University of Tennessee, Knoxville

Trace: Tennessee Research and Creative Exchange

12-2017

Analysis of NFIRS Data for Sensitivity to Foreclosure and Other Select Features

Alexander Meadows Asbury

University of Tennessee, Knoxville, aasbury4@vols.utk.edu

Recommended Citation

Asbury, Alexander Meadows, "Analysis of NFIRS Data for Sensitivity to Foreclosure and Other Select Features. " Master's Thesis, University of Tennessee, 2017.
https://trace.tennessee.edu/utk_gradthes/4944

To the Graduate Council:

I am submitting herewith a thesis written by Alexander Meadows Asbury entitled "Analysis of NFIRS Data for Sensitivity to Foreclosure and Other Select Features." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

David J. Icove, Major Professor

We have read this thesis and recommend its acceptance:

Benjamin J. Blalock, Michael A. Langston

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Analysis of NFIRS Data for Sensitivity to Foreclosure and Other Select Features

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Alexander Meadows Asbury

December 2017

To my parents, none of this would have happened without you

Acknowledgments

I would like to thank my parents, Pam and Mike, who have supported and encouraged all my interests. I know I would not have made it this far without such generous support.

I would like to thank Dr. David Icove. His generosity with his time and experience have been critical factors in the completion of this work and my understanding of Fire Protection Engineering. I would also like to thank the members of my thesis committee: Drs. Benjamin J. Blalock and Michael A. Langston. Additionally, I would like to thank Dr. Charles Phillips for sharing his knowledge of statistics and graph theoretical algorithms towards the completion of this work. I also feel it is important to acknowledge all of the authors whose works are cited in this thesis. The bodies of knowledge they have assembled and distributed have been critical in not only the the completion of this thesis, but surely many others as well.

I would like to thank my friends, particularly Michael, Jake, and Harley, who have endured all my anxieties and frustrations throughout this process. If it takes a village to raise a child, it at least takes a few patient friends to raise a Master's degree.

Abstract

Arson is a grave threat to life and property. In the United States, fire information is collected and disseminated through the National Fire Incident Reporting System (NFIRS). Fire records obtained from NFIRS contain a full range of available information. This information includes the initial incident details in addition to investigative information regarding the cause of ignition and factors contributing to ignition. Combating the arson problem is accomplished in large part by understanding the motives and opportunities of those who commit arson. A common motive for arson is financial gain through insurance fraud. By connecting NFIRS data with mortgage and foreclosure information from RealtyTrac, insight into potential incidents of insurance fraud may be obtained.

Understanding the features that intentional fires have in common is necessary to assess the vulnerability of structures to intentional burning. One historically utilized method of predicting arson prone structures is linear discriminant analysis (LDA). LDA is a method of separating objects or events into two or more categories using a combination of features. Through feature analysis and selection, a discriminant function is proposed that incorporates foreclosure as an independent variable to classify fires as intentional or unintentional.

Additionally, graph theoretical algorithms for clustering are applied in support of the discovery of novel relationships between fires. In this thesis we leverage the paraclique algorithm, which has previously been applied to biological data, to help reveal latent associations within the NFIRS datasets.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In [4], arson is defined as “the willful and malicious burning of another’s property or the burning of one’s own property for some improper purpose such as defrauding the insurer.” People commit arson for various reasons, such as financial gain, revenge, and for the thrill of crime. The arson rate in the US may be dramatically unreported as shown in [14], a fact that underscores the seriousness of this topic. However disparate the reasons, the structures targeted for arson have been shown in the past to be similar across several variables as discovered in [4] and [13].

Currently, the United States collects data from fire departments in every state through the National Fire Incident Reporting System (NFIRS). The U.S. Fire Administration estimates that NFIRS receives data on 75% of the fires that occur annually [30]. Combining this fire data with financial data from RealtyTrac [3], a leading realty data provider, creates the potential to gain new insights into the fire problem in the United States, since it contains information on mortgages and foreclosures.

There is a saying that the definition of spontaneous ignition is when a paid up insurance premium rubs up against a past-due notice. The goal of this thesis is to delve into a dataset and extract meaningful relationships between intentional fires, foreclosure, and other key factors. Discriminant analysis is a well-known and time-honored approach to classification both in the field of Fire Protection and in many others. In contrast, graph theoretical algorithms have shown an ability to highlight latent relationships within biological data. In this thesis, we seek to demonstrate that the graph theoretical paraclique algorithm [10] can provide meaningful results when applied to NFIRS fire data.

This thesis is structured as follows.

- Chapter 2 provides a review of topical literature on foreclosure and arson research in the United States.
- Chapter 3 discusses the methods explored in this thesis.
- Chapter 4 reports all of the results generated by the methods from Chapter 3 and provides a discussion of these results.
- Chapter 5 reviews the results and findings in brief and provides indications for future work.

Chapter 2

Literature Review

The purpose of this chapter is to provide a survey of the various papers and articles produced concerning the topics and methods researched and implemented for this work. Thus, this chapter is broken into sections concerning foreclosure, arson, and other works where similar analysis tools are applied.

Previous research has looked into the development of an algorithm for estimating arson risk. As the subject area generally presents itself as a pattern recognition and statistics problem, this thesis focuses on literature following this approach to arson analysis. This work leverages the historical approach taken by [8] where Ronald Fisher utilized discriminant analysis to build a classification scheme to predict the species of Iris flowers based upon the flower's measured features. Discriminant analysis weighs and combines the input variables to produce decision functions. These decision functions are used to classify new samples by discriminating them into the appropriate class [12]. With Fisher's approach, the classic example of discriminant analysis, various Iris species are discriminated into the correct classification through the study of four variables: sepal width and length and petal width and length. This methodology is well suited and historically used for analyzing arson cases.

2.1 Foreclosure and the Effects of Financial Distress

The financial crisis of the late 2000s led to an increase in residential foreclosures across the country. There is evidence that certain types of crime may be associated with an increased rate of foreclosure for residential properties.

In [7], an attempt is made to link foreclosure to crime. The authors speculate on the many ways that foreclosure might affect crime, such as deterioration of the building from lack of maintenance and weakened community bonds from high resident turnover. Additionally, the authors speculate that foreclosures can cause buildings to be vacant for extended periods. These vacant structures are then a convenient target for criminals. To study these trends, the authors opted to use information from New York City and a difference-in-differences model to analyze crime, foreclosure, and other data. The model focuses on how decisions are made by potential criminals. The authors looked at how foreclosures and their after-effects alter the opportunities available to potential offenders. In the case of lengthy foreclosures, the property is more likely to be unoccupied. These vacant structures can be easily targeted by vandals and trespassers. The authors suggest that further research will be needed to understand whether foreclosures move crime from one area to another or encourage new crimes to take place.

2.2 Arson in the United States

2.2.1 Baxi's Work from the 1980s

In [4], Baxi applied various sources of data to construct a discriminant function toward predicting structures that are prone to arson. The purpose of this paper was to aid in the development of an Early Warning System for the city of Newark, New Jersey as per the interest of the Newark Fire Department. Baxi asserts that discriminant analysis is a common technique in the field of arson prevention. His study included dwellings and commercial or industrial properties. Garages, even when not physically attached to the home, were considered a part of the dwelling. Arsons of automobiles and fires outside of a structure were excluded. To ensure accuracy, Baxi carefully considered the arson and nonarson cases to be included in the study. Each case of arson was matched with an as closely related nonarson structure as possible. In the years before this study, [19] and [27] had undertaken similar endeavors. The results and effectiveness of the author's effort are compared to techniques from [19] and [27]; this study's results are found to be superior. Baxi notes some of the findings from [27] were useful in his study:

- Buildings where an arson took place were nearly twice as likely to be located on block corners.
- In general, nonarson structures were smaller than arson structures.
- Nonarson structures had fewer building code violations.
- Nonarson structures tended to make lower claims on insurance than structures that suffered arson.

Baxi made the conscious decision to select all of the matched buildings carefully. Each of the matches would have no history of arson (though a nonarson fire history was acceptable), be located in a similar area, and be used for the same purpose. Brief mention is made that the stability of the neighborhood may be related to the number of occupied structures in the neighborhood or the economic stability, similar to the findings of [7].

For the analysis, 150 structures were selected at random from a total 987 available. Of those 150, matched cases were available for 127. However, 25 cases were discarded as either unsuitable or had outlying values for the variables under study.

The data used for this study was collected from various sources, mostly departments of the City of Newark. Information about the number of fires was collected from "White Cards" belonging to the Newark Fire Department for the period January 1, 1978 to April 30, 1981. The Fire Department also provided the number of fire code violations for the period January 1, 1979 to April 30, 1981. The Code Enforcement Department provided building code, electrical code, and health code violations. For about half of the cases studied, the Code Enforcement Department was unable to provide a record of any violations. Violations were then divided into serious and nonserious categories based on the author's intuitive discretion. Crime data was provided by the Newark Police Department. Information obtained was only recorded in instances where the police made a visit to each of the structure's addresses. The crime data for the period January 1, 1977 to April 30, 1981 was then divided into Indexed (Part I) and Non-indexed (Part II) offenses. Available tax information for the structures under study was collected from the Newark Department of Tax Collection and Assessment and the Newark Water Department. Finally, information on the structure's insurance status, insurance amount, property loss claim value, and ownership was acquired through the New Jersey Insurance Underwriter's Association. The author had particular interest in the amount of insurance coverage the structure had before its first arson. Insurance data was available to the author as far back as 1976. With the insurance information in hand, the author then developed a simple "insurance score" based on the number of owners, the change of insurance coverage, and the amount of loss claimed due to fire.

The author uses all of the various data sources mentioned above to begin his analysis. Some of the significant relationships uncovered include the following:

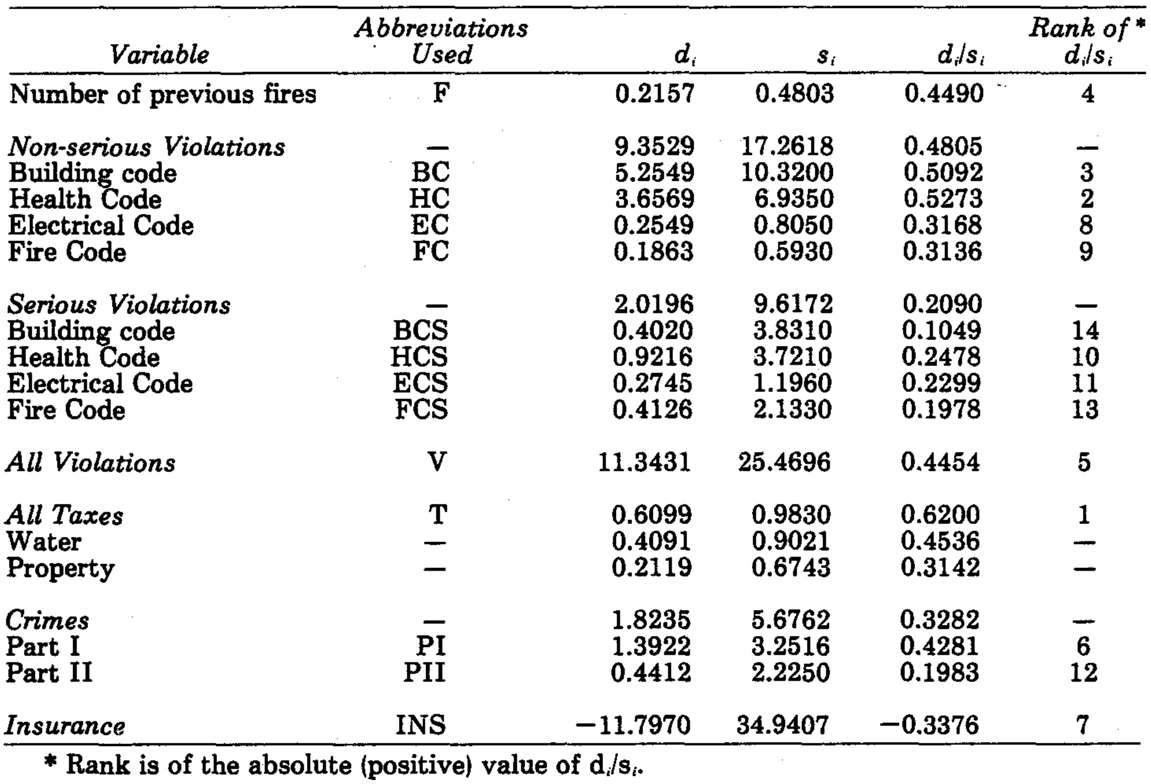
- A distribution of the "total amount of all taxes due" shows significant difference between the arson structures and the matched nonarson structures. The mean amount of taxes due also bears out this relationship.

- The "total number of previous fires" does not appear to differ significantly between the two samples.

- For the two populations, the distribution of Part II crimes does not differ significantly. However, for Part I crimes, the distribution is significantly different. Additionally, the total number of crimes between the two populations differs significantly.

- The average insurance scores between the arson structures and the matched structures is significantly different.

Baxi concludes that the following features are important for identifying arson and match cases: Total tax amount due, nonserious code violations of all types, indexed crimes visited upon the structure, and the insurance score. To begin his discriminant analysis, Baxi looks at 14 variables related to arson of a structure; his table for these values shown in Figure 2.1.



Variable	*Abbreviations Used*	d_i	s_i	d_i/s_i	*Rank of** d_i/s_i
Number of previous fires	F	0.2157	0.4803	0.4490	4
Non-serious Violations	—	9.3529	17.2618	0.4805	—
Building code	BC	5.2549	10.3200	0.5092	3
Health Code	HC	3.6569	6.9350	0.5273	2
Electrical Code	EC	0.2549	0.8050	0.3168	8
Fire Code	FC	0.1863	0.5930	0.3136	9
Serious Violations	—	2.0196	9.6172	0.2090	—
Building code	BCS	0.4020	3.8310	0.1049	14
Health Code	HCS	0.9216	3.7210	0.2478	10
Electrical Code	ECS	0.2745	1.1960	0.2299	11
Fire Code	FCS	0.4126	2.1330	0.1978	13
All Violations	V	11.3431	25.4696	0.4454	5
All Taxes	T	0.6099	0.9830	0.6200	1
Water	—	0.4091	0.9021	0.4536	—
Property	—	0.2119	0.6743	0.3142	—
Crimes	—	1.8235	5.6762	0.3282	—
Part I	PI	1.3922	3.2516	0.4281	6
Part II	PII	0.4412	2.2250	0.1983	12
Insurance	INS	−11.7970	34.9407	−0.3376	7

* Rank is of the absolute (positive) value of d_i/s_i.

