NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

UTILIZATION OF MACHINE LEARNING TECHNIQUES TO DETECT ANOMALIES IN DEPARTMENT OF DEFENSE CONTRACT DATA

by

Nicholas L. Lee

March 2019

Thesis Advisor:	Robert A. Koyak
Co-Advisor:	Jonathan K. Alt
Second Reader:	Anthony D. Smith

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE				*Form Approved OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.				
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** March 2019	**3. REPORT TYPE AND DATES COVERED** Master's thesis		
4. TITLE AND SUBTITLE UTILIZATION OF MACHINE LEARNING TECHNIQUES TO DETECT ANOMALIES IN DEPARTMENT OF DEFENSE CONTRACT DATA				**5. FUNDING NUMBERS**
6. AUTHOR(S) Nicholas L. Lee				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000				**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A				**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.				**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The Federal Funding Accountability and Transparency Act (FFATA) of 2006 makes available to public view information on all federal contracts beginning in 2006. This transparency presents an opportunity to examine large volumes of procurement data, in particular to infer whether anomalies or irregularities are present. In this thesis, we examine direct-order purchases made by the U.S. Army between calendar years 2013 and 2017. A total of 73,570 direct-order contracts were issued by the Army during this period, with a total obligation value of over $36 billion. We use supervised machine learning techniques to detect trends regarding levels of competition, set-aside programs used, sole sourcing, and monies spent both in individual contracts and in the awarding offices that issued the contracts. We also identify specific contracts that warrant further inspection. The suite of analytical tools that we develop can be applied generally to direct-order contracts issued by other DoD service branches. Application of these tools would allow an investigator to identify DoD contracts that warrant further scrutiny, and would allow contracting activities to be monitored with respect to criteria that are identified with best spending practices.

14. SUBJECT TERMS anomaly detection, machine learning, procurement, contract data, financial data, spending			**15. NUMBER OF PAGES** 90
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

UTILIZATION OF MACHINE LEARNING TECHNIQUES TO DETECT ANOMALIES IN DEPARTMENT OF DEFENSE CONTRACT DATA

Nicholas L. Lee
Major, United States Army
BS, Westminster College, 2003

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2019

Approved by: Robert A. Koyak
Advisor

Jonathan K. Alt
Co-Advisor

Anthony D. Smith
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Federal Funding Accountability and Transparency Act (FFATA) of 2006 makes available to public view information on all federal contracts beginning in 2006. This transparency presents an opportunity to examine large volumes of procurement data, in particular to infer whether anomalies or irregularities are present. In this thesis, we examine direct-order purchases made by the U.S. Army between calendar years 2013 and 2017. A total of 73,570 direct-order contracts were issued by the Army during this period, with a total obligation value of over $36 billion. We use supervised machine learning techniques to detect trends regarding levels of competition, set-aside programs used, sole sourcing, and monies spent both in individual contracts and in the awarding offices that issued the contracts. We also identify specific contracts that warrant further inspection. The suite of analytical tools that we develop can be applied generally to direct-order contracts issued by other DoD service branches. Application of these tools would allow an investigator to identify DoD contracts that warrant further scrutiny, and would allow contracting activities to be monitored with respect to criteria that are identified with best spending practices.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION 1
- A. BACKGROUND 1
- B. THESIS PURPOSE 2
- C. RESEARCH QUESTIONS 2
- D. SCOPE, LIMITATIONS, AND ASSUMPTIONS 3
- E. ORGANIZATION OF THESIS 3

II. BACKGROUND AND RELATED WORK 5
- A. ANOMALY DETECTION 5
- B. APPLICATIONS OF MACHINE LEARNING TO DETECT FRAUD 5
- C. INDICATORS OF FRAUD 6
- D. EXAMPLES OF FRAUD IN DOD PROCUREMENT 7
- E. FEDERAL AGENCY INITIATIVES 10
- F. CONCLUSIONS 12

III. DATA AND METHODOLOGY 13
- A. DATA 13
  - 1. Structure 13
  - 2. Naming Hierarchies 13
  - 3. Subsets 14
- B. STATISTICAL LEARNING TECHNIQUES 17
  - 1. Response Variables 19
  - 2. Trend Modeling 20
  - 3. Outlier Modeling 22
  - 4. Analysis of Awarding Offices by Contract Attributes 24

IV. RESULTS AND ANALYSIS 28
- A. SPENDING PATTERN ANALYSIS 29
- B. TREND ANALYSIS 30
  - 1. Monitoring Set-Aside Awards 31
  - 2. Monitoring Multi-year Awards 32
  - 3. Monitoring Non-commercial Awards 33
  - 4. Monitoring Sole-Source Awards 34
  - 5. Monitoring Non-competed Awards 35
  - 6. Monitoring the Designation of Small Recipients 36
  - 7. Recipient Focus 37

C. ANALYSIS OF P-VALUES....40
D. COMBINED APPROACH....41
1. Random Selection....42
2. Compelling Trends....44

V. CONCLUSIONS, RECOMMENDATIONS, AND FUTURE WORK....48
A. FUTURE WORK....48
1. Application....48
2. Relationships....49
3. Data....49

APPENDIX A. SIX CATEGORIES OF CONTRACT FRAUD....50

APPENDIX B. NAICS CODES AND DESCRIPTIONS....52

APPENDIX C. GENERAL FRAUD INDICATORS....54

APPENDIX D. MANAGEMENT FRAUD INDICATORS....56

APPENDIX E. DATA COLUMN NAMES....58

LIST OF REFERENCES....66

INITIAL DISTRIBUTION LIST....70

LIST OF FIGURES

Figure 1. Federal Organizational Quarterly Competition Report Card. Source: Defense Pricing and Contracting (2019)....11

Figure 2. Plot of Yearly Spending for Chosen NAICS Coded Awards. Adapted from USASpending (2019)....15

Figure 3. Histogram of Chosen NAICS Award Amounts....16

Figure 4. Histogram of Chosen NAICS Logarithmic Award Amounts....17

Figure 5. Random Forest Spending Plot of Boxplot Attributes for Amounts of Awards by Awarding Offices....29

