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53-65, nov. 1987.

RUNDMO, T.; HALE, A. R. Managers' attitudes towards safety and accident prevention. **Safety Science**, v. 41, p. 557-574, ago. 2003.

SANDELOWSKI, M. Combining Qualitative and Quantitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method Studies**. Research in Nursing & Health**, v. 223, p. 246–255, jun. 2000.

SANDELOWSKI, M. Real Qualitative Researchers Do Not Count: The Use of Numbers in Qualitative Research. **Research in Nursing & Health**, v. 24, p. 230–240, jul. 2001.

SCAWTHORN, C.; EIDINGER, J.; SCHIFF, A. J. The 1906 San Francisco, California Earthquake and Fires, **Fire Following Earthquake**. v. 11. Reston, VA: American Society of Civil Engineers, 2005.

TABOADA, H. A.; COIT, D. W. Data clustering of solutions for multiple objective system reliability optimization problems. **Quality Technology & Quantitative Management**, v. 4, ed. 2, p. 191–210, jan. 2007.

TASHAKKORI, A.; TEDDLIE, C. **Handbook of Mixed Methods in Social & Behavioral Research**. Thousand Oaks: Sage Publications, 2002.

TROITZSCH, J. H. Fires, statistics, ignition sources, and passive fire protection measures. **Journal of Fire Sciences**, v. 34 (3), p. 171-198, fev. 2016.

TUNG, T. X.; KIM, J. M. An effective four-stage smoke-detection algorithm using video images for early fire-alarm systems. **Fire Safety Journal**, v. 46, p. 276-282, jul. 2011.

UNITED STATES FIRE ADMINISTRATION. Federal Emergency Management Agency. **Fire in the United States, 2005-2014**. 18 ed., jan. 2017.

UNITED STATES FIRE ADMINISTRATION. National Fire Data Center. **National Fire Incident Reporting System 5.0** - Complete Reference Guide, 515 p., jan. 2015.

WUSCHKE, K.; CLARE, J.; GARIS, L. Temporal and geographic clustering of residential structure fires: A theoretical platform for targeted fire prevention. **Fire Safety Journal**, v. 62, parte A, p. 3-12, nov. 2013.

XIN, J.; HUANG, C. Fire risk analysis of residential buildings based on scenario clusters and its application in fire risk management. **Fire Safety Journal**, v. 62, parte A, p. 72-78, nov. 2013.

Anexo A: Quantificação das observações para transformar dados qualitativos em dados quantitativos

Variável	Quantificação da observação	
	Qualitativo	Quantitativo
área de origem do fogo	áreas estruturais	0,10
	áreas de ingresso	0,30
	áreas funcionais	0,50
	áreas de estocagem	0,80
fonte de calor	equipamentos em funcionamento	0,30
	objetos aquecidos	0,50
	fumo e assemelhados	0,70
primeiro item a sofrer ignição	componentes estruturais da edificação	0,10
	suplementos estocados	0,30
	mobiliário	0,40
	vestuário	0,50
	adorno, enfeite e material decorativo	0,60
	materiais gerais pesados (como fios elétricos, transformadores ou pneus)	0,70
	materiais gerais leves (como livros, revistas, papéis ou tecidos)	0,80
causa da ignição	incêndios intencionais	0,10
	causas sob investigação	0,30
	falhas do equipamento ou da fonte de energia	0,50
	causas naturais	0,80
	causas não intencionais	1,00
fator humano	possível incapacidade mental	0,10
	incapacidade física	0,20
	possivelmente sob efeito de álcool ou drogas	0,30
	idade foi um fator agravante	0,40
	adormecido	0,50
	desacompanhado ou sozinho	0,70
	múltiplas pessoas envolvidas no incidente	0,80
	não houve nenhum fator humano relacionado ao incêndio	1,00
propagação do fogo	chamas foram além da edificação de origem	0,00
	fogo permaneceu confinado na edificação de origem	0,25
	fogo permaneceu confinado no pavimento de origem	0,50
	fogo permaneceu confinado na dependência de origem	0,75
	chama não se propagou e permaneceu confinada no objeto que ocorreu a ignição	1,00
condições de uso e ocupação da edificação	em demolição	0,30
	desabitada e em condições inseguras (abandonada)	0,40
	em construção	0,50
	em reforma	0,60
	desabitada e segura	0,70
	desocupada (não utilizada frequentemente)	0,80
	em uso e ocupação normal	1,00
grau do dano causado pelo fogo ao pavimento	danos extremos	0,10
	danos severos	0,30
	danos moderados	0,60
	danos leves	0,90
presença de sistema de detecção automática do fogo	ausente	0,00
	presente	1,00
tipo de detector automático do fogo	detecção de calor	0,60
	combinação de detecção de calor e fumaça	0,90
	detecção de fumaça	1,00
sistema de alimentação do detector automático do fogo	somente a bateria	0,20
	ligado à tomada	0,40
	ligado direto à fiação elétrica	0,60
	ligado à tomada com bateria	0,80
	ligado direto à fiação elétrica com bateria	1,00
operação do detector automático do fogo	falhou	0,10
	não foi ativado devido à baixa magnitude do fogo	0,50
	operou	1,00
efetividade do detector automático do fogo	houve falha no alerta aos ocupantes	0,10
	não havia ocupantes na edificação durante o incêndio	0,50
	sistema alertou os ocupantes, mas ocupantes falharam na resposta	0,80
	sistema alertou os ocupantes e os ocupantes responderam	1,00
presença de sistemas automáticos de extinção do fogo	ausente	0,20
	parcialmente instalado	0,90
	presente	1,00

3 SEGUNDO ARTIGO

SISTEMÁTICA PARA SELEÇÃO DE VARIÁVEIS COM VISTAS À CLASSIFICAÇÃO DE EVENTOS DE INCÊNDIO DE ACORDO COM A SUA CAUSA

Resumo: *Embora representem menos de um terço das ocorrências anuais, incêndios residenciais respondem por aproximadamente oitenta por cento das mortes e das lesões em civis nos EUA. Estatísticas de incêndio têm sido analisadas para o entendimento do comportamento de tais eventos, a fim de auxiliar no direcionamento de intervenções e alocação de recursos de segurança. As pesquisas desenvolvidas na esfera de gerenciamento de incêndio utilizam-se, em sua maioria, de técnicas de agrupamento por clusterização (não supervisionadas) ao invés de sistemáticas supervisionadas para classificação de ocorrências em categorias. Este artigo propõe o emprego de ferramentas de mineração de dados para selecionar as variáveis mais relevantes para classificação das ocorrências de incêndio em classes. Por meio da combinação da técnica "omita uma variável de cada vez" com a ferramenta de classificação K-Nearest Neighbor (KNN), pretende-se apontar o melhor subconjunto de variáveis independentes (que descrevem os eventos de incêndio em termos da existência de sistemas automáticos de detecção e de extinção do fogo presentes nas residências, entre outras variáveis) para a predição da variável de resposta (classe de causa do incêndio). A sistemática proposta foi aplicada no conjunto de observações de incêndios residenciais reportados nos estados americanos da Flórida e do Texas, no período de 2010 a 2014, e associados a cinco principais causas (categorias). A acurácia média de classificação no banco reportando incêndios no Texas foi de 76,07% e de 65,94% na Flórida. Avaliou-se ainda o grau de relevância das variáveis para o modelo de classificação.*

Palavras-chave: Ferramentas de mineração de dados. Seleção de variáveis. Classificação. *K-Nearest Neighbor*. Incêndio.

3.1 Introdução

Anualmente, incêndios residenciais em solo americano implicam em perdas sócio-econômicas significativas, levando a centenas de mortes, milhares de feridos e milhões de dólares de perdas financeiras. Embora não seja possível eliminar completamente a ocorrência de incêndios, o gerenciamento através de atividades que envolvam análise sistemática, planejamento, tomada de decisão e atribuição dos recursos para gerenciar os riscos relacionados ao fogo pode reduzir essas perdas (CEYHAN et al., 2013).

O fogo, mais que fenômeno físico, possui aspetos sociais que transcendem o indivíduo, afetando a sociedade como um todo. A resposta aos incêndios que destruíram parte dos centros urbanos dos Estados Unidos, como Pittsburgh em 1845, Chicago em 1871, Boston

em 1872 e São Francisco em 1906 (ver Chen et al. (2015)), iniciou nas áreas de engenharia e administração pública com o desenvolvimento de técnicas de segurança que vão desde a concepção ao uso das edificações. Apesar desses esforços, a pesquisa sobre incêndio doméstico tem sido fragmentada; progressos importantes na redução das perdas por incêndio são vistos em propriedades não residenciais, visto à regulamentação governamental e às políticas mercadológicas em que estão inseridas (JENNINGS, 2013).

Percebe-se, contudo, que regulamentação e tecnologias, por si só, não atingem, na prática, progresso na prevenção de ocorrência de incêndios. Melhorias adicionais na segurança são vistas quando, associado a legislações e normas, tem-se o entendimento aprofundado da natureza social do problema. Diante disso, Spinardi et al. (2016) defendem que a segurança contra o fogo, afora aspectos técnicos, possui perspectivas influenciadas por fatores sociais e que uma compreensão desses vem a auxiliar os profissionais a utilizar de forma eficaz seu conhecimento. Desastres sociais decorrentes do fogo não devem, portanto, levar somente à construção de regulamentos, mas seu histórico deve orientar soluções de prevenção e serviços direcionados para combate e salvamento.

Madrzykowski e Kerber (2015) destacam a mudança da gravidade dos incêndios residenciais. Edificações modernas podem queimar oito vezes mais rápido devido à maior presença de elementos comburentes presentes nos móveis e materiais de construção atuais, além do uso de elementos construtivos leves que podem reduzir o tempo de colapso estrutural e de escape. No mobiliário, a substituição da madeira maciça por leve ou de algodão por materiais sintéticos pode aumentar a energia combustível em até quatro vezes, acelerando a combustão generalizada. Além disso, a nova configuração aberta entre cômodos facilita a propagação do fogo e difusão da fumaça, agravando o problema.