Figure 2.1: Table 1 from [4]

In Figure 2.1 from [4], d_i is the difference in the average value of the arson and matched populations for each variable i and s_i is the mean error sum of squares from the analysis of variance for each variable i. The ratio of these quantities is used to rank each variables ability to minimize the probability of misclassification. Here, Baxi notes that the insurance variable ranks higher in its ability to discriminate arson prone structures than the all of the code violation variables.

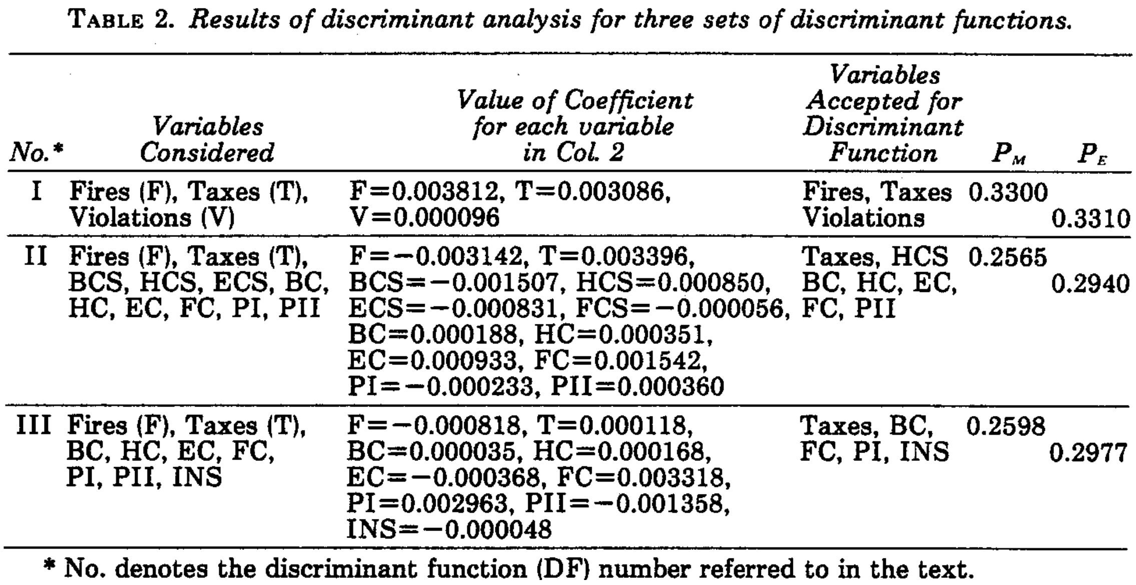


TABLE 2. *Results of discriminant analysis for three sets of discriminant functions.*

*No.**	*Variables Considered*	*Value of Coefficient for each variable in Col. 2*	*Variables Accepted for Discriminant Function*	P_M	P_E
I	Fires (F), Taxes (T), Violations (V)	F=0.003812, T=0.003086, V=0.000096	Fires, Taxes Violations	0.3300	0.3310
II	Fires (F), Taxes (T), BCS, HCS, ECS, BC, HC, EC, FC, PI, PII	F=−0.003142, T=0.003396, BCS=−0.001507, HCS=0.000850, ECS=−0.000831, FCS=−0.000056, BC=0.000188, HC=0.000351, EC=0.000933, FC=0.001542, PI=−0.000233, PII=0.000360	Taxes, HCS BC, HC, EC, FC, PII	0.2565	0.2940
III	Fires (F), Taxes (T), BC, HC, EC, FC, PI, PII, INS	F=−0.000818, T=0.000118, BC=0.000035, HC=0.000168, EC=−0.000368, FC=0.003318, PI=0.002963, PII=−0.001358, INS=−0.000048	Taxes, BC, FC, PI, INS	0.2598	0.2977

* No. denotes the discriminant function (DF) number referred to in the text.

Figure 2.2: Table 2 from [4]

Next, Baxi uses his table of variables to construct several discriminant functions using various subsets of the variables from Figure 2.1. Three functions are constructed and compared in Figure 2.2. The table shows that different variables may be used together to obtain different probabilities of correct classification. Given the slight difference in the classification abilities of Functions II and III, the ultimate decision to choose III is made due to the fewer number of variables used to discriminate. Using few variables to discriminate allows for easier, more efficient analysis due to decreased overhead from collecting less information for each case. Additionally, by reducing the number of features, cases are more likely to contain all the necessary information. To reinforce this determination between his discriminant functions, Baxi turned to a cost-efficiency calculation. The cost-efficiency calculation looks at the probability that an arson prone structure will be misclassified as a nonarson prone structure and vice versa. For his analysis, it is considered a more serious error if an arson prone structure is mislabeled than if a nonarson prone structure is mislabeled. The cost of an incident of arson is assumed to be much greater than the cost of programs designed to prevent arson. By assuming different ratios of cost, the author establishes the effectiveness of his analysis. By setting the cost of an arson to twice the cost of preventing an arson, the author concludes that Function I would cost 11.8 percent more than Function III. Changing the cost ratio to five increases the percentage to 28.8.

Deciding on Function III, Baxi outlines the steps to use his function. Function III from [4] can be seen in Equation 2.1. Using a decision threshold of 50, any structure for which the result of Equation 2.1 is greater than or equal to 50 is a more likely target of arson. If Equation 2.1 is less than 50, then it is less likely the structure will have an arson. Testing of the function was performed on a set of data for 33 structures. Out of the 33 structures, 24 fires were not arson, and 9 were cases of arson. However, for these cases insurance and Part I crime data were unavailable. Part I crime data values were estimated from the other data. Lacking this information, the probability of misclassification for a result was 36 percent. One feature of note is that, even with the missing information, the function was more likely to classify nonarsons as arsons, than it was to classify arsons as nonarsons. In this way, the function, even when under stress from lack of data, errs on the side of caution.

$$Y = 3.37X_1 + 94.80X_2 + 84.66X_3 - 1.37X_4 + X_5 \tag{2.1}$$

where

Y = Discriminant Score

X_1 = Unpaid Taxes in Thousands of Dollars

X_2 = Number of Nonserious Fire Code Violations

X_3 = Number of Indexed Crimes

X_4 = Insurance Score

X_5 = Number of Nonserious Building Code Violations

After establishing the effectiveness of his function, Baxi moves to compare it with that of functions developed in [29] and [27]. His function has a probability of correct classification of 69.1 percent. The discriminant function from [27] has a probability of correct classification of 78 percent. The discriminant function from [29] has a probability of correct classification of 74.4 percent. From this, it is evident that Baxi's Newark function has the lowest probability of correct classification. Addressing this, Baxi then looks at the relative cost efficiency of each function. He finds that his function is much more accurate when misclassifying arsons as nonarsons.

Baxi uses hand analysis to demonstrate the application of discriminant analysis to the problem of predicting arson prone structures. However, Baxi invested considerable amounts of time and resources in his investigation. In the paper, he noted that he visited as many of the structures he chose to study as possible and performed all the calculations for his work manually. These steps are neither an ideal nor realistically repeatable situation for efficient data processing. Baxi's conclusion about the cost of the function precludes any mention of the cost of gathering and processing the data. Instead, his cost of preventing the arson is assumed to be directed at resident awareness and surveillance programs.

2.2.2 Profiling Arsonists

In [1], the authors sought to understand the profile of arsonists and arson structures. When arson was temporarily added as an indexed crime by Congress in 1978 and permanently so in 1982, more data was made available to researchers so that they may understand and assist in the combat of arson. The authors pointed out that, unlike other indexed crimes, arson was under reported. That is, there was difficulty in establishing a motive, victim, or occurrence of a crime in many cases of a suspicious fire. Arson, as a crime, often conceals itself and rarely provides witnesses. The authors noted that at the time approximately one-seventh of all structures that experienced arson were not occupied at the time. Additionally, the portion of arsons cases cleared by law enforcement was the lowest of any indexed crime, only 17 percent for the year of 1983. According to the report, of the cleared cases 34 percent involved minors, greater than the average of all property crimes committed by minors. In 1983, nearly 25 percent of arson-related arrests were of persons under 15 years of age, and persons under 25 years of age accounted for over 60 percent of arrests. Next, the authors looked at the divide between the sexes in the commission of arson. In 1965, male arsonists outnumbered female arsonists 12 to 1. By 1983, this ratio had fallen to 8 to 1. On the topic of race, the authors found that, unlike gender, the pattern had remained relatively nonvolatile in the period of 1965 to 1983. The ratio between nonwhite and white arson arrests was 2 to 1. In the paper, it is also noted the arson has the fewest repeat arrestees of any Indexed crime. In general, the authors found that, at the time, a person arrested for arson was often a young, white male.

2.2.3 Pattern Recognition and the Arson Problem

In [12], arson occurrences were analyzed based on whether the structure was occupied, type of applicant, ISO fire protection class, cash values of the structure, mortgages, and previous losses at the address. Discriminant analysis was used to look at a set of twenty insurance applications submitted to the Nationwide Insurance Company. These applications, evaluated by the company as accept or deny, were used by the author as a training set for the decision function. Each application provided a data point for the variables under study listed above. Figure 2.3 details how a discriminant function may be continually updated as more classifications are made.

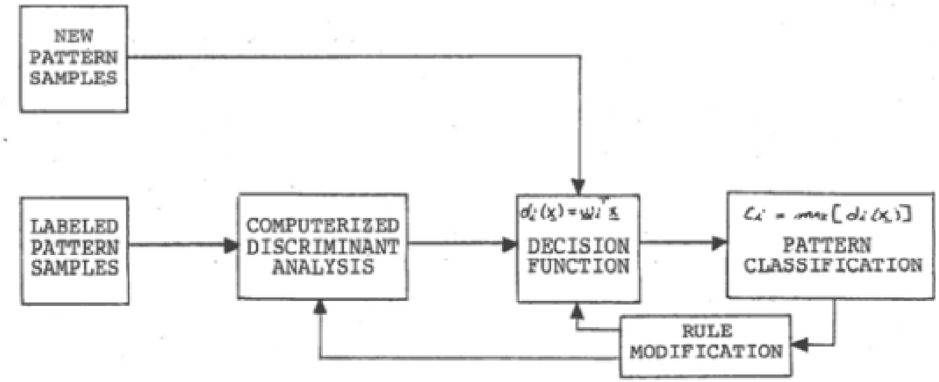

NEW PATTERN SAMPLES
LABELED PATTERN SAMPLES
COMPUTERIZED DISCRIMINANT ANALYSIS
$d_i(\underline{x}) = \underline{w}_i^T \underline{x}$
DECISION FUNCTION
$C_i = \max[d_i(\underline{x})]$
PATTERN CLASSIFICATION
RULE MODIFICATION

Figure 2.3: Discriminant Analysis Flowchart from [12]

In [13], a set of decision rules based on pattern recognition were established. This paper outlines arson pattern recognition systems such as the Fire Engineering Data Analysis Program, which was implemented alongside the Modular Information Reporting System in Prince Georges County, Maryland. Clusters of arson attacks were analyzed to develop patterns that could aid fire investigators through the discovery of previously unseen connections between fires. This research showed that geospatial and temporal data may lead to clues about an arsonists behavior.

2.3 Other Works of Interests

A more contemporary example of discriminant analysis may be seen in [25]. This research used a hybrid technique of Data Envelopment Analysis and Discriminant Analysis to analyze group membership for insurance underwriting. However, this paper focused solely on automobile insurance. The authors studied the correlation of several variables, such as age, sex, driving history. Based on their sample of 6,885 individuals, the authors reported a classification scheme with an overall misclassification error of 5.3%.

Chapter 3

Methods of Data Analysis

The focus of this thesis is to propose a model of arson prediction making use of only two sources of data, NFIRS fire data and RealtyTrac foreclosure data. The analogous feature of Baxi's work is the novel insurance score developed for the paper. The ability of the insurance score to predict arson prone structures could be used in predicting the cost and availability of insurance for similar structures. This thesis asserts that foreclosure can be used as a similar feature. One takeaway from [4] is that functions developed for various places, using different sets of variables and data, can vary widely in their effectiveness of classification. This realization confounds the idea that such functions can perform well over large areas such as the nation as a whole. This thesis will propose using national data that is already collected to address this issue. This chapter presents the methods and their backgrounds used for the analysis featured in this thesis. The initial dataset contained over 100 variables for more than one million fires. Pre-processing was necessary to extract a set of fires featuring the variables under consideration for this thesis. Following pre-processing, statistical tests, including some elements of exploratory data analysis (EDA), were conducted. After this initial work, a two-prong approach was applied to extract additional usefulness and meaning from the data. First, a classification scheme, Linear Discriminant Analysis (LDA), was implemented. Following LDA, a graph-based clustering technique, paraclique, was employed.

3.1 Data Pre-processing

Careful attention was given to the labeling and factoring of the NFIRS and RealtyTrac data presented in this thesis. This section details the initial steps taken to select features for study. This part of the analysis was conducted inside of IBM's Statistical Package for the Social Sciences (SPSS) Version 24 [11].

Starting with all the geocoded fire and foreclosure data from NFIRS and RealtyTrac for the years 2006 to 2011, several variables not under consideration for the study were removed. Examples of these removed variables include the street address, casualty and injury information, and building dimensions. To narrow the focus of this study, and better focus the results toward the link between arson and foreclosure, only 1 or 2 family residential structures were considered, property use code 419 in [31]. Fires with an *Exposure Number* greater than zero, meaning the fire was the result of another fire, were removed. Following this, a search was conducted for outliers. For example, some cases listed the *Story of Origin* as 999. As it seemed unlikely a typical 1 to 2 family residential structure will have even twelve, let alone 999 stories, these cases were removed. Issues such as this were likely due to miscoding and are attributed to human error. Additionally, outliers within the foreclosure recording dates were discovered. Some structures listed their Foreclosure Recording Date as January 1st, 1900. Fires in the datasets were limited to only those in structures that experienced foreclosure during the NFRIS sample period plus two years. Some variables were added based on existing variables, such as *Alarm Time* and *Days Between Fire and Foreclosure*. Values for these features were extracted and calculated from existing fields. Due to the nature of some types of the implemented analysis, some variables were simplified or replaced with dummy variables. Next, variables with too many missing cases were eliminated from consideration. An example of an eliminated variable is *First Factor Leading to Ignition.* Despite potentially holding some value in the determination of an intentional fire, many cases (greater than 50%) under study lacked values for this field. Only complete cases were selected for further analysis after EDA.

3.2 EDA and Initial Analysis

Before attempting the classification or clustering of fires, steps must be taken to ensure the features selected as classifiers are significant in their ability to discriminate intentional fires. This section outlines some of the tests and observations that were made in this determination of quality features. For categorical data, the initial step taken was to inspect the frequencies present in the data. Additionally at this step, the Pearson Chi-Square value for each variable was calculated with respect to *Foreclosure* and *Cause of Ignition*. For the *Foreclosure* and *Intentional* fire status, the odds ratio was calculated. Continuous variables, such as *Alarm Time*, were examined with histograms. This step of the analysis was conducted inside of SPSS 24.