Figure 6. Categorical Random Forest Model Output of the Probability of a Set-Aside Award (Black) for the Awarding Office W25G1V versus the Average Probabilities of all Awarding Offices (Red), across Time....32

Figure 7. Categorical Random Forest Model Output of the Probability of a Multi-year Award (Black) for Awarding Office W25G1V versus Average Probabilities of All Awarding Offices (Red), across Time....33

Figure 8. Categorical Random Forest Model Results for the Probability of the Items or Services Purchased Are Commercial (Black) for the Awarding Office W25G1V versus Average Probabilities of All Awarding Offices (Red), across Time....34

Figure 9. Categorical Random Forest Model Results for the Probability of a Sole-Source Award (Black) for Awarding Office W25G1V versus Average Probabilities of All Awarding Offices (Red), across Time....35

Figure 10. Categorical Random Forest Model Results for the Probability of a Fully and Openly Competed Award (Black) for Awarding Office W91ZLK versus Average Probabilities of All Awarding Offices (Red), across Time....36

Figure 11. Categorical Random Forest Model Results for the Probability of a Contracting Officer Categorizing a Business as Small (Black) for Awarding Office W912BU versus Average Probabilities of All Awarding Offices (Red), across Time....37

Figure 12. Categorical Random Forest Predicted Residuals of a Recipient's Determination of Small Business (Black) Plotted against All Other Recipient Averages (Red)....39

Figure 13. Amount of Annual Revenue for Recipient 806026852 in Millions of Dollars, through Time 40

Figure 14. Categorical Random Forest Model for the Probability of a Set-Aside Award (Black) For the Awarding Office W25G1V versus the Average Probabilities of All Awarding Offices (Red), over Time 42

Figure 15. Categorical Random Forest Plot of Predicted Residual Probabilities of Awards Categorized as Fully and Openly Competed (Black) for Awarding Office W91ZLK, versus Averages (Red) of All Awarding Offices 44

Figure 16. Categorical Random Forest Plot of Predicted Residual Probabilities of Awards Categorized as Sole-Source (Black) for Awarding Office W25G1V, versus Averages (Red) of All Awarding Offices 45

LIST OF TABLES

Table 1. Response Variables 20

Table 2. Categorical Random Forest Data Predictor Variables 21

Table 3. Random Forest Predictor Variables Used to Model Spending 23

Table 4. Analytical Tools 28

Table 5. Top Ten Contract Awards by Outlier Extremity 30

Table 6. NAICS Codes, Descriptions, and the SBA Lower Limit of Consideration for Small Business Designation for FY 2018 38

Table 7. Top Ten Awarding Offices by Extremity of the Number if Set-Aside Awards Made 41

Table 8. Anomalies for Awarding Office W25G1V Expressed in Thousands of Dollars 43

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AC4S	Advanced C4 Technologies
App	Application
BPA	Blanket Purchase Award
CSV	Comma Separated Values
DC	Definitive Contract
DFAS	Defense Finance and Accountability Services
DO	Delivery Order
DPC	Defense Procurement Clearinghouse
DA	Department of the Army
DoD	Department of Defense
DPC	Defense Pricing and Contracting
EMALL	Electronic Mall
FAR	Federal Acquisitions Regulation
FPDS-NG	Federal Procurement Data System-Next Generation
FFATA	Federal Funding Accountability and Transparency Act
FY	Fiscal Year
GAO	Government Accountability Office
GSA	General Services Administration
GUI	Graphical User Interface
IBT	Iron Bow Technologies
IDV	Indefinite Delivery Vehicle
IG	Inspector General
IT	Information Technology
IQR	Inter-Quartile Range
LLC	Limited Liability Company
MART	Multiple Additive Regression Trees
NAICS	North American Industry Classification System
PIID	Procurement Instrument Identification
PO	Purchase Order

SBA	Small Business Administration
SCSI	Superior Communications Solutions Incorporated
SDVOSB	Service-Disable Veteran-Owned Small Business
SML	Supervised Machine Learning
UMDB	United Medical Design Builders
UML	Unsupervised Machine Learning

EXECUTIVE SUMMARY

In fiscal year 2019, the U.S. federal budget exceeded $4.407 trillion (Amadeo, 2018), of which approximately $500 billion was expended in the form of federal contracts (USASpending, 2019). There is a fundamental public interest in knowing how the federal government spends its money. Instances of potential mis- or malfeasance often are currently discovered by manual inspection of financial records that limits the amount of records that can be examined. A systematic approach, based on expenditure data and statistical machine learning enables the application of investigative resources for those instances that generate enough justification for a deeper (and costlier) level of information-gathering. This approach could be used to supplement traditional random auditing.

In this research, we develop statistical methods to identify anomalous U.S. Army contracts over a five-year period. We find that variables explaining the competitive environment of awards give useful insights regarding unusual purchase transactions and trends. The competitive variables assist with the examination of government agency initiatives created to ensure best spending practices.

We define normal Department of Defense contractual behavior through the use of statistical and machine learning for visual representation and comparison.

Using our methods, we find that anomalies are present in the contractual data. We highlight these in our results giving auditors multiple avenues of approach to question deviations from baseline conditions and affords them the chance to acquire details regarding observations of interest not contained within the USASpending data.

Although the USASpending data can produce useful insights, the information that it contains about the nature of transactions is limited. We recommend the addition of pricing and quantity lists for each award to assist trend analysis. We also recommend the addition of naming the competing bidders for each award to gain an understanding of the circumstances that produce a finding of interest in the USASpending data.

References

Amadeo, K. (2018). U.S. federal budget breakdown: The budget components and impact on the U.S. economy. *The Balance*, 1. Retrieved from https://www.thebalance.com/u-s-federal-budget-breakdown-3305789

USASpending. (2019). Award data archive. Retrieved from https://www.usaspending.gov/#/download_center/award_data_archive

ACKNOWLEDGMENTS

To my thesis advisor, Professor Robert M. Koyak: because I owe it all to you. Many thanks for the countless hours spent with me on this project. You gave direction and insight when necessary, and facilitated the expansion of my knowledge.