Atualmente, a compilação e análise de registros de incêndios tem sido visto como recurso promissor para o entendimento do comportamento do fogo e na geração de estratégias para seu combate. A determinação do método estatístico apropriado para tratamento dessas informações se dá com base nos tipos de dados, no objetivo da análise e na aplicação de requisitos (LIZHI; AIZHU, 2008). A edição mais recente do relatório estatístico *Fire in the United States* traz números alarmantes a respeito do fogo em residências nos EUA: entre 2005 a 2014, estima-se a ocorrência de 375.400 incêndios por ano, resultando em uma média anual de 2.620 mil civis mortos, 13.000 civis feridos e US$ 7,6 bilhões em danos à propriedade (UFSA, 2017).

No âmbito do gerenciamento de ações voltadas ao controle de incêndios, tem se tornado comum a utilização de técnicas estatísticas para análise das ocorrências de eventos, o que auxilia na priorização de recursos e no desenvolvimento de políticas de segurança. Nesse sentido**,** Corcoran et al. (2007), para examinar padrões para diferentes tipos de incidentes de incêndio, aplicaram técnicas multivariadas para explorar as relações com variáveis socioeconômicas. Asgary et al. (2010) e Ceyhan et al. (2013) realizaram análises temporais e espaciais de incidentes de incêndio, o primeiro objetivando ilustrar padrões horários, e o segundo compreender padrões de agrupamento de incêndio residencial. USFA (1998) estudou, por regressão múltipla, a relação entre fatores climáticos, demográficos e socioeconômicos e taxas de incêndio residencial, objetivando determinar a relevância das características de uma cidade para prever as taxas de incêndio doméstico atribuíveis a causas específicas. Hastie e Searle (2016), por análise de componentes principais e regressão de mínimos quadrados ordinários, desenvolveram um modelo para explicar a variação nas taxas de incêndio acidental usando variáveis socioeconômicas.

Com vistas a compreender as tendências das ocorrências de incêndio, técnicas de classificação também têm sido utilizadas na análise de dados, embora não tenha se encontrado um estudo que fizesse uso da técnica dos k vizinhos mais próximos (KNN – K-*Nearest Neighbor*) para classificação de observações, diferentemente da análise de *clusters*, que tem sido repetidamente empregada com esse fim. Também não se identificou a aplicação da sistemática "omita uma variável por vez" para seleção de variáveis em bancos relativos a eventos de incêndios. Nessa lacuna se posiciona esse artigo, propondo o uso de ferramentas de mineração de dados para identificar as variáveis mais informativas para classificação de eventos de incêndio em classes (classes essas entendidas como distintas causas do incêndio). Ademais, a escolha em categorizar os incidentes de incêndio conforme suas causas é fomentada pelo fato de que ações de melhoria podem ter maior eficácia quando aplicadas em um grupo de eventos com características similares do que em eventos isolados. Justifica-se ainda, para essa pesquisa, a seleção de variáveis, haja vista que um número elevado de variáveis utilizadas para caracterizar os incêndios tende a reduzir o desempenho dos modelos de classificação.

O presente artigo propõe uma abordagem apoiada em ferramentas de mineração de dados para selecionar as variáveis mais relevantes para classificação das observações (eventos de incêndio) em classes (causas específicas dos eventos). Através da combinação da técnica "omita uma variável de cada vez" com a ferramenta de classificação KNN,

identifica-se o melhor subconjunto de variáveis independentes (variáveis descritivas do tipo de residência, dentre outras) para a classificação do incidente de acordo com a sua causa. Em termos de seus passos, o método proposto consiste na (i) divisão do banco de dados em conjuntos de treino e teste, (ii) aplicação do KNN combinado com técnica "omita uma variável por vez" para seleção de variáveis, (iii) geração do gráfico de frequência de retenção de variáveis, e (iv) análise qualitativa do impacto das variáveis retidas sobre os argumentos gerados. As variáveis identificadas como mais relevantes são então qualitativamente discutidas.

O restante do artigo está organizado em quatro seções. A Seção 2 aprofunda nos princípios teóricos desses trabalhos realizados e dos incêndios residenciais, abordando a técnica de mineração de dados KNN e aspectos da seleção de variáveis. A Seção 3 apresenta a metodologia desenvolvida nessa pesquisa. A Seção 4 descreve os dados avaliados no estudo e expõe os resultados obtidos. A Seção 5 discute esses resultados com base no referencial teórico de segurança contra incêndio. A Seção 6 conclui a pesquisa e aponta oportunidades futuras de continuidade desse trabalho.

3.2 Referencial teórico

Essa seção traz os fundamentos teóricos da técnica de classificação KNN, além de apresentar abordagens de seleção de variáveis em contexto de análise de eventos de incêndio.

3.2.1 Classificação K-Nearest Neighbor

Entre as técnicas de mineração de dados para classificar observações, destaca-se o *K-Nearest Neighbor* (KNN); sua aplicação torna-se atraente por ser conceitualmente mais simples e intuitivo do que outras técnicas de classificação, requerendo apenas um parâmetro *K* que se refere ao número de vizinhos mais próximos para sua operacionalização (ANZANELLO et al., 2009; 2011; 2013; DUDA et al., 2001). Considere as observações em um conjunto de dados, sendo que parte das observações pertence à classe A e parte à classe B. O objetivo da técnica é classificar uma nova observação em A ou B com base nas variáveis independentes que descrevem tal observação. Para tanto, consideram-se os *K*-vizinhos mais próximos da nova observação, onde a proximidade é medida através da distância Euclidiana. Para cada um dos *K*-vizinhos, a classe A ou B é identificada através do seguinte curso de ação: a nova observação pertence à classe A se a maioria de seus *K*-vizinhos mais próximos estiver em

A. Se *K*=1, então a observação é simplesmente atribuída à classe de seu *K*-vizinho mais próximo. O número de vizinhos, *K*, que é um inteiro positivo tipicamente pequeno, é definido maximizando a acurácia da classificação no conjunto de dados onde a classe de cada observação é conhecida (ANZANELLO et al., 2009; 2011; 2013).

A determinação de alguns fatores pode afetar o desempenho do método KNN, como a escolha do valor *K*; se *K* for um valor muito pequeno, o resultado pode ser sensível a ruídos, porém, se for muito grande, pode incluir muitos pontos de outras classes. Outro ponto de atenção é a combinação dos rótulos de classificação: o método mais simples é tomar um voto majoritário, mas isso pode ser problemático caso os *K*-vizinhos mais próximos tenham sua distância variando amplamente e os *K*-vizinhos mais próximos mais confiáveis indicarem a classe do objeto. A escolha da medida de distância é outro aspecto: embora existam várias métricas, a medida de distância mais adequada é aquela em que uma menor distância entre dois objetos implica maior probabilidade de ter a mesma classe. Algumas métricas de distância podem ser afetadas pela alta dimensionalidade dos dados; o uso da medida de distância Euclidiana, portanto, consagrou-se por ser menos discriminante à medida que o número de atributos aumenta. Além disso, os atributos devem ser escalonados para evitar que um atributo de maior escala domine o cálculo da distância e, por conseguinte, a atribuição de rótulos de classe (WU et al., 2008).

Wan et al. (2012) corroboram que o KNN deve sua ampla aplicação ao baixo custo de implementação e ao elevado grau de eficácia. No entanto, salientam que a necessidade de determinar o valor adequado do parâmetro pode restringir seu uso na classificação de objetos. De acordo com Geng et al. (2008), a precisão do classificador KNN aumenta proporcionalmente com o aumento do *K*, contudo há um limite para esse incremento no valor de *K* não comprometer o desempenho do classificador KNN com o uso de muitos *K*-vizinhos.

3.2.2 Abordagens para seleção de variáveis

A seleção de variáveis tornou-se o foco de estudos que manipulam conjuntos de dados com elevado número de variáveis preditoras (independentes). Com os objetivos de otimizar o desempenho da predição, fornecer melhor custo-benefício e proporcionar melhor compreensão do processo que gerou os dados, a seleção de variáveis tende a facilitar a visualização dos dados, restringir os requisitos de medição e reduzir o tempo de coleta dos dados (GUYSON; ELISSEEFF, 2003).

Variáveis irrelevantes, de impacto exíguo sobre o modelo, podem deturpar a interação dos dados, enfraquecendo a exatidão da predição (GAUCHI; CHAGNON, 2001). A remoção de variáveis irrelevantes, ruidosas ou não confiáveis aprimora o desempenho do modelo, podendo reduzir sua complexidade; a melhoria das propriedades estatísticas também é argumento para realizar a seleção de variáveis, bem como minimização do risco de *overfitting* (ANDERSEN; BRO, 2010). Embora realizar todas as combinações de variáveis optando pelas melhores seja um modo de selecioná-las, tal sistemática pode ser computacionalmente infactível dependendo do número de variáveis em análise. Ainda que fosse possível, testar todas as combinações de variáveis pode elevar o risco de *overfitting*, salvo casos em que o número de amostras é muito superior ao número de combinações de variáveis (GAUCHI; CHAGNON, 2001; ANDERSEN; BRO, 2010).