3.2.1 Pearson Chi-Square & Fisher's Exact

From [11], Pearson chi-squared tests the independence of variables against a null hypothesis. The null hypothesis states that the variables are independent. From [21], chi-square tests whether the proportions of one variable differ for varying values of another variable. Observed counts for a variable value are compared to expected counts based on the portion seen in the sample. The important metric of Pearson chi-squared is the significance value (p-value). The p-value may be established by comparing the calculated chi-square value to the chi-square distribution for a given degrees of freedom. The Pearson chi-square may be calculated by Equation 3.1. One stumbling block is the use of expected counts.

In the case of categorical variables with many levels, these expected counts can lead to inaccuracies, if the number of expected counts less than five is greater than twenty percent of the number of expected counts. In the case where too many expected counts were less than five, the data was reduced into more general categories. This is easily accomplished due to the way that [31] encodes the data. For example, the *Area of Fire Origin* variable is coded into ten categories that each encompass several sub categories. For example, 02 is recorded for an exterior stairway while 03 is recorded for an interior stairway. All 0X codes are included in the "Means of Egress" category. More information about the variables from [31] may be found in Appendix B.

$$\chi^2 = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i} \tag{3.1}$$

where

$$\begin{aligned} N &= \text{Total Number of Observations} \\ n &= \text{Number of Categories, See Equation 3.3} \\ O_i &= \text{Number of Observations for Category } i \\ p_i &= \text{Portion of } N \text{ of Type } i \end{aligned}$$

Table 3.1: A 2 By 2 Table

	$\mathbf{Y_1}$	$\mathbf{Y_2}$
$\mathbf{X_1}$	a	b
$\mathbf{X_2}$	c	d

Fisher's exact test is similar but more powerful in practice due to its ability to calculate an exact significance value. This test can be particularly useful in cases where expected counts are unavailable and create problems for Pearson's chi-square. [20] speaks to the computational cost of computing significance values for tables larger than two by two. However, due to its exact nature and usefulness in small sample situations, Fisher's exact is widely used [21]. For a two by two table, such as Table 3.1, the p-value maybe calculated by Equation 3.2 where n is given by Equation 3.3.

$$p = \frac{\binom{a+b}{a}\binom{c}{c+d}}{\binom{n}{a+c}} \tag{3.2}$$

$$n = a + b + c + d \tag{3.3}$$

3.2.2 Odds Ratio & Relative Risk

The odds ratio may be used to measure the size of effect, or strength of association between two binary features. In the case of nondichotomous categorical variables, dummy variables may be created. For a two by two table such as Table 3.1 the odds ratio may be calculated by Equation 3.4 [26]. Similarly, the relative risk may be calculated by Equation 3.5 [17]. In the scope of health related fields, Table 3.1 may be seen as outcome versus risk factor. The rows of the table may be populated by the binary risk variable, whose values are X_1 and X_2, while the columns represent the binary outcome variable, whose outcomes are Y_1 and Y_2. For this thesis, foreclosure will be treated as a risk factor for an intentional fire outcome. In circumstances where the risk factor seldom occurs, the odds ratio approaches the relative risk [9]. This is known as the rare disease assumption.

$$OR = \frac{ad}{bc} \tag{3.4}$$

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \tag{3.5}$$

3.3 Discriminant Analysis

In [8], the classic example of discriminant analysis is found. In this seminal paper, Fisher classifies three species of iris based on four independent features. As seen in Section 2.2, there have been several attempts to apply discriminant analysis and other pattern recognition schemes to the intentional fire problem. This thesis presents discriminant analysis as a method for classifying fires from the NFIRS database based on select features. This approach is similar to [4]. However, by reducing the sources of the data, the analysis gains more practical feasibility. In agreement with [28], the most important features may not be the easiest to acquire. Typically, discriminant analysis requires that data conform to a normal distribution. Not all variables presented in this thesis follow a normal distribution. Baxi's work in [4] does not state whether or not he performed tests for normality. However, in [18], [16], and [23] there is evidence that even while violating the normality assumptions of discriminant analysis, meaningful classification may still be achieved. All of the discriminant analysis for this thesis was conducted inside SPSS 24.

3.4 Graph-based Clustering

Graph theoretical algorithms can be applied across many problem domains with impressive results. While it arose from the hypothesis testing for biological data, the graph-based paraclique algorithm has the potential for wide applications. This thesis applied the paraclique algorithm to the problem of clustering fires. Applying the paraclique algorithm entails a multi-step process, described in the following sections.

3.4.1 Constructing the Graphs

First, variables were selected that relate to the basic elements of the fire data: what, where, and how. For large sets of data, a random sample was taken. Each case in the dataset, or sample thereof, was run pairwise through a similarity measure. As mentioned in the original paraclique paper, a correlation metric can be used to create a similarity measure [6]. The fires in the dataset are identified by a unique ID; these IDs are used to name each vertex in the graph. Next, a threshold was applied to the similarity scores, and for each pair of fires with a similarity greater than or equal to the threshold an edge was placed in the graph between them. The paraclique algorithm was then run on the resulting graph. In graph theory, clique is a complete subset of a graph. The paraclique algorithm looks to increase the size of cliques that already exist within a graph by adding vertices to these cliques [22].

The similarity measure chosen for this thesis was Goodall. This measure was chosen for its ability to handle categorical data and scoring of infrequent events [5]. All variables for this section of the analysis were categorical. In the case of a constructed variable *Days Between Fire and Foreclosure*, the continuous values were discretized using the Freedman-Diaconis rule, resulting in 66 bins with a width of 76 days. Using datasets exported from SPSS 24, R scripts were written to apply Goodall to the fire cases and generate graphs.

3.4.2 Analyzing the Resulting Paracliques

The result of the paraclique algorithm was a list of grouped fires. These fires were then matched back to the cases from the dataset or sample using a script written in Python. The matched sets were then tested for enrichment for each variable value using a hypergeometric distribution. The hypergeometric test performed in this thesis is built into the R statistics package [24]. The variables for which each paraclique is enriched, as reported in Section 4.3, are then used for analysis.

To demonstrate the application and effectiveness of the paraclique algorithm, the iris data from [8] was processed as above, using Pearson correlation as the similarity metric since all variables are continuous. As displayed in Table 3.2, the paraclique algorithm output 4 paracliques. While not all of the 150 iris observations were placed in a paraclique, the algorithm successfully enriched each of the four paracliques for only one type of iris. Figure 3.1 displays the similarity of the iris features as related by a Pearson Correlation.

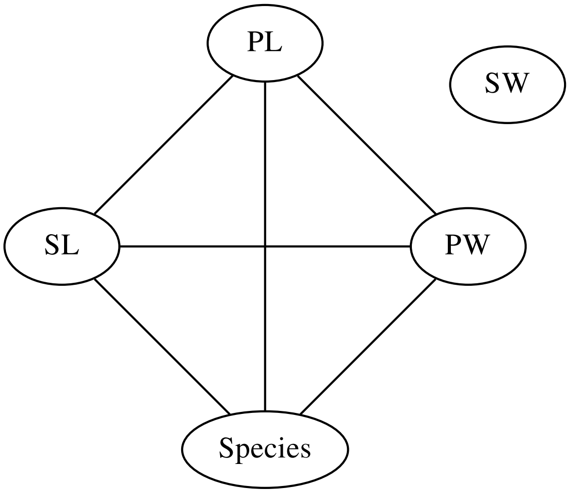


Figure 3.1: Iris Features as Related by Pearson Correlation

Table 3.2: Iris Data Paraclique Results

Paraclique Number	**Paraclique Size**	**Iris Type**	**p-value**
1	7	0	3.40E-04
2	5	1	4.25E-02
3	5	1	3.58E-03
4	5	2	3.58E-03

Chapter 4

Results of Applied Methods

In this chapter, we report and discuss results from the methods described in Chapter 3.

4.1 Basic Statistical Characteristics

Section 4.1 lays out many variables and values associated with the NFIRS and RealtyTrac data; this section reports numbers, proportions, odds ratios, p-values, and other statistical metrics for various features of interest. An enrichment p-value is provided as described in 3.4.2. This enrichment score for each table is relative to the set of data containing all fires. For more information on the features shown here, see Appendix B. For correlation tables, the exact significance is provided if available; asymptotic significance is provided otherwise. Fisher's exact is not available for all variables due to computational limitations. While many variables show a significant correlation, particularly in the national dataset, several are missing a high percentage of expected counts. The missing expected counts and expected counts less than one make these correlations unreliable. All histograms show the frequency percent. For frequency tables, the percent column indicates the percentage of all cases in the dataset, not the percentage of valid cases.

4.1.1 Area of Fire Origin

Table 4.1 displays the top five most common values for *Area of Fire Origin* in the the Saginaw and National data, respectively. The top three values for each dataset show relative agreement. However, the values in the fourth position seem to be at odds. In the Saginaw data, the fourth most common place for a residential fire was in a crawl space. Whereas in the national non-foreclosed dataset, the fourth most common location for a fire to occur is in the attic. This may be attributable to the metropolitan nature of Saginaw versus a higher portion of suburban homes in the country overall. It may be noted that for national foreclosed property fires the fourth most common area of origin is the garage. This value does not manifest in the top five of either other data selection studied. The Saginaw data is particularly enriched for fires in bedrooms, dens, and crawl spaces. The non-foreclosed national data is only enriched for fires in attics. The national foreclosed data is particularly enriched for bedrooms and kitchens.

Table 4.2 displays the Pearson chi-square value for *Area of Fire Origin* versus *Foreclosure* and *Intentional* fires. For each dataset, chi-square finds at least of five percent level of significance with each variable. However, when looking at the number of missing cases, a problem can be seen. One result lacks less than twenty percent of missing expected counts. As discussed in Section 3.2.1, this can have widely varying effects on the reliability of the chi-square value. In the case of *Foreclosure* in the Saginaw Data, Fisher's exact does not reveal a significant correlation. Table 4.2 was generated with the reduced *Area of Fire Origin* described in Section 3.2.1.

Table 4.1: Frequency of Area of Fire Origin

(a) Saginaw, MI

Number of Cases	1386		
Value	**Frequency**	**Percent**	**p-value**
Bedroom	240.0	15.5	1.31E-06
Cooking area, kitchen	237.0	15.3	9.66E-01
Common room, den, family room, living room, lounge	147.0	9.5	1.76E-08
Substructure area or space, crawl space	76.0	4.9	1.65E-08
Function area, other	72.0	4.7	6.99E-02

(b) National

Non-foreclosed			
Number of Cases	473380		
Value	**Frequency**	**Percent**	**p-value**
Cooking area, kitchen	89106	16.5	1.00E+00
Bedroom	60113	11.2	1.00E+00
Common room, den, family room, living room, lounge	31074	5.8	1.00E+00
Attic: vacant, crawl space above top story, cupola	28567	5.3	4.14E-07
Laundry area, wash house (laundry)	24253	4.5	5.62E-01
Foreclosed			
Number of Cases	31619		
Value	**Frequency**	**Percent**	**p-value**
Cooking area, kitchen	6634	19.5	4.95E-21
Bedroom	4903	14.4	5.03E-45
Common room, den, family room, living room, lounge	2290	6.7	2.01E-06
Vehicle storage area; garage, carport	1815	5.3	1.09E-10
Attic: vacant, crawl space above top story, cupola	1700	5.0	1.00E+00

Table 4.2: Correlation of Area of Fire Origin

Saginaw, MI				
Variable	**Cases**	**Chi-Square Sig.**	**Missing Exp. Cnt. (%)**	**Fisher's Sig.**
Foreclosure	1386	0.050	45	0.112
Intentional	1330	0.000	20	NA
National				
Foreclosure	506039	0.000	0	NA
Intentional	458896	0.000	0	NA

4.1.2 Building Status

Table 4.3 shows complete agreement between the datasets reviewed for the top three slots of their frequency tables for *Building Status*. However, reviewing the percentages shows that nearly a quarter of the fires in Saginaw occurred at Vacant and Unsecured properties. This frequency of occurrence is over four times the percentage of Vacant and Unsecured fire seen in the national non-foreclosed dataset. Saginaw fires were highly enriched for vacant and unsecured properties. National non-foreclosed fires are enriched to be in normal use. The national foreclosure fire data show enrichment for vacant and unsecured in addition to vacant and secured.

The correlation table for Building Status, Table 4.4, shows, once again, the perils of missing cases. The Saginaw data lack 50 percent and 43.8 percent missing expected count for foreclosure and intentional fires, respectively. For the more robust national data, a significant correlation is found.

While the correlations do not reveal much information or agreement, breaking the variable down into several components reveals a relationship. The odds ratio and relative risk are calculated for three levels of Building Status in Table 4.5, Table 4.6, and Table 4.7. Here, the levels of diminishing supervision reveal escalating levels of increased risk for an intentional fire for both datasets.