For my co-advisor, Professor Jonathan K. Alt: thank you for the problem set. Between your Analytics Case Studies class and your constant availability, you gave me fresh methods to employ, new avenues of approach to explore, and just common sense advice whenever it was warranted.

Thank you to Major Anthony Smith, my second reader. Having graduated and moved on, your knowledge on writing and formatting helped immensely. I proffer that your pep talks and methods for bringing me back to the center of the problem were imperative to my success.

Andrea: there is not enough room here. This thesis is a culmination point of a long list of tests which, throughout two years, you have enabled me to complete. Taking care of the family, never complaining, and always encouraging me, you are so strong. You are my North Star, as always. Thank you. I love you.

To Kay, and my mother, thank you for all your support. Without you, this would have never been possible. Helping Andrea and me get a house, and following us here to help take care of the kids while I study. I am in your debt, so no change. More so, you both installed an appreciation for education, motivating me to continue to educate myself, that has gotten me to this point.

For my mother and father in law, thank you for all your support. It takes a tribe, and this program combined with two small children has proved challenging.

A special thanks goes out to Mr. Richard Boyd, Sr., and his team at USASpending, for their hands-on assistance with many of the random questions I had regarding the data itself. You and your team are doing good things, in my meager opinion, and your work is important on many levels. Thank you.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

Rule No. 1: Never lose money. Rule No. 2: Never forget Rule No. 1.

—Warren Buffet (Friedman, 2018)

In 2019, the U.S. federal budget exceeded $4.407 trillion (Amadeo, 2018) of which approximately $500 billion is expended in the form of federal contracts (USASpending Data Lab, 2019). In fiscal year (FY) 2017, defense contract spending accounted for most of this total, at $329 billion (USASpending Data Lab, 2019). To put this in perspective, only 21 countries have an estimated gross national product exceeding $500 billion (World Bank, n.d.). Estimates of Department of Defense (DoD) funds lost to fraudulent activity vary between twelve and fourteen percent (Sellers, 1996). Approximately $60 billion was lost to contract fraud in Iraq and Afghanistan in the year 2011, signifying the need to identify fraudulent contracts (Shane, 2011). Fraud detection is one example of proper financial management and federal spending. Generally, the public has an interest in ensuring that the government spends its monies efficiently, this includes examining the competitive environment surrounding contracts. The fundamental public interest in knowing where and how the federal government spends public money led to the enactment of the Federal Funding Accountability and Transparency Act (FFATA) in 2006.

Identification of contracts that do not align with best spending practices may improve with the use of tools based on analyses of financial data. An important data resource for this purpose is publicly available from USASpending through their website at usspending.gov. The U.S. Department of the Treasury established USASpending in 2012 (Teefy, 2018). The data archived by USASpending follows the contractual pathway of an award from approval, through all modifications, and ultimately to completion. Transparency mandated by FFATA allows the public to view and download data on all contracts starting in 2008 within USASpending. FFATA matured with the enactment of the Data Accountability and Transparency Act (DATA Act) in 2014, which sets standards for the quality of federal spending data. This transparency presents a unique opportunity to examine large volumes of federal procurement data, to infer whether best spending practices are present including anomalous spending occurrences.

Instances of mis- or malfeasance often are discovered by manual inspection of financial records as part of random audits or in response to complaints obtained from various sources. Although these remain important avenues for identifying questionable spending practices they are limited because they cannot be applied systematically, which would exhaust limited investigative resources. There is interest in focusing on best-spending practices that include, for example, ensuring that contracts are awarded through competitive bidding in order to reduce costs.

B. THESIS PURPOSE

The objective of this thesis is to propose a set of statistical methods to identify DoD contracts or spending practices that may justify further scrutiny. We develop these methods on the USASpending data for U.S. Army contracts between calendar years 2013 and 2018. The U.S. Army accounted for about fifteen percent of all DoD contract spending in fiscal year 2017 (USASpending Data Lab, 2019). Our approach, however, is extendable to contracts from other DoD service branches.

C. RESEARCH QUESTIONS

Our study offers an approach to identifying irregular spending patterns in U.S. Army contracts using USASpending data. We address the following research questions:

1. Which attributes obtained from the USASpending data give useful insights regarding unusual purchase transactions and trends?

2. What are the baseline conditions for normal DoD contractual behavior, and what does this look like?

3. Can we use statistical and machine learning to identify anomalies or trends present in the U.S. Army contract data?

D. SCOPE, LIMITATIONS, AND ASSUMPTIONS

Data quality limitations such as missing values or non-standard data entry procedures can produce anomalies during initial detection. The USASpending data does not provide important features of contract purchases such as the particular items procured, quantities, and unit pricing. Lacking this information limits the ability of an analyst to probe these purchases using only the data. At present, the data does not provide useful information regarding subcontractors. While subcontractor data is available, its exploration is outside the scope of this effort.

E. ORGANIZATION OF THESIS

Organization of the remainder of this thesis is as follows:

- Chapter II is a review of the main fields of study in anomaly detection and related work associated with supervised learning, unsupervised learning methods, government initiatives and historical cases that contribute to our methodology for investigation of best spending practices. We introduce case studies of known fraud to identify which attributes of the USASpending data are indicative of irregular spending practices.

- Chapter III is a discussion of the USASpending data origin and structure. We explain our methodologies for selecting data for analysis, variable selection, and model preparation.

- Chapter IV is a discussion of the results and analysis of our research.

- Chapter V is a discussion of the conclusions of the results and our recommendations for potential future work.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND AND RELATED WORK

Our methods of analysis are motivated by anomaly detection techniques that incorporate machine learning. In particular, we focus on finding anomalous spending patterns that historically have been associated with irregular spending practices. Such irregularities may result from fraudulent activity or represent tendencies that are not consistent with best spending practices. We also review federal agency initiatives that maintain a competitive environment in order to motivate our analysis.