Alguns aspectos são comuns para todos os tipos de seleção de variáveis sendo importante sua discussão. Problemas de *overfitting* podem ocorrer quando há mais variáveis do que amostras, resultando em predições deficitárias; de tal forma, torna-se imprescindível a validação da predição através do uso de um conjunto real de dados para teste, da validação de modelo cruzado ou de testes de permutação. Faz-se também fundamental a manipulação dos *outliers*, já que muitos métodos de seleção baseiam-se na avaliação de pequenas diferenças na qualidade do modelo, devendo, por isso, haver uma detecção desses valores atípicos (sendo aconselhado removê-los antes da seleção da variável). Embora não acarrete necessariamente uma melhora na predição, variáveis redundantes podem ser removidas quando se almeja um modelo mais enxuto, contudo tal remoção pode ser arriscada quando a interpretação do modelo resultante é o objetivo da análise. Sugere-se ainda, quando necessário, o uso de técnicas de pré-processamento como escalonamento, correção de dispersão multiplicativa, entre outros. Para um determinado conjunto de dados, pode haver vários métodos de seleção de variáveis com aplicabilidade semelhante, sendo útil comparar os resultados desses vários tipos (ANDERSEN; BRO, 2010).

Diversas pesquisas fazem uso de técnicas baseadas na integração de abordagens de mineração de dados, sendo o *K-Nearest Neighbor* (KNN) um dos métodos mais reconhecidos para seleção de variáveis devido sua fácil implementação, resultados satisfatórios e aplicação nos mais variados contextos. Chaovalitwongse et al. (2007) desenvolveram uma técnica de classificação para detecção de atividade cerebral anormal via KNN, enquanto que Anzanello et al. (2009), fazendo uso de índices de importância

das variáveis (IIV), testaram diferentes técnicas de classificação com vistas à categorização de bateladas produtivas. Anzanello et al. (2011) apresentaram um método baseado em mineração de dados para seleção de atributos em painéis sensoriais descritivos composto pela Análise de Componentes Principais (PCA) e KNN, em conjunto com a análise do Pareto Ótimo (PO), calculando os índices de importância de atributo com base nos pesos de PCA. Em consonância, Anzanello et al. (2013), no âmbito da averiguação da autenticidade de medicamentos, propuseram um novo método para selecionar os subconjuntos mais relevantes de número de comprimentos de ondas (variáveis), integrando PCA ao KNN para obtenção de um índice de importância variáveis.

De forma semelhante à geração de índices de importância, a técnica “omita uma variável por vez” tem sido amplamente empregada em diferentes cenários para seleção de variáveis. Essa abordagem omite uma variável por vez do procedimento de classificação, sendo a precisão avaliada em cada estágio. Uma vez que todas as variáveis forem omitidas, a variável responsável pela maior precisão é eliminada em definitivo, visto que contribuiu menos para a classificação da amostra. O procedimento é repetido para as variáveis remanescentes e a precisão é reavaliada sempre que cada variável for momentaneamente removida. O subconjunto de variáveis com maior precisão é o escolhido (ANZANELLO; FOGLIATTO, 2011; ANZANELLO et al., 2014; 2017). Kudo e Sklansky (2000), para medir o desempenho de um subconjunto de características, compararam metodologias para seleção de variáveis; o melhor subconjunto foi identificado através da combinação da “omita uma variável por vez” e KNN. De forma semelhante, Anzanello e Fogliatto (2011) combinaram o procedimento “omita uma variável por vez” com a técnica de agrupamento *k-mean*s para identificação das variáveis de *clustering* relevantes. Gunvig et al. (2013), para desenvolver um modelo matemático para prever o crescimento ou não crescimento de uma bactéria anaeróbica em produtos de carne pasteurizados embalados em atmosfera modificada, executaram uma validação cruzada para determinar o número ótimo de fatores PLS no modelo e um algoritmo "omita uma variável por vez" como uma forma de eliminar variáveis sem importância. Anzanello et al. (2014), para selecionar subconjuntos de número de ondas objetivando classificar amostras de medicamentos em dois grupos, autênticos ou fraudulentos, empregaram PCA aos dados gerando dois índices de importância de variável para orientar a eliminação de variáveis através da abordagem "omita uma variável por vez", aplicando técnicas de

agrupamento não-hierárquicas (*k-means* e *C-means Fuzzy*). Por fim, Anzanello et al. (2017), visando a identificar subconjuntos relevantes de comprimentos de onda para classificar amostras em duas categorias, propuseram um método dividido em duas etapas: a primeira com a classificação ocorrendo via KNN e a segunda com um refinamento com os intervalos que forneceram as maiores precisões de classificação através do uso de algoritmo genético.

3.2.3 Ferramentas multivariadas no contexto de incêndio

Corcoran et al. (2007) examinaram os padrões espaciais em diferentes tipos de incidentes de incêndio e suas relações com variáveis socioeconômicas usando estimativas de densidade *kernel* em dados de 2000 a 2004 de uma região do País de Gales. Seguindo a mesma linha, Asgary et al. (2010), com os dados de incêndios de 2000 a 2006 em Toronto (Canadá), investigaram a relação entre as localidades das ocorrências, sua variação no tempo (hora, semana e mês) e seus efeitos sobre a distribuição espacial das diferentes causas de incêndio. Essas análises mostraram que os eventos de incêndio seguem padrões espaciais e temporais e que tendências de incêndio diferem conforme a causa do incidente.

A modelagem preditiva dos padrões espaço-temporais de incêndios residenciais apresentada por Ceyhan et al. (2013) foi aplicada aos dados de Çankaya (Turquia) em 1998 e de 2005 a 2009. O estudo realizou o agrupamento de incêndio por parcelas de dispersão confrontando locais onde ocorreram incidentes com locais em que não ocorreram; para tanto, foram utilizados na análise gráficos de intensidade, análises do tipo espacial, temporal e espaço-temporal do agrupamento de incêndios através de testes de hipóteses de função-K espaciais e de Monte Carlo, buscando explicar as tendências sobre padrões de incêndio.

O estudo realizado pela USFA (1998) analisou 27 cidades americanas, considerando o número de incêndios atribuíveis às principais causas de incêndios domésticos. As taxas de incêndio referem-se ao período de 1993 a 1995 e foram relacionadas com as características climáticas, demográficas e socioeconômicas de cada localidade. A análise apoiou-se em coeficientes de correlação e regressão *stepwise*, identificando as variáveis independentes que melhor explicaram a variação em cada variável dependente. Uma análise fatorial também foi realizada incluindo todas as variáveis independentes significativamente relacionadas com as variáveis dependentes. As matrizes de correlação

apontaram que determinados indicadores climáticos, demográficos e socioeconômicos estavam relacionados a pelo menos uma das causas de incêndio investigada.

Por fim, uma análise aplicada aos dados dos incêndios de West Midlands (Inglaterra), entre os anos de 2010 e 2013, explorou a correlação entre as taxas de incêndio acidental e as características socioeconômicas. Usando a PCA, buscou-se compreender as relações entre as variáveis preditoras e minimizar a colinearidade entre tais variáveis. A análise de regressão por mínimos quadrados ordinários foi aplicada às variáveis fortemente relacionadas aos componentes identificados (HASTIE; SEARLE, 2016).

Na procura das pesquisas desenvolvidas no contexto de incêndio, não se encontrou um estudo que fizesse uso da metodologia *KNN* para classificação de observações; percebeu-se prevalência de ferramentas de classificação não supervisionadas, como clusterização. Do mesmo modo, não se identificou um trabalho que empregasse a sistemática "omita uma variável por vez" para seleção de variáveis relacionadas a bancos de dados de incêndios.

3.3 Método

O método sugerido é dividido em 4 etapas: (i) divisão do banco de dados descritivos dos incêndios em conjuntos de treino e teste, (ii) aplicação da ferramenta de mineração de dados *K*-vizinhos mais próximos combinada à técnica "omita uma variável por vez" para seleção de variáveis, (iii) geração do gráfico de frequência de retenção de variáveis, e (iv) análise qualitativa das variáveis retidas. Tais etapas são agora detalhadas.

Etapa 1: Divisão do banco de dados em conjuntos de treino e teste

Dividir aleatoriamente o conjunto com *N* observações em um conjunto de treino com N_{tr} observações e um conjunto de testes com N_{ts} observações, onde $N_{tr} + N_{ts} = N$. Nesse estudo, as amostras foram particionadas aleatoriamente em um conjunto de treino (consistindo em 75% das amostras) e um conjunto de teste (consistindo nos 25% restantes); a parcela de treino foi utilizada para selecionar as variáveis mais relevantes e a de teste representa novas amostras a serem classificadas usando as variáveis selecionadas. A fim de reduzir tendências em função do particionamento do banco de dados em treino e teste, Chaovalitwongse et al. (2007) sugerem, para a seleção do parâmetro *K,* o emprego da técnica de validação cruzada.

Etapa 2: Aplicação da ferramenta de mineração de dados K-vizinhos mais próximos combinada à técnica "omita uma variável por vez" para seleção de variáveis

A abordagem para seleção de variáveis visando à classificação das observações de incêndios integra a técnica de classificação *K*-vizinhos mais próximos à técnica "omita uma variável por vez" para inserção das observações em 5 classes: *incêndio intencional*, *descuido ao cozinhar*, *mau funcionamento na rede elétrica*, *falhas em eletrodomésticos* e *incêndio a partir de chamas abertas*. Tais causas de incêndio foram selecionadas para a investigação por representarem juntas 81% das ocorrências de incêndios domésticos nos estados da Flórida e do Texas de 2010 a 2014. Na técnica "omita uma variável por vez", cada variável é momentaneamente omitida do conjunto de treino com N_{tr} observações e uma classificação via KNN é conduzida sem aquela variável; a acurácia de classificação é avaliada. Na sequência, a variável é devolvida ao banco de dados e a próxima é omitida. Uma vez que todas as variáveis foram omitidas momentaneamente do conjunto de dados, a variável que conduziu à maior acurácia quando fora do procedimento de classificação deve ser removida permanentemente, já que a sua omissão resultou na classificação com melhor acurácia. Esse procedimento deve ser repetido iterativamente para as demais variáveis até que reste apenas uma, e a precisão reavaliada depois de cada uma ser omitida momentaneamente; o subconjunto de variáveis que resultar na maior precisão é o escolhido (ANZANELLO et al., 2017).