Table 4.3: Frequency of Building Status

(a) Saginaw, MI

Number of Cases	1527		
Value	**Frequency**	**Percent**	**p-value**
In normal use	934	60.4	1.00E+00
Vacant and unsecured	373	24.1	2.75E-117
Vacant and secured	158	10.2	2.05E-13
Idle, not routinely used	37	2.4	1.41E-03
Under major renovation	12	.8	9.04E-01

(b) National

Non-foreclosed			
Number of Cases	515516		
Value	**Frequency**	**Percent**	**p-value**
In normal use	432117	80.2	4.05E-161
Vacant and unsecured	30791	5.7	1.00E+00
Vacant and secured	27968	5.2	1.00E+00
Idle, not routinely used	7297	1.4	1.93E-01
Under major renovation	5495	1.0	1.00E+00
Foreclosed			
Number of Cases	32791		
Value	**Frequency**	**Percent**	**p-value**
In normal use	25515	74.9	1.00E+00
Vacant and unsecured	3007	8.8	8.11E-107
Vacant and secured	2712	8.0	2.81E-93
Under major renovation	484	1.4	2.42E-11
Idle, not routinely used	452	1.3	7.08E-01

Table 4.4: Correlation of Building Status

Saginaw, MI				
Variable	**Cases**	**Chi-Square Sig.**	**Missing Exp. Cnt. (%)**	**Fisher's Sig.**
Foreclosure	1527	0.25	50	0.082
Intentional	1465	0.000	43.8	NA
National				
Foreclosure	549358	0.000	0	NA
Intentional	483606	0.000	0	NA

Occupied & Operating

Table 4.5: Odds Ratio & Relative Risk for Occupied & Operating

Saginaw, MI				
		Intentional		
		Yes	**No**	**Total**
Occupied & Operating	**Yes**	68	819	887
	No	341	237	578
	Total	409	1056	1465
Odds Ratio	0.058		**Relative Risk**	0.130
National				
		Intentional		
		Yes	**No**	**Total**
Occupied & Operating	**Yes**	18334	391275	409609
	No	18912	55085	73997
	Total	37246	446360	483606
Odds Ratio	0.136		**Relative Risk**	0.175

Vacant & Secured

Table 4.6: Odds Ratio & Relative Risk for Vacant & Secured

Saginaw, MI				
		Intentional		
		Yes	**No**	**Total**
Vacant & Secured	**Yes**	75	74	149
	No	334	982	1316
	Total	409	1056	1465
Odds Ratio	2.980		**Relative Risk**	1.983
National				
		Intentional		
		Yes	**No**	**Total**
Vacant & Secured	**Yes**	6257	19049	25306
	No	30989	427311	458300
	Total	37246	446360	483606
Odds Ratio		4.529	**Relative Risk**	3.657

Vacant & Unsecured

Table 4.7: Odds Ratio & Relative Risk for Vacant & Unsecured

Saginaw, MI				
		Intentional		
		Yes	**No**	**Total**
Vacant & Unsecured	**Yes**	241	128	369
	No	168	928	1096
	Total	409	1056	1465
Odds Ratio	10.400		**Relative Risk**	4.261
National				
		Intentional		
		Yes	**No**	**Total**
Vacant & Unsecured	**Yes**	9551	17178	26729
	No	27695	429182	456877
	Total	37246	446360	483606
Odds Ratio		8.616	**Relative Risk**	5.895

4.1.3 Cause of Ignition

In Table 4.8, the promise of the Saginaw data reveals itself. Following with the findings of [14], the levels of reported intentional fires is much higher in the Saginaw dataset. This assessment is reinforced by Saginaw's enrichment for intentional fires. Both subsets of the national data list "Case under investigation" as their second value. This occurrence may be an instance where either it was not possible to determine the cause of the fire or the NFIRS records were not updated with current information when the investigation was completed. National non-foreclosed data displays enrichment for fires due to a failure of equipment and acts of nature. The foreclosure only data reveals enrichment for intentional and pending causes of ignition.

Correlations for *Cause of Ignition* were not included in this section as they are included as part of the correlations of other variables.

Table 4.8: Frequency of Cause of Ignition

(a) Saginaw, MI

Number of Cases	1483		
Value	**Frequency**	**Percent**	**p-value**
Unintentional	622	40.2	1.00E+00
Intentional	414	26.8	6.10E-121
Cause under investigation	347	22.4	3.36E-02
Failure of equipment or heat source	71	4.6	1.00E+00
Cause, other	15	1.0	8.88E-01

(b) National

Non-foreclosed			
Number of Cases	471187		
Value	**Frequency**	**Percent**	**p-value**
Unintentional	242133	44.9	9.47E-13
Cause under investigation	100205	18.6	1.00E+00
Failure of equipment or heat source	74957	13.9	3.68E-42
Intentional	34693	6.4	1.00E+00
Act of nature	12850	2.4	5.37E-57
Foreclosed			
Number of Cases	30405		
Value	**Frequency**	**Percent**	**p-value**
Unintentional	14969	43.9	1.00E+00
Cause under investigation	7155	21.0	1.38E-20
Failure of equipment or heat source	3954	11.6	1.00E+00
Intentional	3566	10.5	1.16E-149
Act of nature	420	1.2	1.00E+00

4.1.4 Fire Spread

Table 4.9 displays the frequency values for the extent of fire spread for all datasets. These results show that there is little consistency between the three datasets for the levels of *Fire Spread*. For the top position, foreclosed homes in the national data show that fires were slightly more often contained to the room rather than the building where the fire originated. The one point of consensus for the three datasets was on the least numerous outcome, where all three datasets show that the fewest number of fires went beyond their building of origin. Saginaw data conveys an enrichment for fires confined to their buildings and floors of origin. National non-foreclosure data is enriched for fires contained to the building and object of origin. The fires including foreclosure show an enrichment for confinement to the room and floor of origin.

Correlating *Foreclosure* with *Fire Spread* in Table 4.10 for the Saginaw data maintains foreclosure's elusive history. The result is not statistically significant and failed to yield a result for Fisher's exact. However, for intentional fires in Saginaw and both variables in the national dataset, there is a statistically significant correlation.

Table 4.9: Frequency of Fire Spread

(a) Saginaw, MI

Number of Cases	1542		
Value	**Frequency**	**Percent**	**p-value**
Confined to building of origin	701	45.3	2.43E-20
Confined to room of origin	433	28.0	9.98E-01
Confined to floor of origin	200	12.9	8.44E-08
Confined to object of origin	144	9.3	1.00E+00
Beyond building of origin	64	4.1	9.62E-01

(b) National

Non-foreclosed			
Number of Cases	524262		
Value	**Frequency**	**Percent**	**p-value**
Confined to building of origin	180498	33.5	4.16E-71
Confined to room of origin	162993	30.3	1.00E+00
Confined to object of origin	107369	19.9	2.66E-12
Confined to floor of origin	46321	8.6	1.00E+00
Beyond building of origin	27081	5.0	9.70E-28
Foreclosed			
Number of Cases	33427		
Value	**Frequency**	**Percent**	**p-value**
Confined to room of origin	12443	36.5	3.97E-118
Confined to building of origin	9973	29.3	1.00E+00
Confined to object of origin	6284	18.4	1.00E+00
Confined to floor of origin	3439	10.1	5.43E-19
Beyond building of origin	1288	3.8	1.00E+00

Table 4.10: Correlation of Fire Spread

Saginaw, MI				
Variable	**Cases**	**Chi-Square Sig.**	**Missing Exp. Cnt. (%)**	**Fisher's Sig.**
Foreclosure	15542	0.142	20	NA
Intentional	1479	0.000	0	NA
National				
Foreclosure	558779	0.000	0	NA
Intentional	490165	0.000	0	NA

4.1.5 First Item Ignited

Table 4.11 reveals a slight difference in the top position between the national foreclosed and non-foreclosed fires for *First Item Ignited*. Here, cooking materials gain 2.7 percentage points of frequency over non-foreclosed fires. Multiple items first ignited takes third place in the Saginaw dataset frequency table. As for the *Type of Material First Ignited*, Table 4.13 shows relative consistency around sawn wood being the top material for national non-foreclosed and Saginaw data, with fabric taking the top spot in national foreclosed fires by 2.6 percentage points. Trash fires and fires with multiple first items are enriched in the Saginaw data. The materials in this dataset show enrichment for sawn wood and fabric. National non-foreclosed fires bear an enrichment for framing and exterior wall covering. In this dataset, sawn wood appears especially enriched. The only enriched item category in the foreclosed national data is cooking material. These foreclosure inflicted property fires show enrichment for fabric and cooking oil.

In Table 4.12, all correlations show a high level of significance. However, even using a reduced number of levels for this variable resulted in an unacceptable number of missing expected counts in the Saginaw dataset. Additionally, the correlation with intentional fires for the Saginaw Data resulted in an expected count being less than 1, which also made the result dubious. The national data did not suffer this flaw and had a valid correlation for both variables. This statement also held true for Table 4.14. Tables 4.12 and 4.14 were generated with a reduced *First Item Ignited* and *Type of Material First Ignited* as described in Section 3.2.1.

Table 4.11: Frequency of First Item Ignited

(a) Saginaw, MI

Number of Cases	954		
Value	**Frequency**	**Percent**	**p-value**
Structural member or framing	122	7.9	1.25E-01
Cooking materials, including edible materials	83	5.4	7.95E-01
Multiple items first ignited	71	4.6	5.91E-15
Rubbish, trash, or waste	67	4.3	9.70E-13
Exterior wall covering or finish	65	4.2	5.03E-02

(b) National

Non-foreclosed			
Number of Cases	350648		
Value	**Frequency**	**Percent**	**p-value**
Structural member or framing	41012	7.6	4.57E-31
Cooking materials, including edible materials	32295	6.0	1.00E+00
Electrical wire, cable insulation	30059	5.6	6.86E-13
Exterior wall covering or finish	19742	3.7	3.63E-30
Item First Ignited, Other	17218	3.2	6.33E-04
Foreclosed			
Number of Cases	24406		
Value	**Frequency**	**Percent**	**p-value**
Cooking materials, including edible materials	3073	8.7	2.58E-63
Structural member or framing	2275	6.5	1.00E+00
Electrical wire, cable insulation	1779	5.1	1.00E+00
Item First Ignited, Other	1087	3.1	9.99E-01
Exterior wall covering or finish	973	2.8	1.00E+00

Table 4.12: Correlation of First Item Ignited

Saginaw, MI				
Variable	**Cases**	**Chi-Square Sig.**	**Missing Exp. Cnt. (%)**	**Fisher's Sig.**
Foreclosure	954	0.000	50	NA
Intentional	937	0.000	20	NA
National				
Foreclosure	375054	0.000	0	NA
Intentional	356707	0.000	0	NA

Table 4.13: Frequency of Type of Material First Ignited

(a) Saginaw, MI

Number of Cases	776		
Value	**Frequency**	**Percent**	**p-value**
Sawn wood, including all finished lumber	210	13.6	5.10E-06
Fabric, fiber, cotton, blends, rayon, wool	195	12.6	2.18E-07
Plastic	51	3.3	1.00E+00
Paper, including cellulose, waxed paper	43	2.8	2.96E-04
Multiple types of material	36	2.3	9.65E-01

(b) National

Non-foreclosed			
Number of Cases	292713		
Value	**Frequency**	**Percent**	**p-value**
Sawn wood, including all finished lumber	60581	11.2	1.10E-56
Fabric, fiber, cotton, blends, rayon, wool	51602	9.6	1.00E+00
Plastic	29699	5.5	5.77E-02
Multiple types of material	17635	3.3	1.00E+00
Type of material first ignited, other	16909	3.1	1.54E-05
Foreclosed			
Number of Cases	19548		
Value	**Frequency**	**Percent**	**p-value**
Fabric, fiber, cotton, blends, rayon, wool	4042	11.9	1.99E-26
Sawn wood, including all finished lumber	3159	9.3	1.00E+00
Plastic	1901	5.6	9.73E-01
Multiple types of material	1340	3.9	2.11E-06
Cooking oil, transformer or lubricating oil	1231	3.6	9.35E-25

Table 4.14: Correlation of Type of Material First Ignited

Saginaw, MI				
Variable	**Cases**	**Chi-Square Sig.**	**Missing Exp. Cnt. (%)**	**Fisher's Sig.**
Foreclosure	776	0.051	45	NA
Intentional	760	0.003	10	NA
National				
Foreclosure	312889	0.000	0	NA
Intentional	297037	0.000	0	NA

4.1.6 Foreclosure

Table 4.15 reveals the relatively low incidence of foreclosure in the Saginaw dataset. This level of incidence is especially unfortunate as it may harm *Foreclosure's* effectiveness as a classifying feature. Expanding the NFIRS data window to include a larger selection of years might increase the number of foreclosure cases here.

The national data reveals an increased risk of intentional fires in structures under foreclosure in Table 4.16. The sample size in the Saginaw data appears too low for such a relationship to be revealed, however, since only four intentional fires occurred in structures under foreclosure.

The creation of the *Days Between Fire and Foreclosure* variable shows a relationship between the temporal proximity of the foreclosure recording date to the fire. This relationship may be caused by the financial stress the foreclosure creates for the homeowner. This relationship is visible in both the Saginaw data, Figure 4.1a, and the national data, Figure 4.1b. For *Days Between Fire and Foreclosure*, the result is negative if the fire occurred before the foreclosure.

Table 4.15: Frequency of Foreclosure

(a) Saginaw, MI

Number of Cases	1547	
Value	**Frequency**	**Percent**
Yes	39	2.5
No	1508	97.5

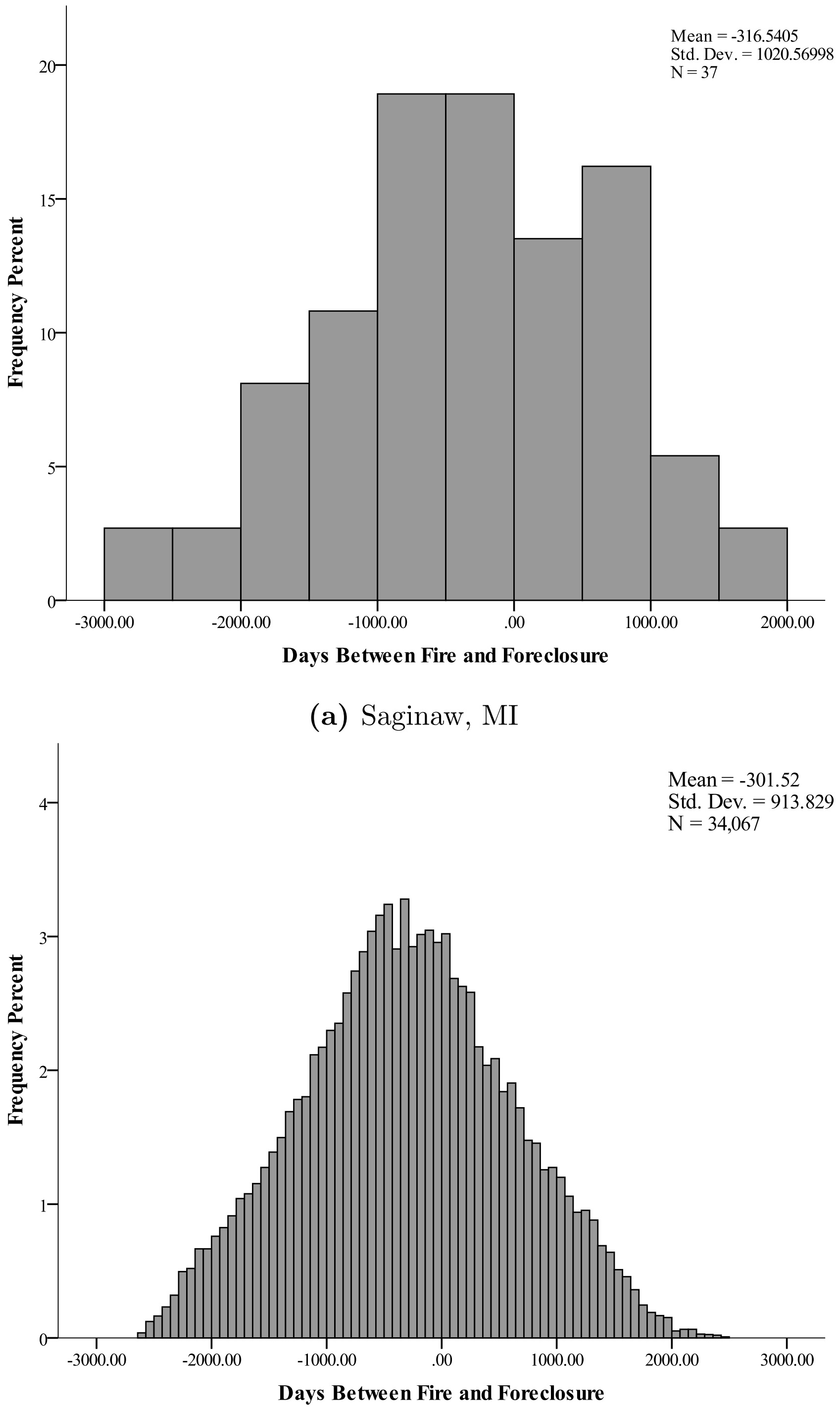
(b) National

Number of Cases	573981	
Value	**Frequency**	**Percent**
Yes	35191	6.1
No	538790	93.9