A. ANOMALY DETECTION

The statistical literature in the area of anomaly detection discusses a number of approaches that are applicable to financial data. Two application areas that are prominent in financial anomaly detection are network security and fraud detection. Network security is concerned with the identification of abnormal network traffic signals among a large collection of standard signals. The flow of network traffic is immense and ever-changing, which makes the analysis of such data challenging (Garcia, Diaz, Marcia & Vazquez, 2009). Fraud detection mainly targets suspicious changes in spending behavior over time. Classic approaches to fraudulent activity include solutions based upon intuition, experience and knowledge levels of investigators (Baesens, Van Vlasselaer, & Verbeke, 2015). Anomaly detection in fraud detection and network security operations follow similar strategies: identify patterns that establish “normal” baseline conditions, and then compare individual observations against the baseline to detect anomalous behavior.

B. APPLICATIONS OF MACHINE LEARNING TO DETECT FRAUD

Rouillard (2003) uses unsupervised learning (UML) methods to detect fraudulent vendor payments in Defense Finance and Accounting Services (DFAS) data. The author uses cluster analysis to determine the optimal model choice from a variety of statistical models. His utilization of UML methods incorporates explanatory variables previously left out of historical supervised learning models in order to enhance fraudulent detection measures. Rouillard’s work expands upon previous work by Monteiro (2002). Monteiro examines fraudulent DFAS vendor payments and uses multiple Additive Regression Trees (MART) to develop a predictor for them. MART pools a forest of hundreds of regression trees and uses machine learning (ML) to identify important classes of explanatory variables. Rouillard's method differs from that of Monteiro in that it identifies appropriate UML models rather than the model variables themselves, and includes all explanatory variables left out of Monteiro's model due to his MART selection criteria.

C. INDICATORS OF FRAUD

According to Baesens et al. (2015), markers of fraud include pressure, which may be financial, social, or of a different form; opportunity, for which there must exist an oversight gap allowing a fraudulent act to occur; and rationalization, by which an individual or group justifies their actions. Contract auditors rely on experience-based knowledge to identify situations in which pressure, opportunity, and rationalization may create opportunities for fraud. Anders (2015) differentiates two categories of contractual fraud: individual and organizational. Appendices C and D list general and managerial fraud indicators, respectively, identified by the U.S. Department of Defense Inspector General (Publications Branch, 1993).

Federal acquisitions must conform to the Federal Acquisitions Regulation (FAR) found in Title 48 of the Code of Federal Regulations. This extensive regulation, which has 37 chapters and comprises thousands of pages, points to the complexity of monitoring federal contracting. Contract auditors must examine individual and organizational internal control measures to ensure that they are in place and are effective. Upon finding deficiencies, auditors may recommend an investigation, and civil or criminal charges may be brought if warranted by the evidence.

Gayton (2004) examines historical cases of adjudicated contract fraud for indicators of violations standards established by the U.S. House Code of Ethics resolution, passed into Public Law 66–303 July 1980 (Code of Ethics for Government Services, 1980). He cites Public Law 96–303 to recognize that substantial responsibility for fraud prevention rests on the federal contracting office. This observation coincides with the markers identified by Baesens et al. (2015). Gayton classifies contract fraud into six categories:

defective products and product substitution; defective testing; bid rigging; bribery and public corruption; defective pricing; and, false invoices. The categories are listed and defined in Appendix A. He classifies each case into one of his six categories and identifies the corresponding code of ethics violations, from which he gives recommendations on internal control measures to prevent repeats of fraudulent actions.

Tan (2013) examines case studies compiled from the Encyclopedia of Ethical Failure issued by the DoD Defense General's Council. Tan's objective is to identify areas of weakness within internal control systems and to provide recommendations regarding their repair. He cites twenty cases of fraud within the federal contracting process beginning with procurement planning and ending at contract closeout. Sole-source contracting often is implicated in cases with solicitation irregularities. Thirteen of the 20 contracts considered by Tan came from solicitations that were designated as sole-source either legally or fraudulently. Fraud associated with sole-source contracts stems from inappropriate competition elimination to obtain a contract, or from an ability to leverage an inappropriate relationship between the awardee and the awarding office.

D. EXAMPLES OF FRAUD IN DOD PROCUREMENT

We seek metrics observable in case studies that also are observable with the USASpending data. Although most of the indicators shown in Appendices C and D cannot be derived from the USASpending data, trends regarding related attributes may indicate anomalies. We define anomalies as any trend or metric that appears highly unusual in comparison to baseline conditions. Such attributes of an award include whether it is a set-aside to a designated group, whether it is competed fully and openly, and whether the contract is awarded to a sole-source bidder. We examine five separate cases in which an awardee was found guilty of contract fraud. In each of those cases either sole sourcing, limited competition methods, or set-aside program abuses were present. Set-aside programs are reserved for businesses, owned (at least 51 percent) by individuals belonging to socially or economically disadvantaged groups (Murray, 2016). For example, a Service-Disabled Veteran-Owned Small Business (SDVOSB) may qualify for a set-aside. In two of the five cases, the bidder misrepresented itself as an SDVOSB to restrict competition in the solicitation and bidding processes.

The first case involves three actors and three businesses. Between 2010 and 2011 select personnel working for the companies Advanced C4 Solutions (AC4S), Superior Communications Solutions Inc. (SCSI), and Iron Bow Technologies were found guilty of conspiracy connected to contractual fraud. Three individuals conspired to commit fraudulent transactions in order to provide telecommunications products, at inflated prices, to the government, in a non-competitive manner. One of the individuals, while working as a federal employee, accepted employment at SCSI. In order to secure unfair competitive advantages, the three individuals worked together to drive contracts towards their companies and companies they had an affiliation with by tailoring specifications of awards. The individuals also engaged in collusive bidding schemes (Kramer, 2012) where one of their companies secretly agreed to submit a high subcontract bid, allowing another one of their companies to win the subcontract by bidding lower. Upon award of the subcontract, the awarded company would then subcontract the award to another one of their companies. The individuals would then submit statements of work on behalf of both subcontracted companies, earning double wages.