A fim de evitar tendências nos resultados por alguma divisão específica do conjunto de dados de treino e teste, replicou-se o processo iterativo de classificação e remoção de variáveis 500 vezes; os conjuntos de treino e teste foram aleatoriamente misturados para então serem divididos de forma distinta a cada repetição. A acurácia média de classificação oriunda das 500 replicações foi armazenada para avaliação da contribuição das variáveis ao modelo.

Etapa 3: Geração do gráfico de frequência de retenção de variáveis

Sugere-se então construir um gráfico relacionando percentual de variáveis retidas (abcissa) com a acurácia de classificação (ordenada), de forma a avaliar incrementos/reduções no desempenho de classificação com a remoção sistemática das variáveis menos relevantes. O subconjunto escolhido é aquele que conduz à máxima acurácia. Com vistas à identificação das variáveis que mais foram retidas nas 500 replicações, elabora-se um gráfico de barras, mostrando o percentual de retenção de cada

variável após a realização das 500 interações. Identificar o conjunto de variáveis retido com maior frequência, avaliado pelo percentual de ocorrências em que a variável foi retida no subconjunto de variáveis selecionadas

Etapa 4: Análise qualitativa das variáveis retidas

Por fim, interpretar a influência das variáveis retidas nos agrupamentos gerados, de forma a analisar qualitativamente o impacto dessas variáveis nas observações de incêndios nas classes *incêndio intencional*, *descuido ao cozinhar*, *mau funcionamento na rede elétrica*, *falhas em eletrodomésticos* e *incêndio a partir de chamas abertas.*

3.4 Estudo de caso

3.4.1 Bancos de dados analisados e pré-processamento dos dados

O *National Fire Incident Reporting System* (NFIRS) é um padrão de relatório utilizado por bombeiros americanos para reportar atividades de combate a incêndios. Participam desse sistema todos estados dos Estados Unidos e o Distrito de Columbia; anualmente, mais de 30.000 departamentos de bombeiros reportam ao NFIRS, somando quase 1 milhão de incidentes de incêndios (USFA, 2015; 2017). Os dados utilizados nesse estudo são provenientes de dois dos onze relatórios disponibilizados pela NFIRS: o Módulo de Incêndio (NFIRS-2), que descreve cada evento de incêndio, e o Módulo de Incêndio de Estrutura (NFIRS-3), que especifica a estrutura que sofreu essa ocorrência. Os dados coletados referem-se somente aos incêndios residenciais ocorridos nos Estados da Flórida e do Texas no período de 2010 a 2014 associados às seguintes causas: 1 - incêndio intencional, 2 - descuido ao cozinhar, 3 - mau funcionamento na rede elétrica, 4 - falhas em eletrodomésticos e 5 - incêndio a partir de chamas abertas. Optou-se por realizar o estudo sobre os dados desses Estados devido à representatividade de seus registros.

As 14 variáveis que compõem os bancos de dados e caracterizam o evento de incêndio são: (i) *área de origem do fogo*, (ii) *fonte de calor*, (iii) *primeiro item a sofrer ignição*, (iv) *causa da ignição*, (v) *fator humano*, (vi) *propagação do fogo*, (vii) *condições de uso e ocupação da edificação*, (viii) *grau do dano causado pelo fogo ao pavimento*, (ix) *presença de sistema de detecção automática do fogo*, (x) *tipo de detector automático do fogo*, (xi) *sistema de alimentação do detector automático do fogo*, (xii) *operação do detector automático do fogo*, (xiii) *efetividade do detector automático do fogo* e (xiv) *presença de sistemas automáticos de extinção do fogo*.

No tratamento das observações, foram eliminadas aquelas identificadas como espúrias por conta de registros indeterminados nas variáveis ou por registros conflitantes. Restaram 782 observações no banco de dados do Estado da Flórida, e 1596 do Texas. Além disso, as ocorrências registradas de forma incompleta ou inconsistente foram excluídas do conjunto de dados. Em termos do escalonamento das variáveis (para evitar a influência da magnitude das mesmas) Tashakkori e Teddlie (2002) e Sandelowski (2000; 2001) sugerem que temas qualitativos sejam representados numericamente em escores ou escalas para interpretar o fenômeno. Dessa forma, as variáveis de cunho qualitativo da amostra foram quantificadas através de escala, sendo representadas numericamente no intervalo 0 a 1, seguindo a razão do tipo "maior-é-melhor" (ver Anexo B).

3.4.2 Resultados

O método proposto para seleção das variáveis mais informativas foi aplicado aos dois bancos de dados descritos por 14 variáveis. A Figura 3.1 ilusta o perfil da acurácia de classificação à medida que as variáveis foram eliminadas em uma das 500 repetições realizadas (perfis distintos foram obtidos para cada replicação por conta da divisão treino-teste obtida). No contexto de todas as repetições, a acurácia média de classificação foi de 76,07% para dados de incêndios ocorridos no Texas; para a Flórida, obteve-se acurácia média de 65,94%. Para ambos os bancos de dados, não houve uma situação, durante as 500 replicações do método, que gerasse subconjuntos com 1, 2 ou 14 variáveis retidas. A Figura 3.2 mostra a distribuição do número de variáveis retidas pelo método para cada banco de dados. Para os dados do Texas, verificou-se maiores ocorrências de acurácia máxima nos subconjuntos que reteram 3 a 5 variáveis; para a Flórida, encontrou-se maior precisão nos subconjuntos de 7 a 9 variáveis.

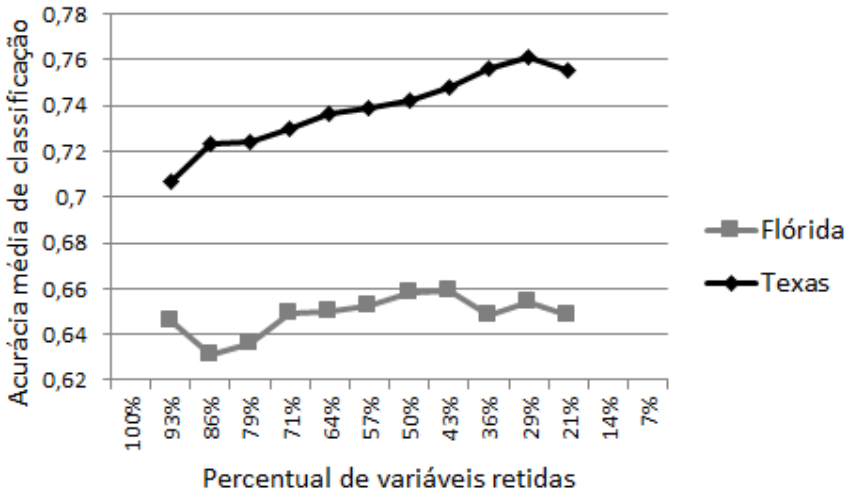

Figura 3.1: Perfil de acurácia x Percentual de variáveis retidas

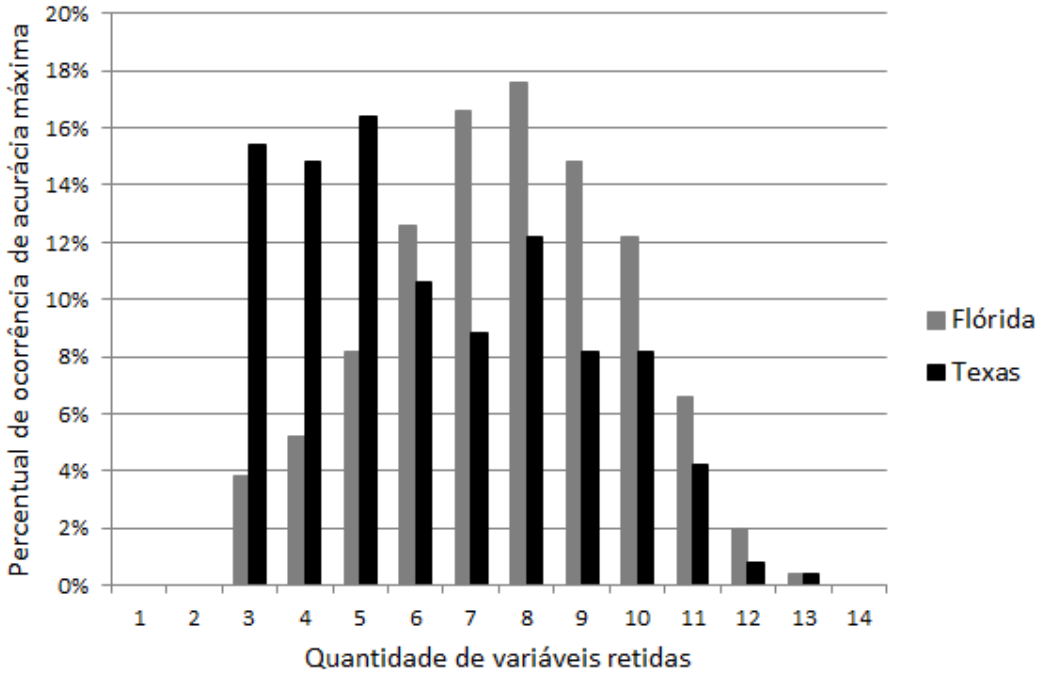

Figura 3.2: Percentual de ocorrência de acurácia máxima x Quantidade de variáveis retidas

As Figuras 3.3 e 3.4 demonstram o percentual de retenção de cada variável analisada nas 500 repetições para os dados da Flórida e Texas, respectivamente. O grau de relevância que as variáveis ofereceram ao modelo, mensurado pelo percentual de vezes em que a variável foi retida no subconjunto de variáveis selecionadas, apresentou valores semelhantes para os dois bancos de dados. A ordem de contribuição das variáveis ao modelo manteve-se praticamente idêntica para os dados da Flórida e Texas, uma vez que

as variáveis identificadas como de máxima importância são iguais para os dois bancos de dados, assim como as de média e baixa relevância.