Table 4.16: Odds Ratio & Relative Risk for Foreclosure

Saginaw, MI				
		Intentional		
		Yes	**No**	**Total**
Foreclosure	**Yes**	4	35	39
	No	410	1034	1444
	Total	414	1069	1483
Odds Ratio	0.288		**Relative Risk**	0.361
National				
		Intentional		
		Yes	**No**	**Total**
Foreclosure	**Yes**	3647	27763	31410
	No	34693	436494	471187
	Total	38340	464257	502597
Odds Ratio	1.653		**Relative Risk**	1.577

Temporal Proximity Between Fire and Foreclosure



(a) Saginaw, MI

(b) National Data

Figure 4.1: Days Between Fire and Foreclosure

4.1.7 Heat Source

Table 4.17 demonstrates relative agreement for national non-foreclosed versus foreclosed. For Saginaw, radiated heat takes the top position. Other and hot ember sources of heat are enriched in this dataset. For national non-foreclosed data, arcing and hot embers are enriched. In the foreclosure based data set, heat from an open flame shows particular enrichment. Additionally, it may be noted that the percentages between the Saginaw and national data do not show relative agreement for the top three positions. Table 4.18 was generated with the reduced *Heat Source* described in Section 3.2.1.

Table 4.17: Frequency of Heat Source

(a) Saginaw, MI

Number of Cases	782		
Value	**Frequency**	**Percent**	**p-value**
Radiated, conducted heat from operating equipment	105	6.8	8.42E-01
Arcing	91	5.9	1.00E+00
Hot ember or ash	70	4.5	9.53E-03
Heat source: other	69	4.5	9.75E-03
Heat from powered equipment, other	61	3.9	1.00E+00

(b) National

Non-foreclosed			
Number of Cases	349858		
Value	**Frequency**	**Percent**	**p-value**
Arcing	62986	11.7	8.38E-24
Radiated, conducted heat from operating equipment	51147	9.5	7.24E-01
Heat from powered equipment, other	45667	8.5	1.00E+00
Hot or smoldering object, other	25118	4.7	2.46E-02
Hot ember or ash	23881	4.4	1.29E-28
Foreclosed			
Number of Cases	22748		
Value	**Frequency**	**Percent**	**p-value**
Arcing	3506	10.3	1.00E+00
Radiated, conducted heat from operating equipment	3370	9.9	2.08E-01
Heat from powered equipment, other	3231	9.5	5.07E-07
Hot or smoldering object, other	1558	4.6	9.70E-01
Heat from other open flame or smoking materials	1495	4.4	1.56E-65

Table 4.18: Correlation of Heat Source

Saginaw, MI				
Variable	**Cases**	**Chi-Square Sig.**	**Missing Exp. Cnt. (%)**	**Fisher's Sig.**
Foreclosure	782	0.273	37.5	NA
Intentional	776	0.000	12.5	NA
National				
Foreclosure	373390	0.000	0	NA
Intentional	363970	0.000	0	NA

4.1.8 Story of Origin

Table 4.19 reports the frequencies of the floor where fires originated. For this variable, many cases recorded a value of 0 for the *Story of Origin*. These cases were treated as missing due to the inconclusive nature of using 0. It could not be verified if 0 referred to the value missing or if 0 indicated the ground floor. In the case of the Saginaw data, very few cases were recorded as 0 where the national data contained tens of thousands of 0s. Negative values indicate the floor was under grade. Fires in Saginaw, MI only showed enrichment for the basement level. Residential properties in the national non-foreclosed data only showed enrichment for the first floor. In the properties that had experienced foreclosure, the basement and second floor showed significant enrichment.

Table 4.20 reveals, once again, a lack of expected counts for the Saginaw dataset, leaving inconclusive correlations with *Foreclosure* and *Intentional*. The national data yielded a significant correlation with both variables.

Table 4.19: Frequency of Story of Origin

(a) Saginaw, MI

Number of Cases	1534		
Value	**Frequency**	**Percent**	**p-value**
1	1237	80.0	1.00E+00
2	148	9.6	6.92E-01
-1	145	9.4	2.58E-13
3	4	.3	9.98E-01

(b) National

Non-foreclosed			
Number of Cases	507540		
Value	**Frequency**	**Percent**	**p-value**
1	428298	79.5	5.66E-30
2	50122	9.3	1.00E+00
-1	25110	4.7	9.74E-01
3	4010	.7	8.37E-01
Foreclosed			
Number of Cases	32735		
Value	**Frequency**	**Percent**	**p-value**
1	26823	78.7	1.00E+00
2	3938	11.6	6.91E-35
-1	1696	5.0	3.15E-02
3	278	.8	1.26E-01

Table 4.20: Correlation of Story of Origin

Saginaw, MI				
Variable	**Cases**	**Chi-Square Sig.**	**Missing Exp. Cnt. (%)**	**Fisher's Sig.**
Foreclosure	1543	0.925	50	NA
Intentional	1472	0.000	25	NA
National				
Foreclosure	541331	0.000	0	NA
Intentional	475258	0.000	0	NA

4.2 Discriminant Analysis

This section contains the function coefficients associated with the discriminant function. Section 4.2 displays the results of applying LDA to the Saginaw, MI data. Taking the Saginaw data to be the cleanest, best available for this thesis, LDA function coefficients were generated to classify the Saginaw fires as either intentional or unintentional. The features used to discriminate can be seen in Table 4.21. These function coefficients were then applied to the national data. Tables 4.22 and 4.23 show how well the function with the Saginaw coefficients was able to classify both sets of data. The differences in frequency and enrichment, as revealed in 4.1, point to the overall difficulty of using a single city to train a classifier on a national level.

In Section 4.2.2, an attempt was made to classify fires into foreclosed or not foreclosed. This attempt can be seen as a test of strength for foreclosure itself as a significant feature. Unfortunately, the classification only achieved an accuracy of 56.2% with the national data based on the Saginaw coefficients. This lack of classifying power may be due to the relatively small number of foreclosures available in the Saginaw dataset. Tables 4.24, 4.25, and 4.26 report the results of this classification.

4.2.1 Classifying for Intentional Fires

Table 4.21: Classification Coefficients From Saginaw Data for Intentional Fires

Fisher Coefficients		
	Intentional	
Variable	False	True
Fire Origin	1.708	1.386
Area of Fire Origin	0.060	0.058
Fire Spread	1.907	2.021
Item First Ignited	0.119	0.133
Foreclosure	2.520	1.707
Heat Source	0.025	0.066
Building Status	0.997	2.146
Type of Material First Ignited	0.158	0.161
(Constant)	-13.176	-19.508

Table 4.22: Intentional Fire Classification of Saginaw Data with Saginaw Coefficients

	Predicted		
Intentional	**FALSE**	**TRUE**	**Total**
FALSE	377	57	434
TRUE	35	67	102
Percent correct	82.8%		

Table 4.23: Intentional Fire Classification of National Data with Saginaw Coefficients

	Predicted		
Foreclosure	**FALSE**	**TRUE**	**Total**
FALSE	214031	14521	228552
TRUE	9859	5734	15593
Percent correct	90.0%		

4.2.2 Classifying for Foreclosure

Table 4.24: Classification Coefficients From Saginaw Data for Foreclosure

Fisher Coefficients		
	Intentional	
Variable	**False**	**True**
Fire Origin	1.823	2.059
Area of Fire Origin	0.058	0.067
Cause of Ignition	4.238	4.258
Fire Spread	1.701	1.416
Item First Ignited	0.129	0.133
Heat Source	0.027	0.044
Building Status	1.185	0.847
Type of Material First Ignited	0.169	0.144
(Constant)	-18.258	-16.934

Table 4.25: Foreclosure Classification of Saginaw Data with Saginaw Coefficients

	Predicted		
Foreclosure	**FALSE**	**TRUE**	**Total**
FALSE	340	177	517
TRUE	6	13	19
Percent correct	65.9%		

Table 4.26: Foreclosure Classification of National Data with Saginaw Coefficients

	Predicted		
Foreclosure	**FALSE**	**TRUE**	**Total**
FALSE	130326	98119	228445
TRUE	8724	6976	15700
Percent correct	56.2%		

4.3 Paracliques

This section presents the relevant statistics associated with the graph analysis. The analysis here follows the same structure of Section 3.4. First, the variables are compared to each other. This comparison is implemented with a Goodall measure as described in Section 3.4.1, the results are shown in Table 4.27. Figure 4.2 displays this information in a graph. Next, enrichment analysis was performed on the paraclique to determine which, if any, variable values are overrepresented.

Table 4.27: Feature Similarity in the National Foreclosure Data

	Area of Orig.	**Cause of Ign.**	**1^{st} Itm. Ign.**	**Heat Src.**	**Type of Mat. 1^{st} Ign.**
Area of Orig.	0.00	0.00	0.94	0.46	0.60
Cause of Ign.	0.00	0.00	0.21	0.33	0.00
1^{st} Itm. Ign.	0.94	0.21	0.00	0.58	1.28
Heat Src.	0.46	0.33	0.58	0.00	0.39
Type of Mat. 1^{st} Ign.	0.60	0.00	1.28	0.39	0.00

Table 4.27 displays the similarity of select features using a Goodall measure.

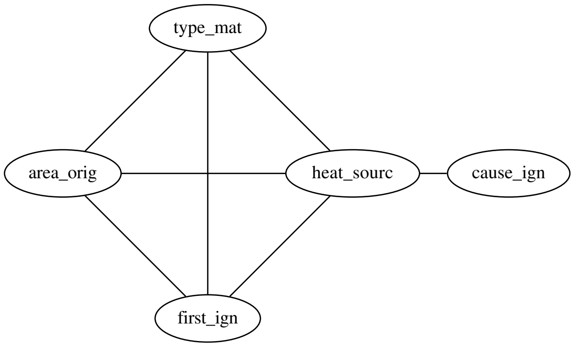


Figure 4.2: Graph Based on Table 4.27

The paracliques presented in Section 4.3 show enrichment for several interesting scenarios. In both datasets, the largest paraclique discovered was enriched for cooking related fires in kitchens. This example demonstrates the power of the algorithm and the implementation to help construct scenarios, even if in this case such a scenario is evident[1] even before paraclique analysis. The value added here is that individual feature values associated with those fires are also enriched. In the case of the cooking fires, paracliques revealed three related values for *Type of Material First Ignited*: cooking oil, fat, and starch. Looking at fires in this way takes a reviewer a step further than frequency tables and correlation scores.

The subsections here are broken down into the datasets from which the paracliques were extracted. Section 4.3.1 displays two paracliques from the national fire dataset. Of the many paracliques generated for this dataset, these were the only two enriched for *Foreclosure*. These paracliques show that fires in foreclosed properties may have some association with the other variable values discovered. Section 4.3.2 reports two paracliques from a subset of the national data with only residential structures that experienced a foreclosure. By looking at paracliques for which different bins of *Days Between Fire and Foreclosure* are enriched, particular points in the cycle of financial distress may be studied. The paracliques shown here were chosen due to the close temporal proximity of the fire and foreclosure recording date. Section 4.3.3 details two paracliques from the Saginaw, MI dataset. Since the Saginaw dataset contains what is considered for this thesis to be clean and reliable data, these paracliques can reveal particularly potent relationships. The first paraclique in this section was chosen as it was the only paraclique to be enriched for *Foreclosure*. The second paraclique presented here was the largest of the paracliques found to be enriched for intentional fires.

[1] The kitchen was the top value for *Area of Fire Origin* in the national data.

4.3.1 National Data

Table 4.28: Paracliques from National Data Enriched for Foreclosure

(a) Paraclique 6

Clique Size		23
Variable	**Value**	**p-value**
Area of Origin	Kitchen	5.96E-17
Cause of Ignition	Unintentional	1.67E-05
Item First Ignited	Cooking Material	6.57E-26
Foreclosure	Yes	1.23E-02
Heat Source	Heat from Operating Equipment	1.09E-03
Heat Source	Heat from Direct Flame	4.26E-02
Type of Material First Ignited	Cooking Oil, Other Oil	5.01E-31

(b) Paraclique 49

Clique Size		6
Variable	**Value**	**p-value**
Area of Origin	Laundry Area	1.72E-08
Fire Spread	Confined to Room of Origin	3.64E-03
Item First Ignited	Dust, Fiber, Lint	8.64E-08
Foreclosure	Yes	5.00E-02
Heat Source	Heat from Operating Equipment	4.96E-02
Heat Source	Spark from Friction	3.26E-02
Type of Material First Ignited	Fabric, Finished Goods	3.02E-05

Table 4.28a shows the enrichment results for a 23 fire paraclique from the national dataset. This paraclique is enriched for unintentional fires originating in the kitchen. Also enriched is an oil based cooking material ignited by radiated heat or heat from a flame. This paraclique was enriched for foreclosure.

Table 4.28b reports the enrichment results for a 6 fire paraclique. This paraclique was enriched for fires originating in a laundry area caused by fabric or dust ignited by radiated heat or spark from friction. The fire spread variable was enriched for fires that were contained in the room the fire originated. This paraclique was also enriched for foreclosure.