The second case involves defrauding the federal government by falsely claiming a disadvantaged status to take advantage of set-aside programs. In FY 2008, individual A was working on a subcontract project at Fairchild Air Force Base. A representative from United Medical Design Builders (UMDB) approached individual A in search of someone to pose as their established disabled-veteran figurehead. Individual A agreed to fill the role, thereby granting the business access to set-aside programs and competition restrictions linked to a disabled-veteran status. The indictment charged that UMDB received awards for four separate construction projects, on four separate Air Force bases, totaling $40 million (United States Attorney's Office District of Kansas, 2017).

The third case demonstrates how an insider threat can operate in the presence of pressure and opportunity markers mentioned by Baesens et al. (2015). An insider threat refers to an individual working within an organization to circumvent regulatory constraints that separate them from personal financial gain (Costa, 2017). During fiscal years 2007 through 2011, Individual B was the mastermind of the largest domestic bid-rigging and bribery scheme in federal contract history (United States Attorney's Office District of Columbia, 2013). As a former program manager and contracting officer technical representative, individual B knew contractual processes. Individual B worked with others to steer contracts to vendors who would reward him with bribes. Individual B and his cohorts schemed to use an Alaskan Native-owned company, EyakTek, to gain a controlled-competition environment securing federal contracts, and then subcontracted them out to larger businesses such as Nova Datacom LLC to receive kickbacks.

Case four again exhibits abuse of set-aside programs. In FY 2006, individual C established Legion Construction, Inc. Individual C recruited a disabled Korean War veteran to be the figurehead of the company. In March of 2010, a bid which Legion Construction won was challenged by another SDVOSB, alleging that individual C appeared to be Legion's owner and operator, not a disabled veteran (United States Attorney's Office District of Maryland, 2017).

The final case involves fraud associated with product substitution. Individuals D and E were the owners of Veteran Logistics, Industrial Xchange, and Boston Laser Technology. Both individuals regularly sold supplies to the DoD through their companies using multiple contracts with the DoD's Defense Logistics Agency, which were used as vehicles to sell products to the federal government through the Electronic Mall (EMALL), now known as the Federal Mall (FedMall) (Mcallister, 2017). FedMall is an online resource for DoD customers to purchase commercial goods. Their contracts allowed them to sell pre-approved goods for maximum prices to authorized EMALL users. In one example of product substitution, the individuals agreed to supply the Maritime Expeditionary Security Group Two with over 10,000 Post-it notepads. After colluding with Naval personnel, the individuals substituted the current order of Post-it notes with 50 electronic transceivers. This gave them the ability to charge a 134 percent markup on products they were not authorized to sell.

E. FEDERAL AGENCY INITIATIVES

The federal government has instituted a number of studies focusing on best practices for contractual award processes. In December of 2014, the Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics published guidelines for creating and maintaining a competitive environment (Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2014). This document emphasizes the importance of competition focusing on how a competitive environment leads to secure lower prices, raise innovation, invoke higher quality, raise performance standards, and offer more opportunity for a wide range of vendors to do business with the government. The Defense Pricing and Contracting (DPC) website contains a link to this document accompanied by report cards concerning organizational quarterly competition metrics. Figure 1 displays the competition metric report card for the first quarter of FY 2018.

Contracting Agency	Total Actions	Total Dollars	Competed Dollars	FY 2018 Compete %	FY 2018 Goal %	Not Competed Dollars	Non-Compete Weight
DEPT OF THE AIR FORCE	80,704	$ 48,134,779,330	$ 17,772,568,997	37%	41%	$ 30,329,832,265	26%
DEPT OF THE ARMY	141,739	$ 57,516,803,631	$ 32,515,533,041	57%	56%	$ 24,751,631,353	21%
DEPT OF THE NAVY	167,733	$ 71,213,007,163	$ 24,598,758,439	35%	40%	$ 46,430,824,064	40%
DEFENSE LOGISTICS AGENCY	2,525,912	$ 30,140,573,414	$ 23,778,676,725	79%	78%	$ 6,332,121,714	5%
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY	1,029	$ 656,344,911	$ 546,302,983	83%	77%	$ 109,878,011	0%
DEFENSE COMMISSARY AGENCY	4,243	$ 177,249,430	$ 89,432,793	50%	47%	$ 87,154,309	0%
DEFENSE CONTRACT MANAGEMENT AGENCY	146	$ 43,738,005	$ 17,868,049	41%	72%	$ 21,760,991	0%
DEFENSE FINANCE AND ACCOUNTING SERVICE	633	$ 200,937,894	$ 168,227,738	84%	83%	$ 29,799,388	0%
DEFENSE HEALTH AGENCY	3,136	$ 12,185,895,871	$ 10,570,327,234	87%	87%	$ 1,615,108,562	1%
DEFENSE HUMAN RESOURCES ACTIVITY	311	$ 107,218,653	$ 84,050,135	78%	73%	$ 23,148,875	0%
DEFENSE INFORMATION SYSTEMS AGENCY	36,516	$ 3,818,884,353	$ 2,701,364,164	71%	73%	$ 1,101,118,302	1%
DEFENSE MEDIA ACTIVITY	296	$ 59,956,140	$ 40,486,217	68%	73%	$ 19,292,540	0%
DEFENSE MICROELECTRONICS ACTIVITY	489	$ 617,996,851	$ 428,167,613	69%	88%	$ 189,829,238	0%
DEFENSE SECURITY COOPERATION AGENCY	220	$ 33,841,772	$ 25,128,450	74%	70%	$ 8,713,322	0%
DEFENSE SECURITY SERVICE	118	$ 44,595,396	$ 28,906,191	65%	75%	$ 15,689,205	0%
DEFENSE THREAT REDUCTION AGENCY	1,182	$ 566,501,964	$ 458,425,158	81%	86%	$ 108,020,369	0%
DEPT OF DEFENSE EDUCATION ACTIVITY	1,159	$ 165,689,138	$ 117,537,383	71%	73%	$ 48,126,935	0%
MISSILE DEFENSE AGENCY	3,074	$ 6,992,292,877	$ 3,196,699,775	46%	46%	$ 3,795,355,101	3%
U.S. CYBER COMMAND	7	$ 7,225,049	$ 5,318,757	74%	N/A	$ 1,906,292	0%
U.S. SPECIAL OPERATIONS COMMAND	5,103	$ 2,373,443,221	$ 1,758,288,305	74%	72%	$ 614,606,134	1%
UNIFORMED SERVICES UNIVERSITY OF THE HEALTH SCIENCES	418	$ 17,410,666	$ 15,075,301	87%	60%	$ 2,335,365	0%
USTRANSCOM	23,812,575	$ 3,426,806,941	$ 3,398,634,815	99%	98%	$ 28,172,126	0%
WASHINGTON HEADQUARTERS SERVICES	2,104	$ 822,587,833	$ 393,646,870	48%	57%	$ 372,003,446	0%
DEPARTMENT OF DEFENSE TOTAL	**26,803,731**	**$ 239,690,465,112**	**$ 122,854,324,101**	**51%**	**53%**	**$ 116,257,207,337**	**100%**