Em ambas situações, a variável *fonte de calor* foi retida em todas as 500 replicações do procedimento classificatório, revelando-se imprescindível para compreensão do processo que deu origem ao incêndio. Igualmente relevante para a gestão de incêndio, a variável *causa da ignição* foi retida em 99% das repetições nos subconjuntos que apresentaram maior acurácia de classificação. A terceira variável tida como relevante é o *primeiro item a sofrer ignição*, que foi retida em 73% das classificações realizadas nos subconjuntos que geraram maior precisão.

No grupo de variáveis de média importância estão as variáveis *área de origem do fogo*, *presença de sistemas automáticos de extinção do fogo* e *tipo de detector automático do fogo* (embora as duas primeiras citadas tenham tido um comportamento diferenciado entre os bancos de dados estudados). Para a Flórida, *área de origem do fogo* e *presença de sistemas automáticos de extinção do fogo* apresentaram relevância superior se comparado ao Texas, tendo sido retidas em aproximadamente 70% das classificações; para o Texas esse valor diminui para aproximadamente 50% das ocorrências.

Em contraposição, a variável *propagação do fogo* não indica significância para a gestão de incêndio em ambos os bancos de dados, já que foi retida em apenas 14% dos subconjuntos para a Flórida e 17% para o Texas. O comportamento da variável *grau do dano causado pelo fogo ao pavimento* também é de baixa importância para análise, principalmente para o Texas, onde foi retida em 18% das verificações; para a Flórida esse valor é 31%. As demais variáveis estudadas (*fator humano*, *condições de uso e ocupação da edificação, presença de sistema de detecção automática do fogo*, *sistema de alimentação do detector automático do fogo*, *operação do detector automático do fogo* e *efetividade do detector automático do fogo*) também compõem esse grupo de baixa relevância em termos de classificação dos incêndios de acordo com sua classe causadora.

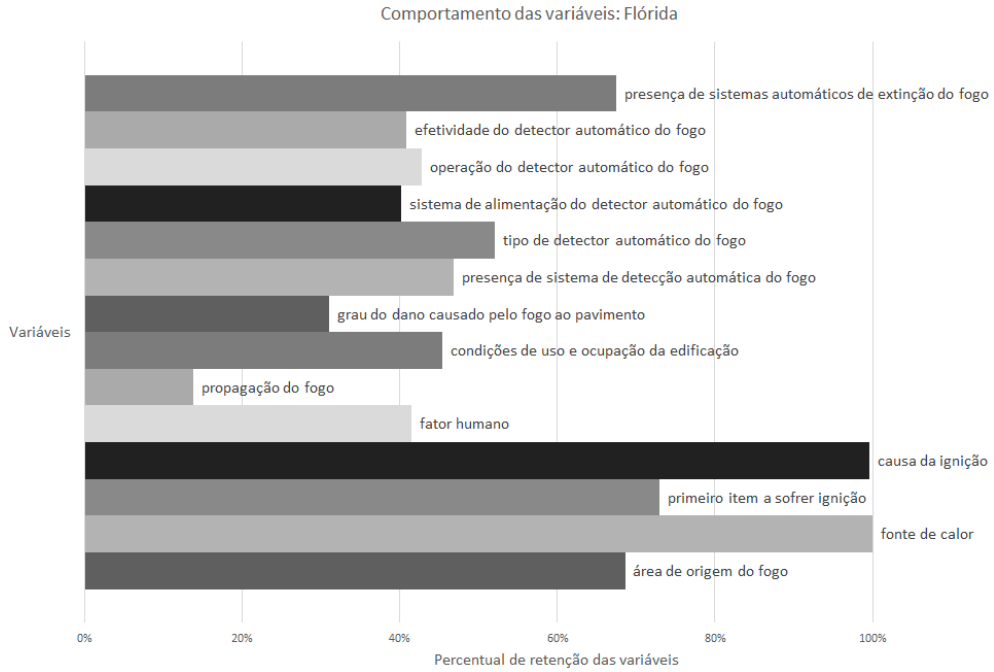

Figura 3.3: Frequência de retenção das variáveis para classificação dos incêndios na Flórida

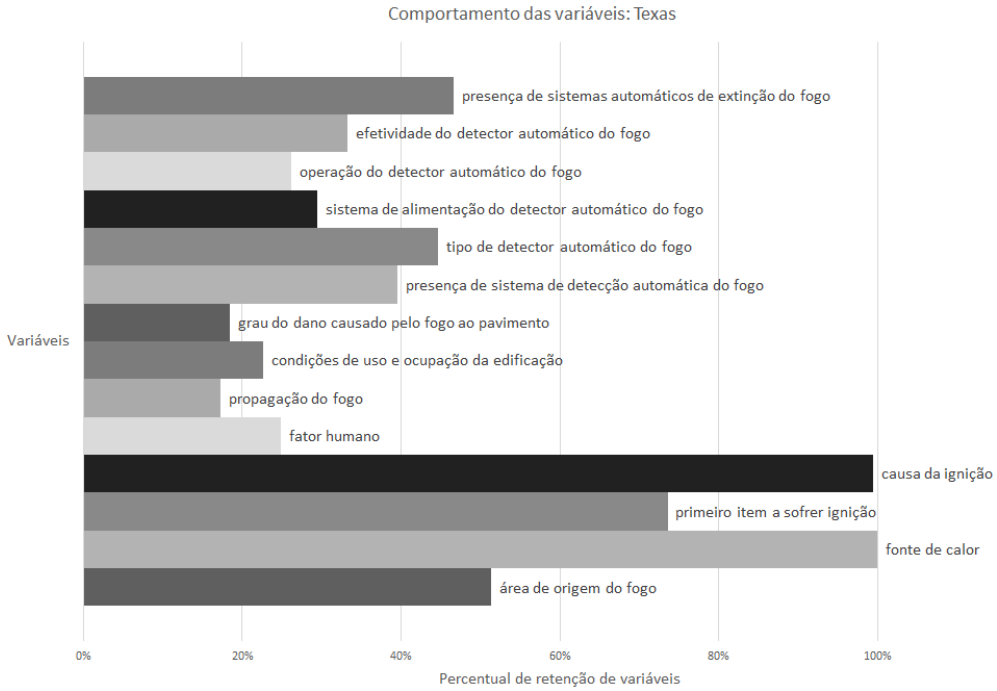

Figura 3.4: Frequência de retenção das variáveis para classificação dos incêndios no Texas

Com relação ao impacto das variáveis retidas nas classes de incêndios estudadas (*incêndio intencional*, *descuido ao cozinhar*, *mau funcionamento na rede elétrica*, *falhas em*

eletrodomésticos e *incêndio a partir de chamas abertas*) percebe-se que as variáveis se comportaram de forma análoga em ambos os bancos de dados e em todas as cinco categorias. As variáveis *fonte de calor* e *causa da ignição*, retidas em praticamente todas as repetições do procedimento classificatório, revelaram-se de elevada importância para o modelo, apresentaram resultados similares para os dois bancos de dados, conforme as Figuras 3.5 e 3.6.

Acerca do impacto das variáveis retidas sobre as categorias, a variável *fonte de calor* influencia as classes *descuido ao cozinhar*, *mau funcionamento na rede elétrica* e *falhas em eletrodomésticos*: as observações mostram que majoritariamente objetos em funcionamento foram a fonte de calor para o início do incêndio. O comportamento da variável *causa da ignição* atendeu à expectativa na classe *incêndio intencional*, haja vista que todas as observações marcaram a pontuação referente a incêndios intencionais. Já para as categorias *mau funcionamento na rede elétrica* e *falhas em eletrodomésticos*, essa mesma variável interveio fortemente mostrando resultados esperados: as pontuações se concentraram em falhas do equipamento ou da fonte de energia e causas não intencionais.

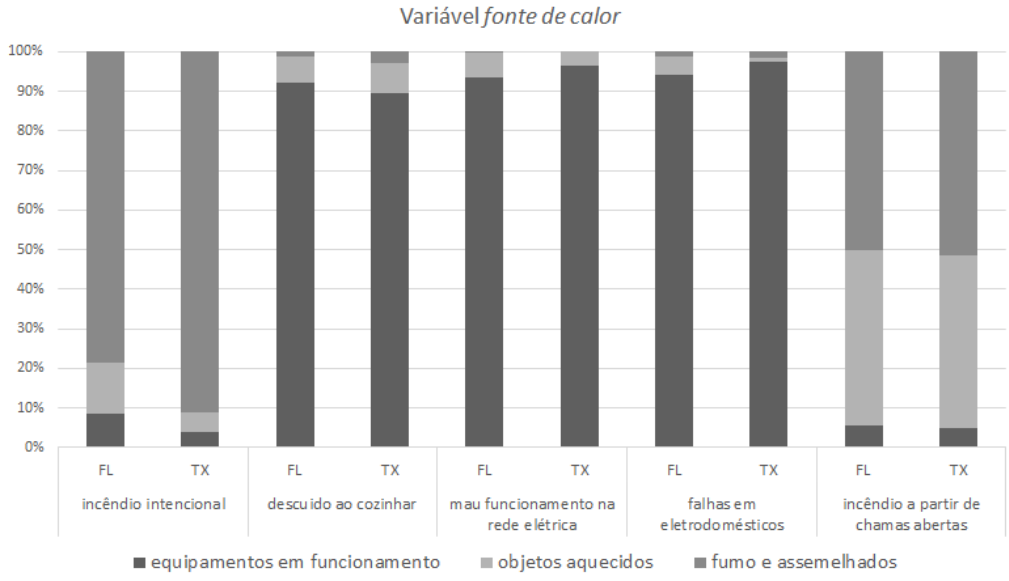

Figura 3.5: Comportamento da variável fonte de calor para os bancos de dados da Flórida (FL) e Texas (TX)