Table 4.29: Paracliques from National Data Enriched for Intentional Fires

(a) Paraclique 28

Clique Size		9
Variable	**Value**	**p-value**
Area of Origin	Bedroom	1.33E-08
Cause of Ignition	Intentional	1.96E-11
Item First Ignited	Bedding	3.24E-09
Heat Source	Lighter	7.68E-15
Type of Material First Ignited	Fiber, Finished Goods	1.61E-07

(b) Paraclique 50

Clique Size		6
Variable	**Value**	**p-value**
Area of Origin	Bedroom (More Than 5 People)	2.08E-02
Area of Origin	Closet	2.74E-03
Cause of Ignition	Intentional	7.77E-08
Fire Spread	Confined to Room of Origin	3.64E-03
Item First Ignited	Wearing Apparel Not on a Person	1.77E-09
Heat Source	Incendiary Device	5.77E-04
Heat Source	Heat from Open Flame, Other	2.08E-02
Type of Material First Ignited	Fabric, Finished Goods	3.02E-05

Table 4.29 contains two paracliques enriched for intentional fires. Paraclique 28 is enriched for fires starting in bedrooms. These fires were enriched to be bedding ignited by a lighter.

Paraclique 50 is enriched for fires starting in closets and bedrooms for more than five people that were confined to the room of origin. These fires are enriched for clothing that was ignited by an unspecified open flame or incendiary device.

4.3.2 National Data with Only Foreclosures Selected

Table 4.30 details the two paracliques from the national foreclosure data selected for analysis. Paraclique 18 contains 13 fires enriched for unintentional fires in attics. The first ignited item was enriched for plastic electrical wire ignited by arcing. These fires were enriched to take place ten days before to 66 days after foreclosure.

Paraclique 30 contains eight fires enriched for intentional fires with *Area of Fire Origin* as storage. The *First Item Ignited* was enriched for flammable liquid and rolled paper or fabric ignited by heat from an open flame. These fires were enriched for a fire spread indicating that the fire was confined to the building of origin. Additionally, these fires were enriched to occur ten days before to 66 days after foreclosure.

Table 4.30: Paracliques from National Data with a Foreclosure

(a) Paraclique 18

Clique Size		13
Variable	**Value**	**p-value**
Area of Origin	Attic	4.84E-02
Days Between Fire and Foreclosure	Fire 10 Days Before to 66 Days After Foreclosure	5.13E-03
Cause of Ignition	Unintentional	2.43E-03
Item First Ignited	Electrical Wire	4.25E-18
Heat Source	Arcing	2.97E-11
Type of Material First Ignited	Plastic	8.52E-14

(b) Paraclique 30

Clique Size		8
Variable	**Value**	**p-value**
Area of Origin	Storage Area, Other	1.63E-03
Days Between Fire and Foreclosure	Fire 10 Days Before to 66 Days After Foreclosure	2.00E-02
Cause of Ignition	Intentional	1.21E-08
Fire Spread	Confined to Building of Origin	6.72E-06
Item First Ignited	Rolled, Wound Material	2.38E-02
Item First Ignited	Flammable Liquid/Gas	1.56E-02
Heat Source	Heat from Open Flame, Other	5.06E-10
Building Status	Vacant and Secured	9.45E-11
Type of Material First Ignited	Flammable or Combustible Liquid, Other	8.70E-03

4.3.3 Saginaw Data

Table 4.31 outlines two paracliques selected from the Saginaw dataset. Paraclique 1 contains 36 fires enriched for unintentional fires in kitchens. The ignited material of these fires is enriched for cooking oil and fat-based cooking material ignited by conducted heat, heat from a direct flame, or an unknown other heat source. These fires were enriched to be contained to the room of origin. The structures were enriched to be occupied and operating. The fires are enriched to start on the first floor.

Paraclique 4 contains 11 fires enriched for intentional fires originating in multiple areas or entrance ways. The structures where these fires occurred was enriched to be vacant and unsecured. The fires were enriched to be from paper rubbish ignited by an open flame or an undetermined smoking material. The fires were enriched to be contained to the building of origin.

Table 4.31: Paracliques from Saginaw Data

(a) Paraclique 1

Clique Size		36
Variable	**Value**	**p-value**
Area of Origin	Kitchen	2.32E-32
Cause of Ignition	Unintentional	1.57E-14
Fire Origin	First Floor	1.92E-04
Fire Spread	Confined to Room of Origin	1.69E-20
Item First Ignited	Cooking Material	1.98E-56
Foreclosure	Yes	1.44E-02
Heat Source	Other	3.45E-02
Heat Source	Heat from Operating Equipment	2.54E-08
Heat Source	Heat from Direct Flame	2.38E-04
Building Status	In Normal Use	1.68E-08
Type of Material First Ignited	Ether, Pentane-type Flammable Liquid	1.78E-26
Type of Material First Ignited	Fat, Grease	2.08E-15

(b) Paraclique 4

Clique Size		11
Variable	**Value**	**p-value**
Area of Origin	Entranceway	4.75E-02
Area of Origin	Multiple Areas	1.90E-04
Cause of Ignition	Intentional	6.43E-07
Fire Spread	Confined to Building of Origin	1.08E-04
Item First Ignited	Rubbish, Trash	1.63E-14
Heat Source	Heat from Open Flame, Other	3.80E-08
Heat Source	Undetermined Smoking Material	2.61E-05
Building Status	Vacant and Unsecured	1.01E-07
Type of Material First Ignited	Paper	2.13E-14

Chapter 5

Conclusion

This thesis explored a portion of the data available from NFIRS and attempted to expose meaningful relationships between several fire variables. The ultimate takeaways from work presented here are the link between foreclosure and fire, the link of temporal proximity on foreclosure and fire, and the benefits of using graph theoretical algorithms to extract relationships from fire data.

In this thesis, the historical methods and applications were discussed with an exploration of the iris data from Ronald Fisher's 1936 paper and Baxi's application of Fisher's LDA to Newark fire data. These applications of discriminant analysis are discussed, and the method was subsequently applied to the NFIRS RealtyTrac-augmented data. Following this analysis, a more contemporary type of analysis was applied to the NFIRS datasets in the form of the graph theoretical paraclique algorithm. Paraclique was applied to the iris data before being used to cluster the fire data. In this way, a common thread of analysis was provided. Seeing the paracliques form around the iris species hints at the usefulness of the paraclique method. This usefulness is further borne out when applied to the fire data where feature rich scenarios are readily extracted. The advantage of the paraclique method as compared to looking strictly at correlations and frequencies is the level of ease with which the analysis is performed. Chi-square is vulnerable to missing counts and incomplete cases. Fisher's test does not scale well with larger numbers of feature levels. The paraclique analysis presented here overcame these stumbling blocks and allowed the data to tell its own stories.

5.1 Reducing the Impact of Foreclosure

A paper [15] made similar findings in agreement with [7]. The authors found there was a statistically significant link between foreclosure and violent crime. However, they were not able to show a statistically significant link between foreclosure and property crimes. The author of [7] notes though that property crime (the classification assigned to arson) is very under reported. This finding reinforces those of [1] and [14] that arson is a slippery target for investigators.

One way to reduce the impact of foreclosure may be to reduce the amount of foreclosure. Alternatives to foreclosure have been in explored in [2]. This paper suggested that by viewing mortgage lenders as agents of profit, seeking to maximize the amounts of money to be extracted from a property, a case for pursuing non-foreclosure related outcomes can be made financially sound. However, the authors noted that this is most effective for large borrowing institutions that are capable of spreading the risk over a large number of borrowers. Even still, by pursuing options that keep borrowers in their homes, such as loan modifications, the asset can remain profitable for the lender.

5.2 Future Work

Following this research, other areas of pattern recognition could be applied, such as neural networks. Logistic regression was considered as a replacement for discriminant analysis but was rejected due to the historical nature of discriminant analysis in connection with arson prone structures. Logistic regression does not suffer from the same constraint assumptions as discriminant analysis, as noted in [23]. The analysis presented here focused on classifying fires into either intentional or unintentional based on a training set with the same categories. However, as shown in [14], the effectiveness of arson detection and investigation may be limited. Using the findings presented there it could be beneficial to construct a profile of the risk factors that increase a fire to the level of suspicious. Fires categorized as suspicious could then be used to train a new decision function. These simple classifications, arson and nonarson and foreclosure and non-foreclosure, can be the building blocks of new methods for insurance underwriting.

Additional development of the graph techniques may also be useful. Applying alternate methods of clustering features or combining features into more robust factors may yield a greater number of high-quality paracliques. An alternative similarity scoring system could also be used to find different links between features and fires.

Bibliography

[1] Akiyama, Y. and Pfeiffer, P. C. (1984). Arson: A statistical profile. *FBI Law Enforcement Bulletin*, 53(10):8–14. 10, 49

[2] Ambrose, B. W. and Capone, C. A. (1996). Cost-Benefit Analysis of Single-Family Foreclosure Alternatives. *Journal of Real Estate Finance and Economics*, 13:105–120. 49

[3] ATTOM Data Solutions (2017). Transforming the Future of Property Data. 1

[4] Baxi, H. (1984). Use of discriminant analysis to predict arson-prone structures. *Fire Technology*, 20(4). x, 1, 4, 7, 8, 9, 13, 17

[5] Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity Measures for Categorical Data: A Comparative Evaluation. *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 243–254. 18

[6] Chesler, E. J. and Langston, M. A. (2006). Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data. *Systems Biology and Regulatory Genomics*, pages 150–165. 18

[7] Ellen, I. G., Lacoe, J., and Sharygin, C. A. (2013). Do foreclosures cause crime? *Journal of Urban Economics*. 4, 5, 49

[8] Fisher, R. A. (1936). THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*, 7(2):179–188. 3, 17, 19

[9] GREENLAND, S. and THOMAS, D. C. (1982). ON THE NEED FOR THE RARE DISEASE ASSUMPTION IN CASE-CONTROL STUDIES. *American Journal of Epidemiology*, 116(3):547–553. 17

[10] Hagan, R. D., Langston, M. A., and Wang, K. (2016). Lower bounds on paraclique density. *Discrete Applied Mathematics*. 2

[11] IBM (2016). SPSS Statistics Version 24. 14, 15

[12] Icove, D. J. (1979). *Principles of Incendiary Crime Analysis*. PhD thesis, The University of Tennessee, Knoxville, TN. x, 3, 11

[13] Icove, D. J. and Crisman, H. J. (1975). Application of pattern recognition in arson investigation. *Fire Technology*, 11(1):35–41. 1, 11

[14] Icove, D. J. and Hargrove, T. K. (2014). Project Arson: Unconvering the True Arson Rate in the United States. In National Association of Fire Investigators, I., editor, *International Symposium on Fire Investigation Science and Technology*. 1, 27, 49

[15] Immergluck, D. and Smith, G. (2006). The Impact of Single-family Mortgage Foreclosures on Neighborhood Crime. *Housing Studies*. 49

[16] Klecka, W. R. (1980). Discriminant Analysis. In *Quantitative Applications in the Social Sciences*. SAGE Publications, Inc. 17

[17] LaMorte, W. W. (2016). Risk Ratios and Rate Ratios (Relative Risk). 17

[18] Li, T., Zhu, S., and Ogihara, M. (2006). Using discriminant analysis for multi-class classification: An experimental investigation. *Knowledge and Information Systems*. 17

[19] Logue, F. (1980). Municipal Anti-Arson Strategy - The New Haven Model. *Fire Journal*, 74(2). 5

[20] Looney, S. W. and Hagan, J. L. (2015). *Analysis of Biomarker Data: A Practical Guide*. Wiley. 16

[21] McDonald, J. H. (2014). *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, Maryland, 3rd edition. 15, 16

[22] Phillips, C. A. (2015). Multipartite Graph Algorithms for the Analysis of Heterogeneous Data. *PhD diss.* 18

[23] Pohar, M., Blas, M., and Turk, S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. *Metodološki zvezki*, 1(1):143–161. 17, 49

[24] R Core Team (2017). R: A Language and Environment for Statistical Computing. 19

[25] Retzlaff-Roberts, D. and Puelz, R. (1996). Classification in automobile insurance using a DEA and discriminant analysis hybrid. *Journal of Productivity Analysis*, 7(4):417–427. 12

[26] Szumilas, M. (2010). Explaining Odds Ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3):227229. 17

[27] The City of New York Arson Strike Force (1980). Predicting Arson in New York City. 5, 9

[28] Tou, J. T. and Gonzalez, R. C. (1974). *Pattern Recognition Principles*. Addison-Wesley Publishing Company, reading, ma edition. 17

[29] Urban Educational Systems, I. (1980). The Research Manual: A Manual of Property Research and Arson Analysis. 9

[30] U.S. Fire Administration (2017). About the National Fire Incident Reporting System. 1

[31] U.S. Fire Administration National Fire Data Center (2010). Complete Reference Guide National Fire Incident Reporting System. 14, 15, 65

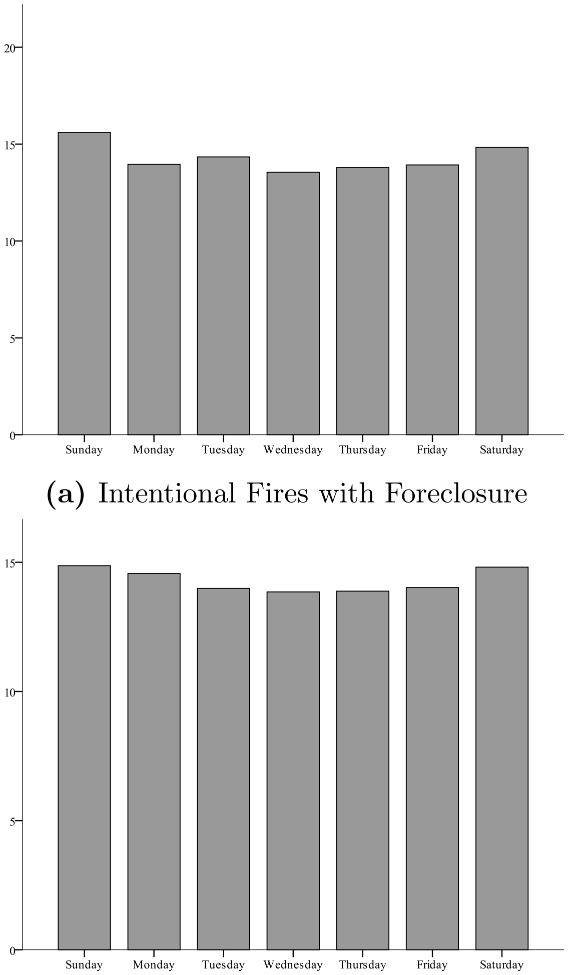
Appendix

A Additional Results

This appendix contains results that were not included in Chapter 4.

Weekday of Fire

Reviewing the temporal repercussions of the day of the week did not reveal a significant relationship. Figures A.1, A.2, and A.3 display this relationship.