Notes:
DCMA only covers procurement center S5121A
U.S. CYBER COMMAND reporting was added in the 2nd quarter FY 2018.

Legend
- **>FY2018 Compete % Goal**
- **<FY2018 Compete % Goal**

Figure 1. Federal Organizational Quarterly Competition Report Card. Source: Defense Pricing and Contracting (2019).

The report card shows the total number of actions, dollar amounts, and competed-for dollar amounts as well as the FY goal, progress percentage towards that goal, and the total amount given to non-competitive contracts and their weight. The report cards show the importance of maintaining a competitive procurement environment. These report cards are also consolidated further into a full competition report available on the DPC website. The competition report further analyzes the effects surrounding the competition process. It details DoD competition trends, the effectiveness of competition, and includes comparative statistical analysis between civilian and DoD contracts.

A separate report created by the U.S. Government Accountability Office (GAO), illustrates the importance of competition in the procurement process. The main findings of the GAO report include the following: savings correlate directly with competition, and a lack of government ability to identify or record data associated with contract service or product costs significantly hinders the ability to carry out a competitive bidding process (Mihm & Holman, 1998). If an item is commercially available, the competitive environment usually is healthy (Office of the Under Secretary of Defense Acquisitions, Technology, and Logistics, 2014).

F. CONCLUSIONS

This review of relevant literature demonstrates an association between levels of competition, solicitation methods, and questionable use of set-aside programs with spending practices that may warrant further scrutiny. The review lends insight into our research question (1): which attributes obtained from the USASpending data give useful insights regarding anomalous purchase transactions and trends? These three areas of focus relate to the proper management of taxpayer money.

III. DATA AND METHODOLOGY

"We estimate that 80% of the time taken in any data analysis problem is taken up just in reading and preparing the data" (Buttrey and Whitaker, 2018). In this chapter, we describe the accessibility, unique attributes, and structure of the USASpending data. We provide a summary of statistical learning techniques and a general discussion of how we create and build our models.

A. DATA

USASpending consolidates data from multiple government systems to improve the transparency of federal government spending. Contract data is supplied daily from the Federal Procurement Data System-Next Generation (FPDS-NG). The data is available through customizable downloads at https://www.usaspending.gov/#/download_center/award_data_archive, allowing users to specify fiscal years, government agencies, recipients, and contracts of interest. Timeframes available for downloadable data are the FY 2008 through the present. Downloads are in the form of Excel comma separated value files (CSV).

1. Structure

Each row of a downloaded CSV file is a financial action of an award. Columns represent details of the award such the contracting office, the recipient of the award, and the total amount of the award. There are 258 columns, and each CSV file contains up to one million rows. Appendix C contains a list of the column names in the USASpending data. The size of the data in its entirety is approximately one terabyte.

2. Naming Hierarchies

In the acquisitions process, there are multiple instances of spending under parent award identifications (IDs). The parent award identification connects the recipient to the original award. Under a parent award ID, it is possible to have multiple award identification procurement instrument identifiers (Award ID PIIDs). The award ID PIIDs are charges against a parent award. A federal agency may not have the ability to predict what supplies or services it will need as a project progresses. Therefore, they may need to make multiple orders against a parent award ID. There are five types of ordering instruments, known as indefinite delivery vehicles (IDV). Linking occurs between IDVs and the types of calls or orders the agency can use to procure goods and services through the duration of the contract life. Separate regulations and constraints apply to each IDV. Modifications of individual parent award IDs and award ID PIIDs are represented by separate records or rows in the CSV file.

3. Subsets

We inspect contracts classified as delivery orders due to their abundance in the data. Approximately 80 percent of the contracts are delivery orders. We seek complete contracts, meaning we are interested in data that includes the contract initiation and every action under that contract until closeout. Our focus is on Department of the Army (DA) contracts for the fiscal years 2013 through 2018.

Approximately 73 percent of delivery orders are firm fixed price, meaning that there is no price negotiation, and the recipient accepts full responsibility concerning costs and resulting profit or loss ("GSA Federal Procurement," 2018). North American Industry Classification System (NAICS) numbers categorize products and services purchased in a contract. We create a subset of data for information technology (IT) services or products to include an umbrella of activities that could be related to IT. Appendix B lists the NAICS codes and descriptions selected.

A cost analysis would prove complicated upon comparing information technology services or products to construction. Figure 2 depicts the amounts spent on goods and services by our chosen NAICS codes per year.