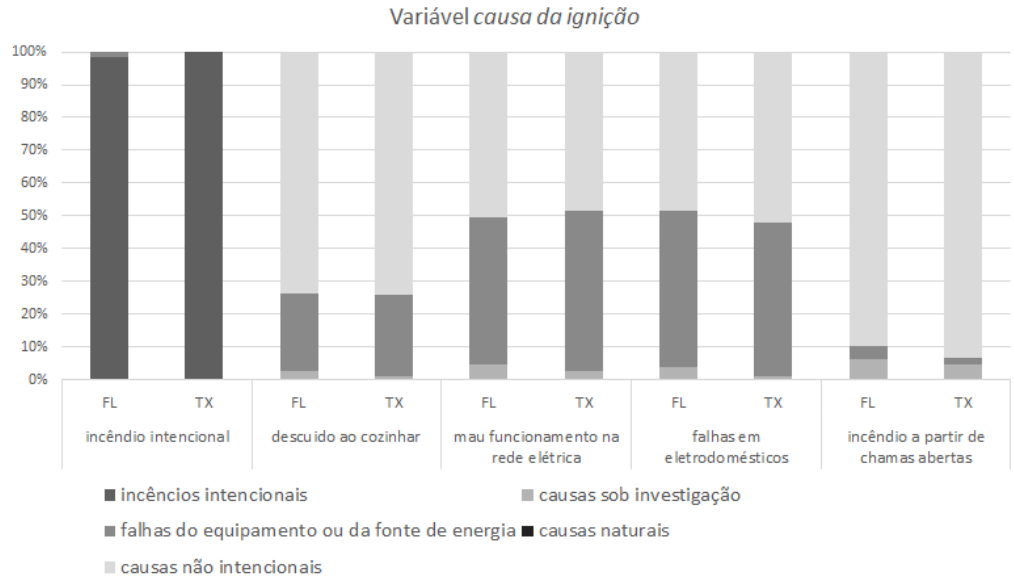

Figura 3.6: Comportamento da variável causa da ignição para os bancos de dados da Flórida (FL) e Texas (TX)

Quanto à análise das variáveis *presença de sistemas automáticos de extinção do fogo* e *tipo de detector automático do fogo*, que apresentaram importância mediana nos resultados obtidos, verificou-se que ambas não sugerem impacto diferenciado em nenhuma das cinco classes. Com um resultado semelhante nos bancos de dados da Flórida e Texas, todas as classes possuíam majoritariamente mais observações com o mesmo tipo de detector automático (detecção de fumaça) e não possuíam sistemas automáticos de extinção do fogo. Cabe ressaltar o comportamento das variáveis *operação do detector automático do fogo* e *efetividade do detector automático do fogo*: em todas as cinco categorias, verificou-se que o sistema automático de detecção do fogo operou de forma satisfatória, haja vista que alertou os ocupantes e gerou resposta ao alerta. As demais variáveis do modelo apresentaram resultados pulverizados nas classes estudadas, não justificando interpretar a influência ou impacto dessas variáveis em cada uma das categorias.

3.5 Discussão dos resultados

Os resultados obtidos com aplicação do método refletem coerência com o gerenciamento do fogo, de acordo com o reportado pela literatura. As variáveis *fonte de calor*, *causa da ignição* e *primeiro item a sofrer ignição*, apontadas como mais relevantes pela abordagem proposta, são frequentemente citadas em relatórios oficiais acerca de incêndios, sendo

prática das investigações indicarem as possíveis causas e os primeiros itens inflamados nos eventos com o fogo. Além disso, *fonte de calor* e *primeiro item a sofrer ignição* compõem permanentemente as estatísticas oficiais dos EUA. Fazem-se comuns também regulamentações exigirem testes de ignibilidade e inflamabilidade com chamas abertas e itens incandescentes para atender aos níveis básicos de segurança contra incêndio de materiais e produtos usados no cotidiano (TROITZSCH, 2016).

Entende-se a expressiva contribuição das variáveis *causa da ignição* e *primeiro item a sofrer ignição* ao modelo, tendo em vista que, conforme Troitzsch (2016), o desenvolvimento global suscitou novos e notáveis riscos, elevando o número de catástrofes causadas pelo fogo, percebendo-se ainda que uma variedade de incêndios catastróficos se deu principalmente por pequenas fontes de ignição. Técnicas modernas de proteção e prevenção contra o fogo tem desempenhado um papel importante na redução dos riscos, ao passo que se percebe que as regulamentações atuais de países desenvolvidos exigem uma série de requisitos para promover a segurança contra incêndio em edifícios altos, preconizando, entre outros, o uso de sistemas de exaustão de fumaça, escadas de fuga, sistemas de detecção e extinção automática de incêndio. Contudo, as normas de prevenção contra o fogo, geralmente, não abrangem a prescrição para a aplicação desses sistemas de segurança para edificações residenciais familiares. Diante dessa lacuna, outras exigências para garantir níveis básicos de segurança contra incêndio são impostas a materiais de construção a fim de assegurar desempenho satisfatório em relação à sua contribuição para ignição, propagação da chama e libertação de calor. De tal forma, torna-se imprescindível o conhecimento e o estudo aprofundado da *causa da ignição* e do *primeiro item a sofrer ignição* para a gestão de risco de incêndio, auxiliando na seleção de medidas de prevenção e combate ao fogo.

Sobre a variável *primeiro item a sofrer ignição*, pontua-se a preocupação constante dos órgãos governamentais na redução da carga de incêndio (soma das energias caloríficas possíveis de serem liberadas pela combustão completa de todos os materiais combustíveis em um espaço, incluindo elementos construtivos e revestimentos), tanto nos objetos interiores às edificações como nos materiais de construção que as compõem. Por exemplo, o National Institute of Standards and Technology (NIST) dos EUA desenvolveu um roteiro estratégico com o objetivo de diminuir a carga de incêndio evitável em até um terço, reduzindo o impacto do fogo através do desenvolvimento de padrões para vestuários, sistemas automáticos de extinção do fogo, cigarros com propensão reduzida à

ignição e materiais de construção. A National Fire Protection Association (NFPA) dos EUA, por sua vez, identifica o mobiliário estofado, como sofás e camas, como um item importante em termos das mortes por incêndio, tendo apresentado um panorama da investigação e propondo atividades para redução da inflamabilidade de tais itens (HALL, 2013; HAMINS et al., 2012; TROITZSCH, 2016).

As variáveis selecionadas e entendidas como detentoras de importância intermediária (*área de origem do fogo* e *tipo de detector automático do fogo*) mostraram-se coerentes à realidade do gerenciamento do fogo, seus relatórios de acompanhamento e campanhas de prevenção. O emprego de alarmes de fumaça nas residências americanas é corriqueiro: pesquisas periódicas realizadas pela NFPA indicam que 96% das edificações familiares contam com esses sistemas e que eles estavam presentes em quase três quartos dos incêndios domésticos reportados à NFIRS no período de 2009 a 2013, tendo sido acionados em mais da metade desses eventos. Salienta-se que 88% dos incidentes reportados dispunham unicamente de dispositivos de detecção de fumaça, enquanto apenas 7% contavam com a combinação de detecção de fumaça e de calor (AHRENS, 2015). Ressalta-se que essas estatísticas poderiam induzir a expectativa de maior relevância para a variável presença de sistema de detecção automática do fogo quando comparada ao atributo tipo de detector automático do fogo. Contudo, percebe-se que essas duas variáveis apresentaram resultados muito próximos em ambos os bancos analisados.

Diferentemente do esperado, a variável *presença de sistemas automáticos de extinção do fogo* (*sprinklers*) apresentou relevância mediana. Embora sejam uma parte altamente eficaz e confiável do sistema de proteção contra incêndios – estatisticamente, no período de 2010 a 2014, os *sprinklers* operaram em 94% dos incêndios residenciais reportados à NFIRS, sendo 96% eficaz nas ocorrências em que foram acionados – sua presença ainda não é frequente em edificações residenciais – apenas 8% dos incêndios residenciais relatados contavam com a presença de *sprinklers*. Apenas os Estados da Califórnia e de Maryland e a cidade de Washington (Distrito de Colúmbia) adotam como obrigatório o uso de *sprinklers* em habitações novas. A Flórida e o Texas não possuem regulamentação estadual que exija o uso de *sprinklers* em residências, mas a Flórida concede que jurisdições locais prescrevam *sprinklers*, enquanto o Texas não permite que legislações locais versem sobre esse assunto (AHRENS, 2017; NFPA, 2017). Tais circunstâncias ajudam a explicar o comportamento distinto apresentado pela variável *presença de sistemas automáticos de extinção do fogo* apresentou entre os bancos de dados, haja vista

que a Flórida permite a descentralização das legislações sobre o uso de *sprinklers*, possivelmente difundindo mais sua aplicação, enquanto no Texas essa regulamentação não pode ser realizada de forma local.

No grupo de variáveis que apresentaram pouca significância ao modelo gerado, a *propagação do fogo* e o *grau do dano causado pelo fogo ao pavimento* foram resultados prováveis, visto os bancos de dados estudados serem compostos apenas por incidentes ocorridos isoladamente em edificações residenciais de pavimento único. As demais variáveis (*fator humano*, *condições de uso e ocupação da edificação*, *presença de sistema de detecção automática do fogo*, *sistema de alimentação do detector automático do fogo*, *operação do detector automático do fogo* e *efetividade do detector automático do fogo*) ofereceram menor grau de relevância ao modelo, embora sejam repetiditamente assunto de artigos técnicos, relatórios oficiais e campanhas de prevenção e combate ao incêndio.