(a) Intentional Fires with Foreclosure

(b) Unintentional Fires with No Foreclosure

Figure A.1: National Fire Days of the Week

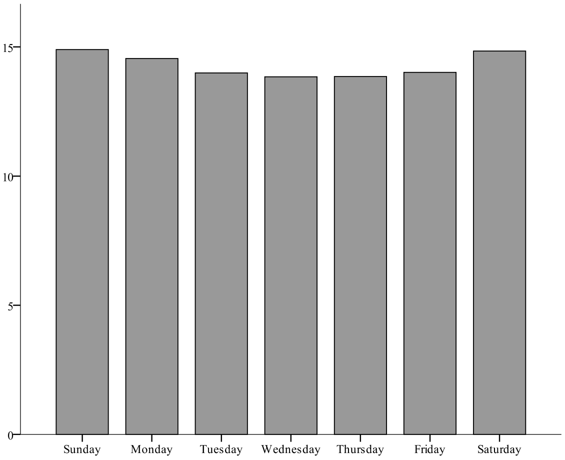


Figure A.2: All National Fires Days of the Week

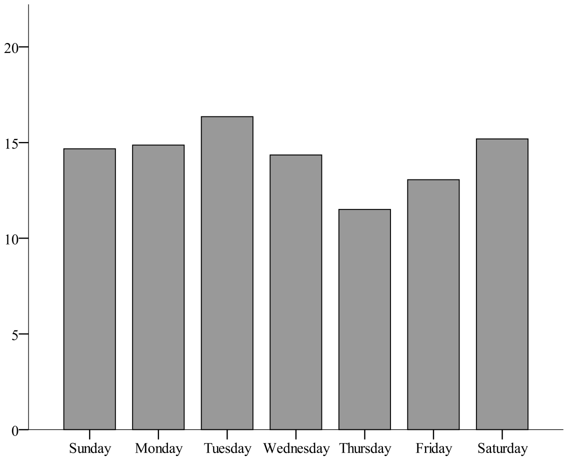


Figure A.3: All Saginaw, MI Fires Days of the Week

Alarm Time

Reviewing the temporal repercussions of the day of the week did not reveal a significant relationship. Figures A.5, A.4, and A.6 display this relationship.

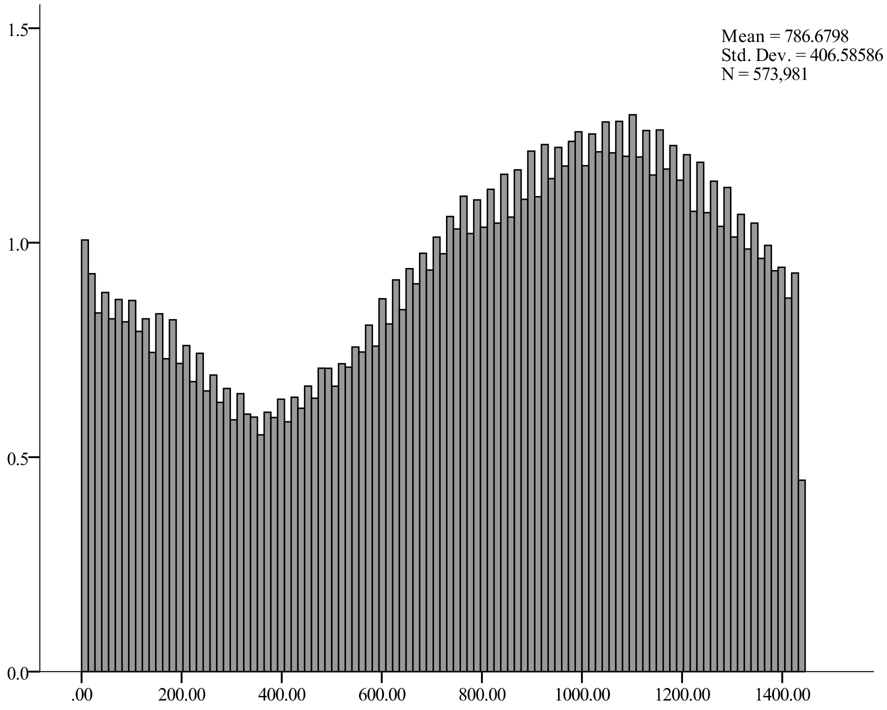
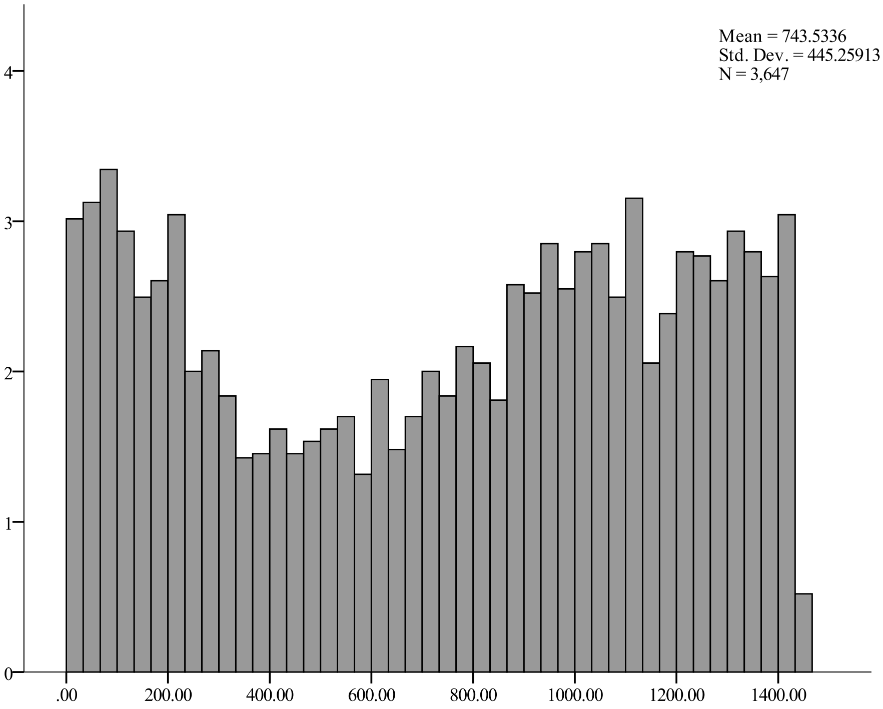
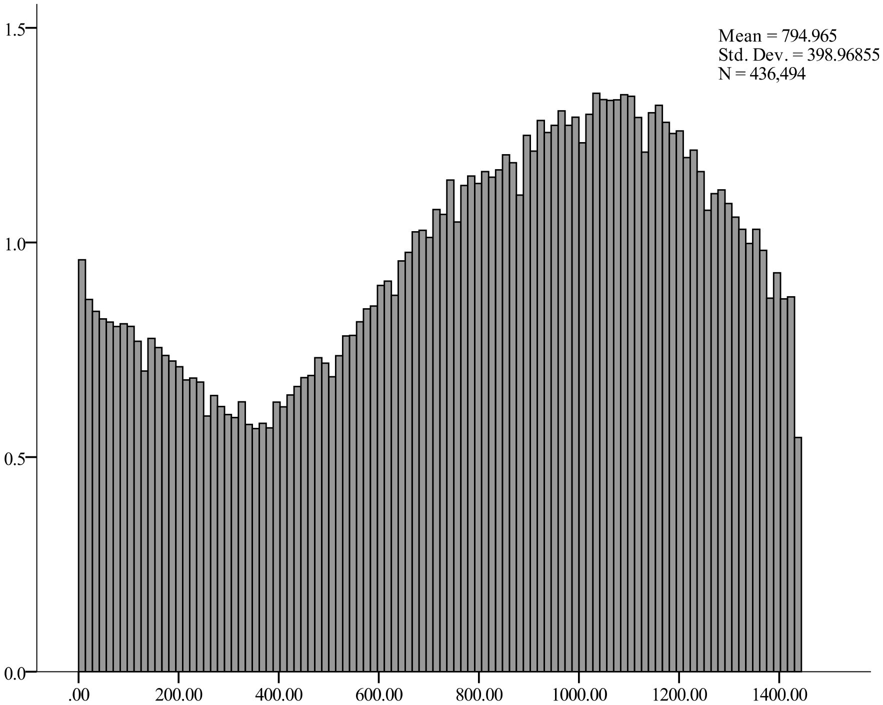


Figure A.4: National All Fires Alarm Time



(a) Intentional Fires with Foreclosure



(b) Unintentional Fires with No Foreclosure

Figure A.5: National Fires Alarm Time

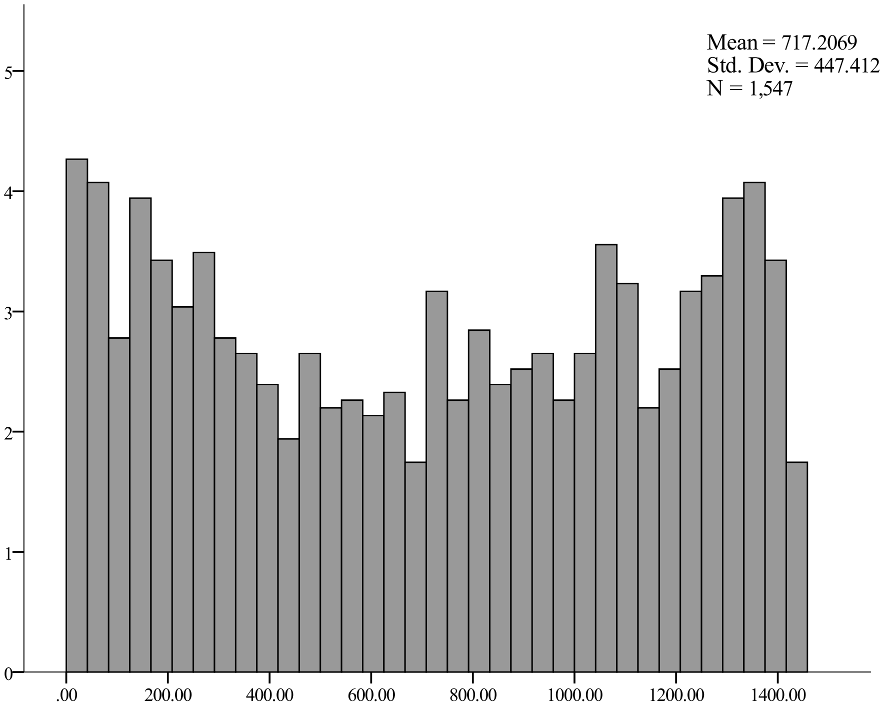
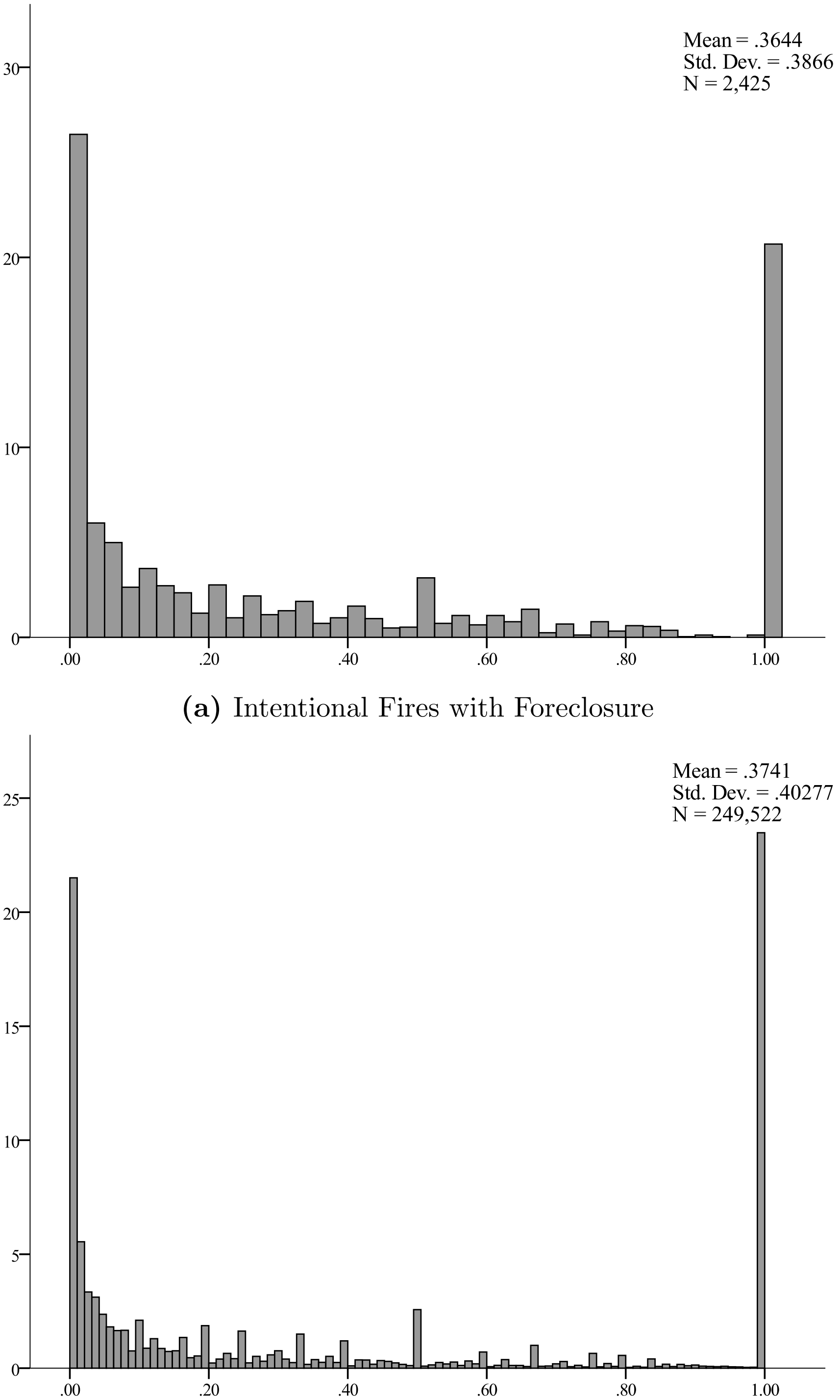


Figure A.6: Saginaw, MI All Fire Alarm Time

Reviewing the monetary estimates from the NFIRS data did not reveal a significant relationship. Figures A.7 through and A.12 display this relationship. The bumps around 50% in these histograms suggested most investigators favor round estimates: 0%, 50%, and 100%.

Property Loss

This section presents descriptive statistics and figures related to the percent value of estimated dollar property lost. For these results any outliers where the loss was greater than the pre-incident property value were removed. Figures A.7, A.8, and A.9 display this relationship.