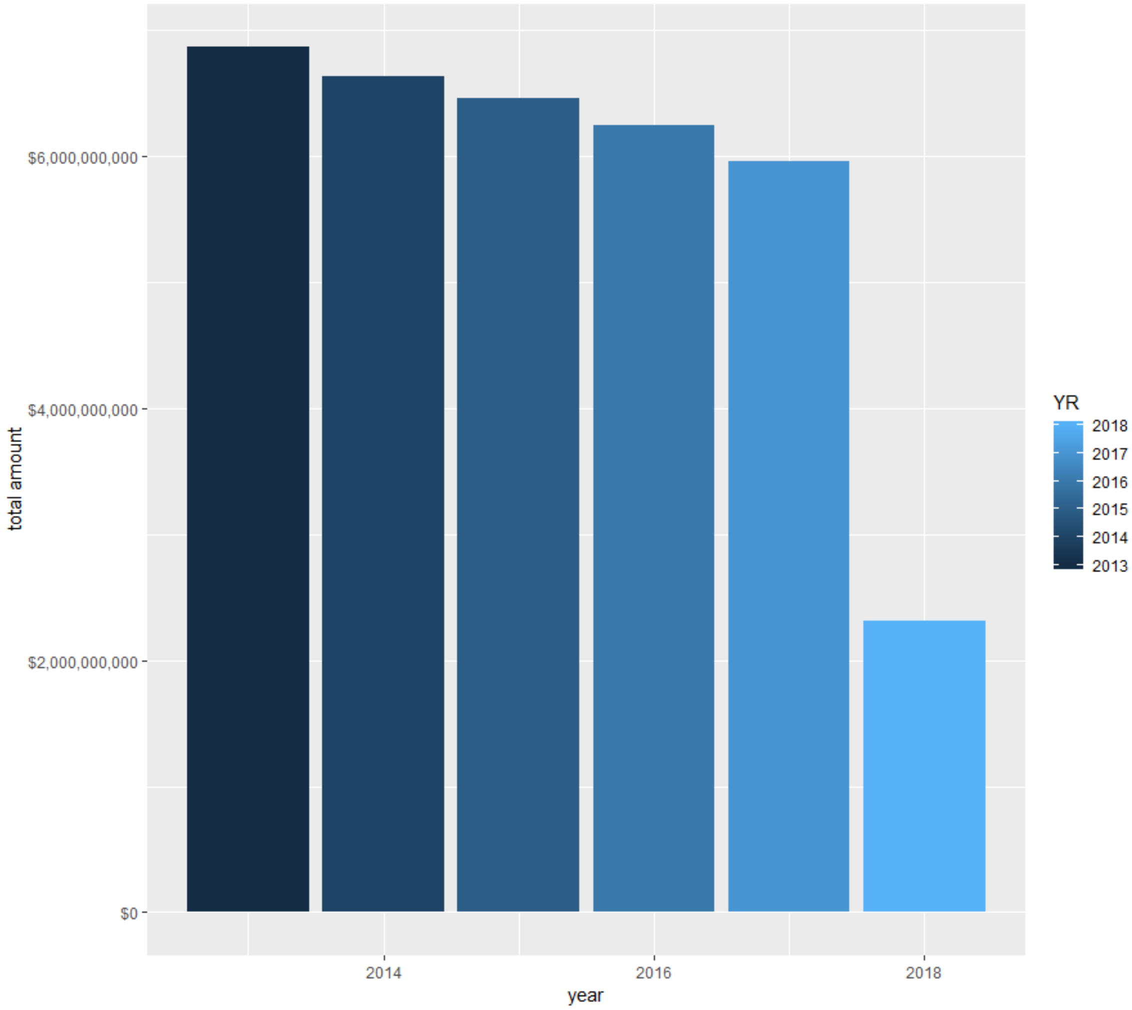

Figure 2. Plot of Yearly Spending for Chosen NAICS Coded Awards. Adapted from USASpending (2019).

The maximum amount spent for one award is approximately $138 million. The total amount spent for our chosen NAICS products or services is approximately $5.7 billion. We convert the award amounts to a logarithmic scale thereby limiting the skewness of potential outliers towards substantially large values. Figure 3 shows the histogram for the frequency of award amounts within our data.