Tendo em consideração a quantidade de observações em cada classe, é possível explicar intuitivamente alguns resultados obtidos ao analisar o impacto das variáveis retidas sobre os argumentos. O comportamento da variável *causa da ignição* sobre a categoria *incêndio intencional* poderia servir como indício para a predição da classe da observação, visto ser esperado que a causa da ignição seja sempre intencional para essa classe. De forma semelhante, a variável *fonte de calor* revela importância para a predição nas categorias *mau funcionamento na rede elétrica* e *falhas em eletrodomésticos,* já que se espera que equipamentos em funcionamento e objetos aquecidos sejam os responsáveis pela ignição no primeiro foco de incêndio. Com relação às categorias de incêndio *descuido ao cozinhar, mau funcionamento na rede elétrica* e *falhas em eletrodomésticos*, a variável *origem do fogo* também pode interferir fortemente na predição das categorias das observações, pois se espera que o início da combustão se dê nas áreas funcionais para essas classes de incêndio.

3.6 Conclusão

Essência das pesquisas que tratam de conjuntos de observações com elevado volume de informações, a seleção de variáveis busca aumentar o desempenho de ferramentas multivariadas em contexto de predição e classificação. Entre as abordagens de mineração de dados para classificar observações, destaca-se o KNN, ordinariamente de baixo custo de implementação e de elevada eficácia. Seguindo a linha de pesquisas que se utilizam da integração de abordagens de mineração de dados, esse artigo combinou as técnicas “omita

uma variável de cada vez" e KNN com o objetivo de identificar o melhor subconjunto de variáveis de processo para inserção das observações (eventos de incêndios) em classes (causas dos eventos).

De tal forma, o método sugerido para seleção de variáveis consistiu em: (i) divisão do banco de dados em conjuntos de treino e teste, (ii) aplicação da ferramenta de mineração de dados K-vizinhos mais próximos combinada à técnica "omita uma variável por vez" para seleção de variáveis, (iii) geração do gráfico de frequência de retenção de variáveis e sua interpretação e (iv) análise qualitativa do impacto das variáveis retidas sobre os argumentos gerados.

O método proposto foi aplicado a dois conjuntos de observações distintos constituídos pelas mesmas 14 variáveis. A precisão média de classificação obtida com as observações do estado do Texas foi de 76,07% e da Flórida foi de 65,94%; a acurácia máxima foi atingida para o banco de dados do Texas com um subconjunto de 4 variáveis do total inicial de 14, enquanto para os dados da Flórida, o subconjunto que resultou maior precisão foi de 6 atributos. Destaca-se a confluência nos resultados obtidos entre os bancos de dados: a significância que os atributos ofereceram ao modelo apresentou valores semelhantes em ambos os estudos de caso, percebendo-se similaridade na hierarquia de contribuição das variáveis ao modelo. As variáveis tidas como relevantes no procedimento classificatório eram intuitivamente esperadas e têm sua importância amparada pela literatura especializada em incêndio.

Para trabalhos futuros, sugere-se a aplicação de ferramentas classificadoras mais robustas para selecionar variáveis com vistas ao aumento da acurácia. Outros desdobramentos possíveis incluem a utilização de índices de importância para seleção de atributos apoiados em outras técnicas multivariadas e posterior comparação com o método de seleção de variáveis proposto.

3.7 Referências

AHRENS, M. Smoke Alarms in U.S. Home Fires. **NFPA Research, Data and Analytics Division**, set. 2015.

AHRENS, M. U.S. Experience with Sprinklers. **NFPA Research, Data and Analytics Division**, jul. 2017.

ANDERSEN, C. M.; BRO, R. Variable selection in regression - a tutorial, **Journal of Chemometrics**, v. 24, ed. 11-12, p. 728-737, nov./dez. 2010.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. A. Selecting the best variables for classifying production batches into two quality levels. **Chemometrics and Intelligent Laboratory Systems**, v. 97, ed. 2, p. 111-117, jul. 2009.

ANZANELLO, M.J.; YAMASHITA, G.; MARCELO, M.; FOGLIATTO, F.S.; ORTIZ, R.S.; MARIOTTI, K.; FERRÃO, M.F. A genetic algorithm-based framework for wavelength selection on sample categorization. Drug. **Testing and Analysis**, fev. 2017.

ANZANELLO, M.J.; FOGLIATTO, F.S.; ORTIZ, R.S.; LIMBERGER, R.; MARIOTTI, K. Selecting relevant Fourier transform infrared spectroscopy wavenumbers for clustering authentic and counterfeit drug samples. **Science and Justice**, v. 54, p. 363-368, set. 2014.

ANZANELLO, M. J.; FOGLIATTO, F. S.; ROSSINI, K. Data mining-based method for identifying discriminant attributes in sensory profiling. **Food Quality and Preference**, v. 22, ed. 1, p. 139-148, jan. 2011.

ANZANELLO, M. J.; FOGLIATTO, F. S. Selecting the best clustering variables for grouping mass-customized products involving workers' learning. **International Journal of Production Economics**, v. 130, ed. 2, p. 268-276, abr. 2011.

ANZANELLO M. J., ORTIZB, R.S., LIMBERGERB, R. P., MAYORGA, P. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v. 83, p. 209-214, set. 2013.

ASGARY, A.; GHAFFARI, A.; LEVY, J. Spatial and temporal analyses of structural fire incidents and their causes: A case of Toronto, Canada. **Fire Safety Journal**, v. 45, p. 44–57, jan. 2010.

CEYHAN, E.; ERTUGAY, K.; DÜZGÜN, S. Exploratory and inferential methods for spatio-temporal analysis of residential fire clustering in urban areas. **Fire Safety Journal**, v. 58, p. 226-239, maio 2013.

CHEN, H.; PITTMAN, W. C; HATANAKA, L. C.; HARDING, B. Z.; BOUSSOUF, A.; MOORE, D. A.; MILKE; J. A.; MANNAN, M. S. Integration of process safety engineering and fire protection engineering for better safety performance. **Journal of Loss Prevention in the Process Industries**, v. 37, p. 74-81, set. 2015.

CHAOVALITWONGSE, W. A.; FAN, Y.; SACHDEO, R. C. On the Time Series K-Nearest Neighbor Classification of Abnormal Brain Activity. **IEEE Transactions on Systems Man and Cybernetics** - Part A Systems and Humans, v. 37, n. 6, p. 1005-1016. dez. 2007.

CORCORAN, J.; HIGGS, G.; BRUNSDON, C.; WARE, A.; NORMAN, P. The use of spatial analytical techniques to explore patterns of fire incidence: A South Wales case study. **Computers, Environment and Urban Systems**, v. 31, ed. 6 p. 623-647, nov. 2007.

DUDA, R.; HART, P.; STORK, D., **Pattern Classification**. ed. 2. New York: Wiley-Interscience, 2001.

GAUCHI, J.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v. 58, ed. 2, p. 171-193, out. 2001.

GENG, X.; LIU, T. Y.; QIN, T.; ARNOLD, A.; LI, H.; SHUM, H. Y.. Query dependent ranking using K-nearest neighbor. Annual ACM Conference on Research and Development in Information Retrieval. **Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval**, SIGIR 2008, Singapore, p. 115–122, 20-24, jul. 2008.

GUNVIG, A.; HANSEN, F.; BORGGAARD, C. A mathematical model for predicting growth/no-growth of psychrotrophic C. botulinum in meat products with five variables. **Food Control**, v. 29, ed. 2, p. 309-317, fev. 2013.

GUYSON, I.; ELISSEEFF, A.; An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research**, v. 3, p. 1157-1182, jan. 2003.

HALL, J. R. **White paper on upholstered furniture flammability**. National Fire Protection Association, set. 2013.

HAMINS, A.; AVERILL, J; BRYNER, N.; GANN, R.; BUTRY, D.; DAVIS, R.; AMON, F.; GILMAN, J.; MARANGHIDES, A.; MELL, W.; MADRZYKOWSKI, D.; MANZELLO, S.; YANG, J.; BUNDY, M. **Reducing the risk of fire in buildings and communities: a strategic roadmap to guide and prioritize research** (NIST Special Publication 1130). National Institute of Standards and Technology, abril 2012.

HASTI, E, C.; SEARLE, R. Socio-economic and demographic predictors of accidental dwelling fire rates. **Fire Safety Journal**, v. 84, p. 50-56, ago. 2016.

JENNINGS, C. R. Social and economic characteristics as determinants of residential fire risk in urban neighborhoods: A review of the literature. **Fire Safety Journal**, v. 62, parte A, p.13–19, nov. 2013.

KUDO, M.; SKLANSKY, J. Comparison of algorithms that select features for pattern classifers. **Pattern Recognition**, v. 33, ed. 1, p. 25-41, jan. 2000.

LIZHI, W.; AIZHU, R. Urban Fire Risk Clustering Method Based on Fire Statistics. **Tsinghua science and technology**, v. 13, suplemento 1, p. 418-422, out. 2008.

MADRZYKOWSKI, D.; KERBER, S. New Fires, New Tactics. **NFPA Journal**, 29 dez. 2014. Entrevista por Jesse Roman. Disponível em: http://www.nfpa.org/news-and-research/publications/nfpa-journal/2015/january-february-2015/features/fire-tactics. Acesso em: 10 abr. 2016.

NATIONAL FIRE PROTECTION ASSOCIATION. Fire Sprinkler Initiative. **Home fire sprinkler requirements at a glance**. Disponível em:

<http://www.firesprinklerinitiative.org/legislation/sprinkler-requirements-by-state.aspx>. Acesso em 30 ago. 2017.

SANDELOWSKI, M. Combining Qualitative and Quantitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method Studies**. Research in Nursing & Health**, v. 223, p. 246–255, jun. 2000.

SANDELOWSKI, M. Real Qualitative Researchers Do Not Count: The Use of Numbers in Qualitative Research. **Research in Nursing & Health**, v. 24, p. 230–240, jul. 2001.

SPINARDI, G.; BISBY, L.; TORERO, J. A Review of Sociological Issues in Fire Safety Regulation. **Fire Technology**, p. 1-27, 26 jul. 2016.