(a) Intentional Fires with Foreclosure

(b) Unintentional Fires without Foreclosure

Figure A.7: National Fires Percent Property Loss Data

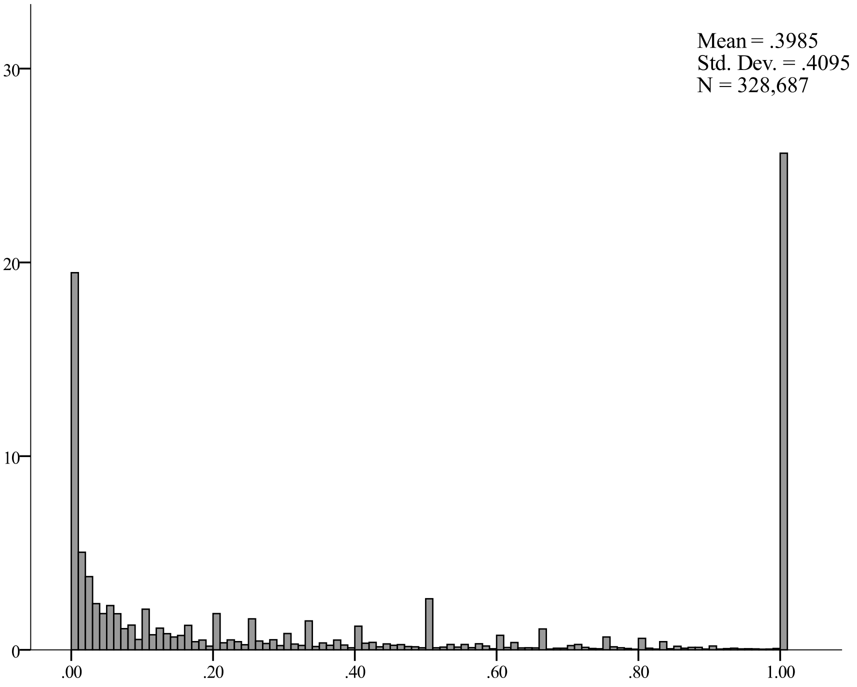


Figure A.8: National All Fires Percent Property Loss Data

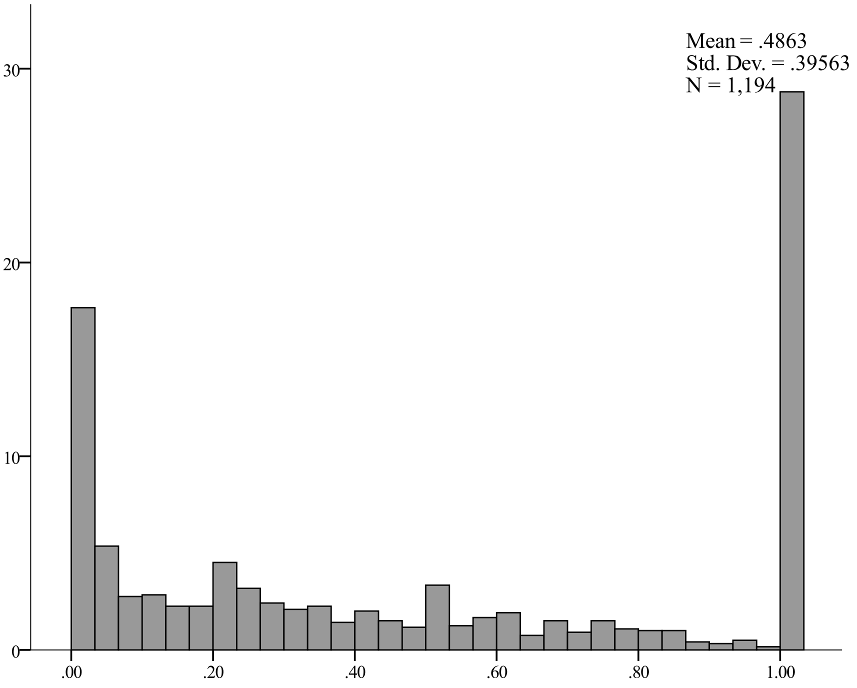
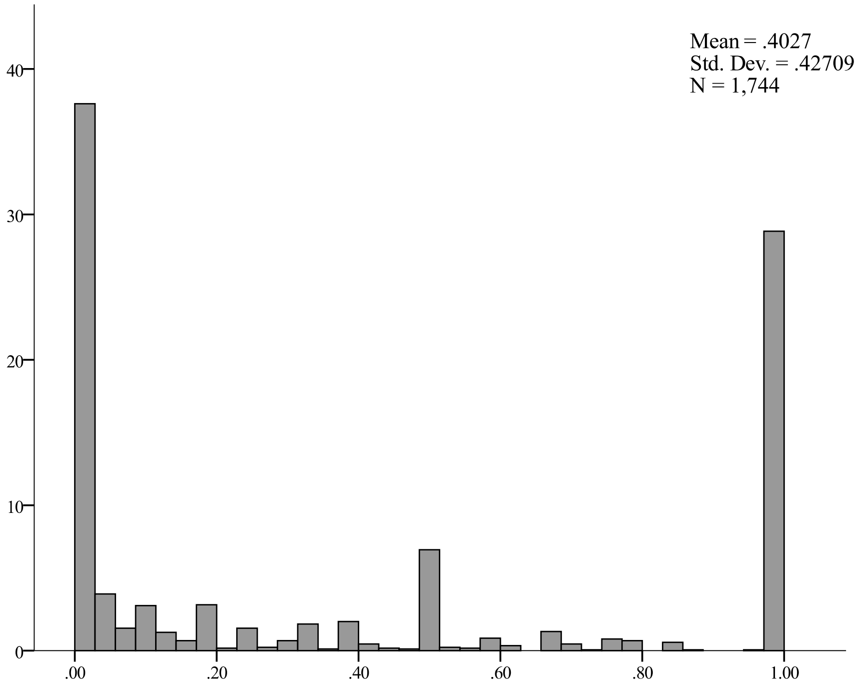


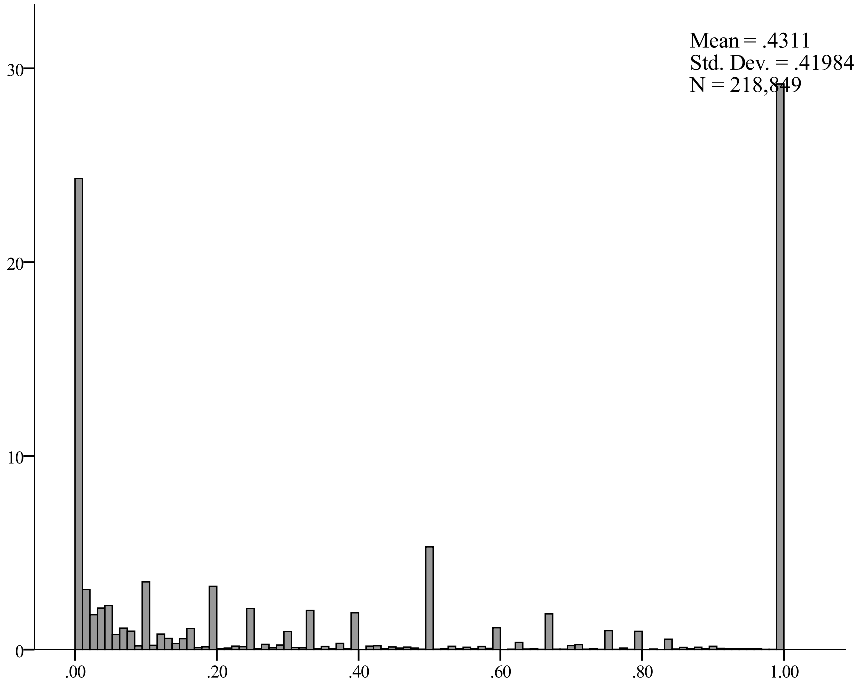
Figure A.9: Saginaw, MI Fires Percent Property Loss Data

Content Loss

Figures A.10, A.11, and A.12 display the percent of content loss for the datasets investigated.



(a) Intentional Fires with Foreclosure



(b) Unintentional Fires with No Foreclosure

Figure A.10: National Fires Percent Content Loss Data

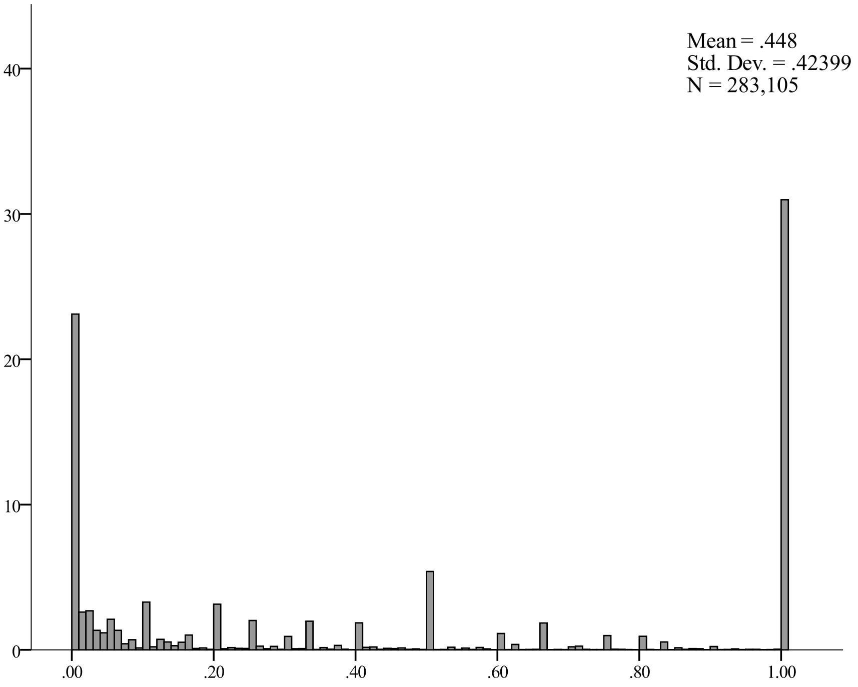


Figure A.11: National All Fires Percent Content Loss Data

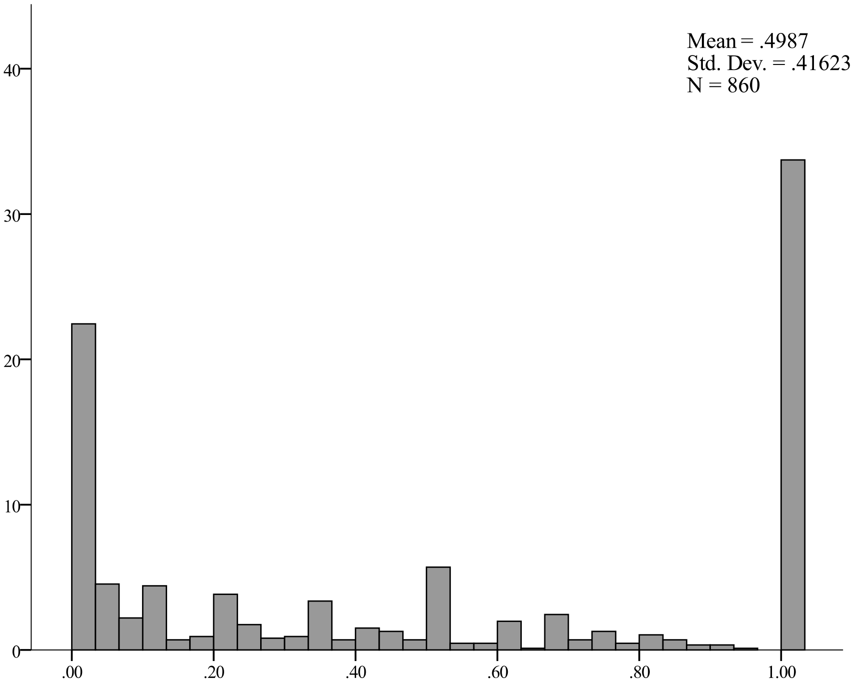


Figure A.12: Saginaw, MI Fires Percent Content Loss Data

B National Fire Incident Reporting System Summary

Variable information as provided by [31]

Alarm Time

The actual month, day, year, and time of day (hour, minute, and (optional in on-line entry) seconds) when the alarm was received by the fire department. This is not an elapsed time.

Area of Fire Origin

The primary use of the area where the fire started within the property. The area of origin may be a room, a portion of a room, a vehicle, a portion of a vehicle, or an open area devoted to a specific use. Every fire has an area of fire origin.

Cause of Ignition

The general causal factor that resulted in a heat source igniting a combustible material. The cause could be the result of a deliberate act, mechanical failure, or act of nature.

Contents Loss

Estimate of the total contents dollar loss. An estimate of the contents dollar loss is required for all fires where the value is known. This estimation of the fire loss includes contents damaged by fire, smoke, water, and overhaul. This does not include indirect loss, such as business interruption.

Contents Value

Pre-incident value estimation of the replacement cost of the contents.

Fire Spread

The extent of fire spread in terms of how far the flame damage extended. The extent of flame damage is the area actually burned or charred and does not include the area receiving only heat, smoke, or water damage.

First Item Ignited

The use or configuration of the item or material first ignited by the heat source. This block identifies the first item that had sufficient volume or heat intensity to extend to uncontrolled or self-perpetuating fire.

Incident Date

The month, day, and year of the incident. This date is when the alarm was received by the fire department and must be the same as the date for the alarm time.

Initial Heat Source

The specific source of the heat energy that started the fire.

Property Loss

Estimate of the total property dollar loss. An estimate of the property dollar loss is required for all fires where the value is known. Losses: Rough estimation of the total loss to the structure, in terms of the cost of replacement in like kind and quantity. This does not include indirect loss, such as business interruption.

Property Value

Estimate the pre-incident value of the property. Pre-incident value estimation of the replacement cost of the structure.

Property Use

Each individual property has a specific use, whether a structure or open land. The intent of this entry is to specify the property use, not the configuration of the building or other details of the property.

Type of Material First Ignited

The composition of the material in the item first ignited by the heat source. The type of material ignited refers to the raw, common, or natural state of the material. The type of material ignited may be a gas, flammable liquid, chemical, plastic, wood, paper, fabric, or any number of other materials.

Vita

Alex Asbury was born in Knoxville, TN, to parents Pamela Meadows and Mike Asbury. Alex attended Jefferson County High School where he graduated with honors in May 2009. Directly following high school, he began studies at Walters State Community College in Morristown, TN. After finding direction and inspiration, he graduated Magna Cum Laude with an Associate of Science in Pre-Professional Engineering in May 2012. Transferring to The University of Tennessee in Knoxville, he attained his Bachelor of Science in Electrical Engineering in May 2015. In the summer of 2015, Alex explored other fields and began his Master's studies with a graduate certificate in Fire Protection Engineering. Fire Protection Engineering became the concentration of his Master's of Science degree. His research was centered around understanding arson and the characteristics of structures that burn.