TASHAKKORI, A.; TEDDLIE, C. **Handbook of Mixed Methods in Social & Behavioral Research**. Thousand Oaks: Sage Publications, 2002.

TROITZSCH, J. H. Fires, statistics, ignition sources, and passive fire protection measures. **Journal of Fire Sciences**, v. 34 (3), p. 171-198, fev. 2016.

UNITED STATES FIRE ADMINISTRATION. Federal Emergency Management Agency. **An NFIRS Analysis: Investigating City Characteristics and Residential Fire Rates**, abr. 1998.

UNITED STATES FIRE ADMINISTRATION. Federal Emergency Management Agency. **Fire in the United States, 2005-2014**. 18 ed., jan. 2017.

UNITED STATES FIRE ADMINISTRATION. National Fire Data Center. **National Fire Incident Reporting System 5.0** - Complete Reference Guide, 515 p., jan. 2015.

WAN, C. H.; LEE, L. H.; RAJKUMAR, R.; ISA, D. A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. **Expert Systems with Applications**, v. 39, ed. 15, p. 11880–11888, nov. 2012.

WU, X.; KUMAR, V.; QUINLAN, R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z. H.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, ed. 1, p. 1-37, jan. 2008.

Anexo B: Quantificação das observações para transformar dados qualitativos em dados quantitativos

Variável	Quantificação da observação	
	Qualitativo	Quantitativo
área de origem do fogo	áreas estruturais	0,10
	áreas de ingresso	0,30
	áreas funcionais	0,50
	áreas de estocagem	0,80
fonte de calor	equipamentos em funcionamento	0,30
	objetos aquecidos	0,50
	fumo e assemelhados	0,70
primeiro item a sofrer ignição	componentes estruturais da edificação	0,10
	suplementos estocados	0,30
	mobiliário	0,40
	vestuário	0,50
	adorno, enfeite e material decorativo	0,60
	materiais gerais pesados (como fios elétricos, transformadores ou pneus)	0,70
	materiais gerais leves (como livros, revistas, papéis ou tecidos)	0,80
causa da ignição	incêndios intencionais	0,10
	causas sob investigação	0,30
	falhas do equipamento ou da fonte de energia	0,50
	causas naturais	0,80
	causas não intencionais	1,00
fator humano	possível incapacidade mental	0,10
	incapacidade física	0,20
	possivelmente sob efeito de álcool ou drogas	0,30
	idade foi um fator agravante	0,40
	adormecido	0,50
	desacompanhado ou sozinho	0,70
	múltiplas pessoas envolvidas no incidente	0,80
	não houve nenhum fator humano relacionado ao incêndio	1,00
propagação do fogo	chamas foram além da edificação de origem	0,00
	fogo permaneceu confinado na edificação de origem	0,25
	fogo permaneceu confinado no pavimento de origem	0,50
	fogo permaneceu confinado na dependência de origem	0,75
	chama não se propagou e permaneceu confinada no objeto que ocorreu a ignição	1,00
condições de uso e ocupação da edificação	em demolição	0,30
	desabitada e em condições inseguras (abandonada)	0,40
	em construção	0,50
	em reforma	0,60
	desabitada e segura	0,70
	desocupada (não utilizada frequentemente)	0,80
	em uso e ocupação normal	1,00
grau do dano causado pelo fogo ao pavimento	danos extremos	0,10
	danos severos	0,30
	danos moderados	0,60
	danos leves	0,90
presença de sistema de detecção automática do fogo	ausente	0,00
	presente	1,00
tipo de detector automático do fogo	detecção de calor	0,60
	combinação de detecção de calor e fumaça	0,90
	detecção de fumaça	1,00
sistema de alimentação do detector automático do fogo	somente a bateria	0,20
	ligado à tomada	0,40
	ligado direto à fiação elétrica	0,60
	ligado à tomada com bateria	0,80
	ligado direto à fiação elétrica com bateria	1,00
operação do detector automático do fogo	falhou	0,10
	não foi ativado devido à baixa magnitude do fogo	0,50
	operou	1,00
efetividade do detector automático do fogo	houve falha no alerta aos ocupantes	0,10
	não havia ocupantes na edificação durante o incêndio	0,50
	sistema alertou os ocupantes, mas ocupantes falharam na resposta	0,80
	sistema alertou os ocupantes e os ocupantes responderam	1,00
presença de sistemas automáticos de extinção do fogo	ausente	0,20
	parcialmente instalado	0,90
	presente	1,00

4 CONSIDERAÇÕES FINAIS

Ferramentas multivaridas têm sido amplamente aplicadas em estudos voltados à melhor compreensão das tendências dos incidentes de incêndio para orientar as políticas de prevenção, proteção e combate ao fogo. A análise de *clusters* constitui-se em uma das técnicas mais empregadas em dados de incêndio para agrupar as observações visando identificar as similaridades que definem cada grupo. Também no sentido de auxiliar a compreensão de conjunto de observações, a seleção de variáveis objetiva aumentar o desempenho de ferramentas multivariadas em contexto de predição e classificação.

Essa dissertação apresentou dois artigos independentes que apresentaram metodologias que se utilizam do ferramental multivariado para analisar as observações em grupos com características similares (primeiro artigo) e identificar as variáveis mais relevantes para classificação de eventos de incêndio em classes, classes essas entendidas como distintas causas do incêndio (segundo artigo). Ressalta-se ainda que a sistemática proposta no segundo artigo preenche uma lacuna identificada nos estudos de ocorrências de incêndio: não se encontrou um estudo que fizesse uso da técnica KNN para classificação de observações, em detrimento do elevado números trabalhos que empregam da análise de *clusters*.

Alinhado com a tendência corrente de análise, o primeiro artigo empregou o método *K-Means* para aprimoramento das técnicas hierárquicas e a ACP para melhorar a qualidade da análise de clusterização, avaliada do através do SI. As etapas dessa pesquisa compreenderam em (i) geração do dendrograma para definição do número de clusters a serem formados, (ii) clusterização via K-means utilizando as variáveis originais que descrevem os eventos de incêndio, (iii) avaliação da qualidade dos agrupamentos gerados através do *Silhouette Index*, (iv) interpretação dos agrupamentos formados; e (v) repetição das etapas (i) a (iv) para dados transformados pela ACP. A utilização das componentes principais como parâmetros de entrada da clusterização, em substituição às variáveis originais, objetiva avaliar um eventual aprimoramento da qualidade dos *clusters* gerados com base na utilização de variáveis não correlacionadas.

Ainda no primeiro artigo, os dendrogramas de ambos bancos de dados recomendaram K=3 para a aplicação do método *K-Means*. A distribuição das observações nesses *clusters* foram similares para a Flórida e Texas, sendo os 3 grupos rotulados como (i) ocorrências de incêndio que não possuíam de sistema de detecção automática do fogo, (ii) ocorrências

de incêndio em que o sistema de detecção automática do fogo operou plenamente, e (iii) ocorrências de incêndio em que a operação do sistema de detecção automática do fogo falhou. As características dos clusters gerados acompanham a literatura especializada e os relatórios oficiais acerca de incêndio. Percebeu-se que, para as observações da Flórida, as correlações tipicamente verificadas nas variáveis originais não comprometem a qualidade da clusterização, já que a qualidade do procedimento é semelhante ao agrupamento gerado com base em 3 CPs. Em contrapartida, percebeu-se que, nos dados do Texas, os SIs médios verificados indicaram que a clusterização realizada sobre variáveis não correlacionadas (CPs) aumenta a qualidade da clusterização.

O segundo artigo associou as técnicas "omita uma variável de cada vez" e KNN com o objetivo de identificar o melhor subconjunto de variáveis de processo para inserção das observações (eventos de incêndios) em classes (causas dos eventos). As etapas da sistemática consistiram em: (i) divisão do banco de dados em conjuntos de treino e teste, (ii) aplicação da ferramenta de mineração de dados K-vizinhos mais próximos combinada à técnica "omita uma variável por vez" para seleção de variáveis, (iii) geração do gráfico de frequência de retenção de variáveis e sua interpretação e (iv) análise qualitativa do impacto das variáveis retidas sobre os argumentos gerados.

A precisão média de classificação obtida no segundo artigo com as observações do Texas foi de 76,07% e da Flórida foi de 65,94%; a acurácia máxima foi alcançada com as observações do Texas com um subconjunto de 4 variáveis, enquanto que, para os dados da Flórida, o subconjunto que resultou em maior precisão reteve 6 atributos. Ressalta-se a convergência nos resultados obtidos entre os dois bancos de dados: a significância que os atributos ofereceram ao modelo apresentou valores semelhantes em ambos os estudos de caso, notando-se semelhança na hierarquia de contribuição das variáveis ao modelo. As variáveis identificadas como relevantes no procedimento de classificação satisfazem a expectativa e têm sua importância amplamente respaldada pela literatura especializada em incêndio.

Frente ao exposto, entende-se que os objetivos descritos no capítulo 1 dessa dissertação foram atingidos, uma vez que as ferramentas multivariadas citadas foram aplicadas satisfatoriamente aos conjuntos de observações estudados, sendo possível gerar resultados com qualidade e interpretá-los com apoio na literatura.

Futuros desdobramentos deste estudo incluem: (i) emprego de outras ferramentas multivariadas para a formação dos agrupamentos, comparando os resultados obtidos com a sistemática proposta no primeiro artigo; (ii) uso de diferentes métricas para avaliar a qualidade dos *clusters* gerados; (iii) aplicação de técnicas classificadoras mais robustas para selecionar variáveis objetivando o aumento da acurácia; (iv) utilização de índices de importância para seleção de atributos fundamentados em outras técnicas multivariadas e posterior comparação com o método de seleção de variáveis proposto no segundo artigo; e (v) aplicação dos métodos propostos utilizando registros de incêndio no Brasil.