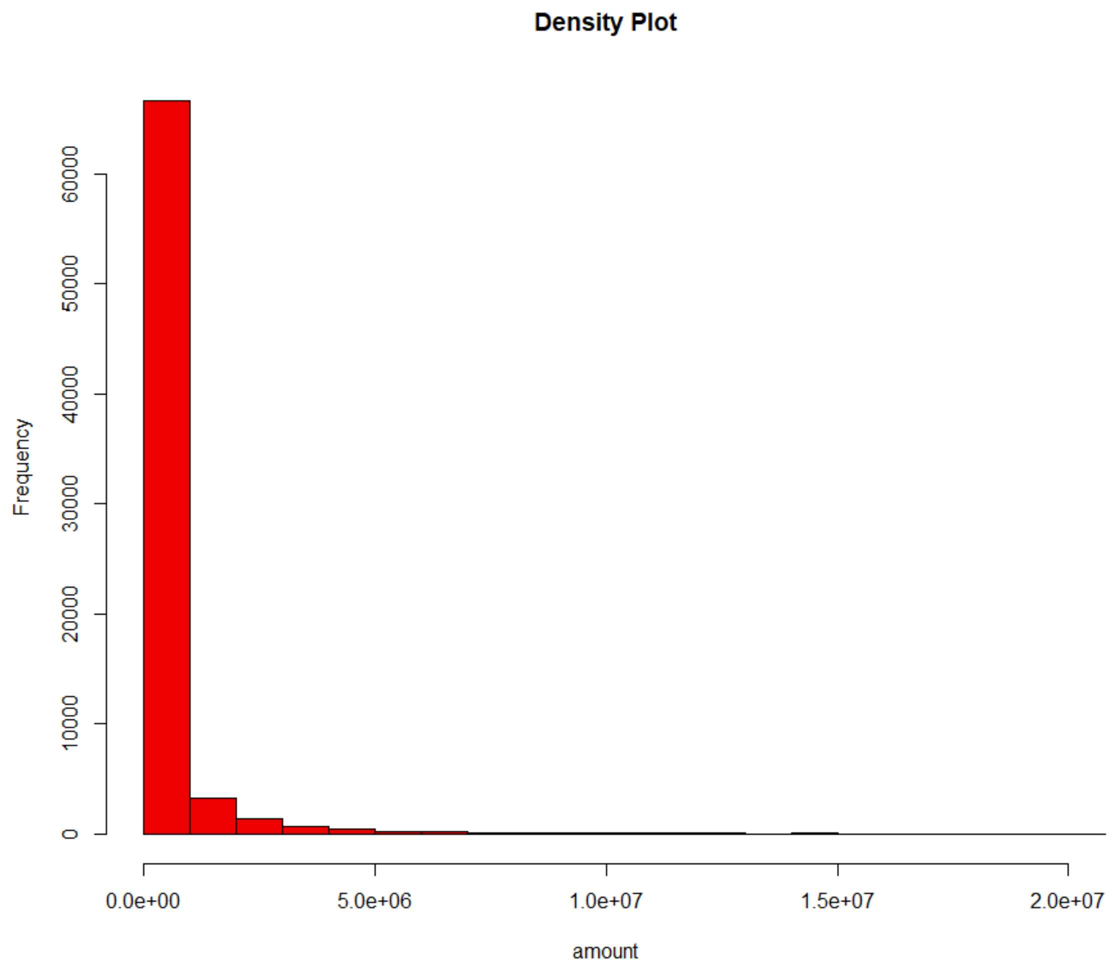

Figure 3. Histogram of Chosen NAICS Award Amounts

The logarithmic transformation eliminates right tail skewness shown in Figure 3 and balances trend analysis by evening out the difference between extremely large and small amounts as shown in Figure 4.

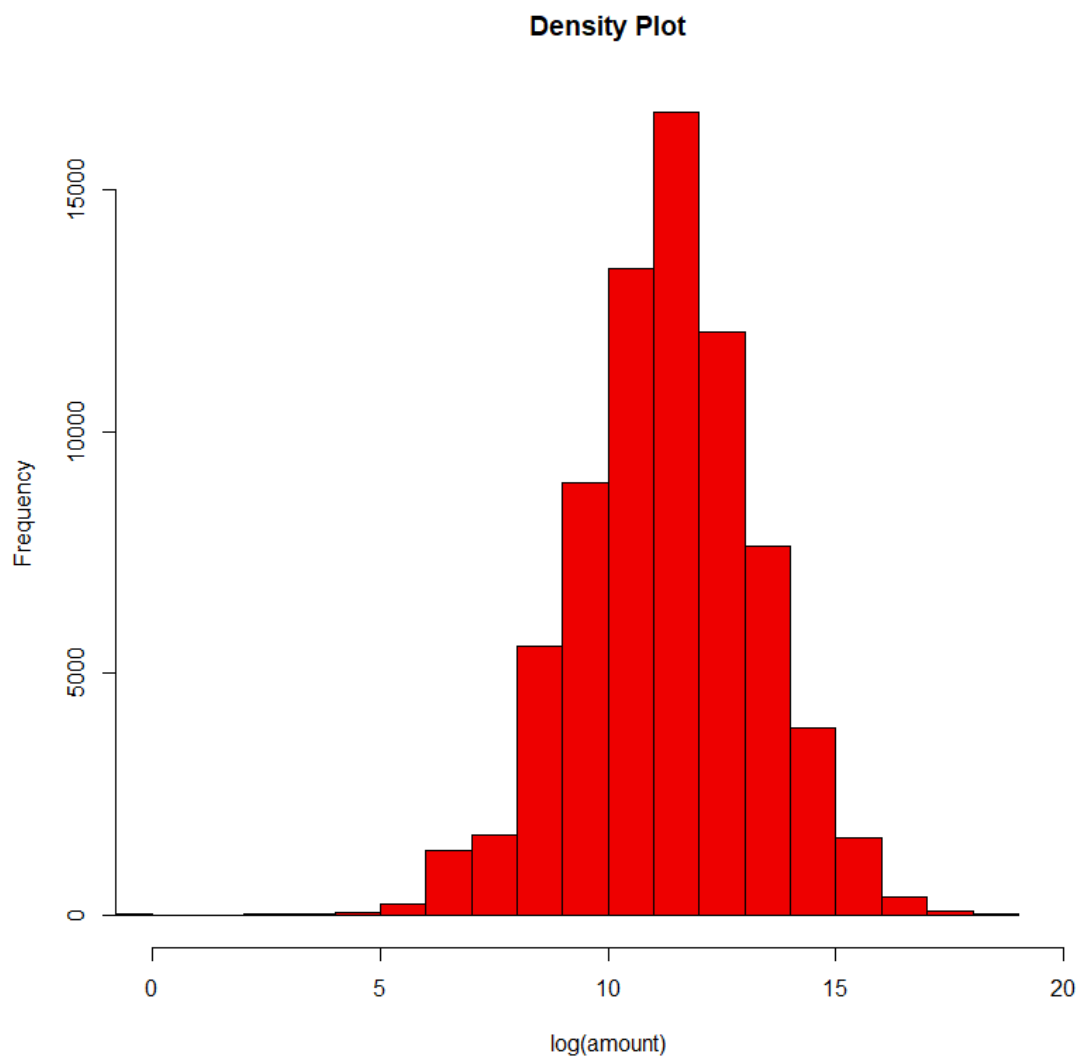

Figure 4. Histogram of Chosen NAICS Logarithmic Award Amounts

B. STATISTICAL LEARNING TECHNIQUES

Statistical learning techniques are classified as either SML or UML. SML relates the data to known cases of anomalous behavior, in contrast to UML in which such labeling is absent. UML seeks to identify trends and patterns that may point to anomalies without explicit labeling of cases as being anomalous.

A machine-learning analysis involves the use of a collection of tools that include statistical profiling with histograms, neural networks, clustering, and a mixture of methods (Cheng, 2013). Simple statistical methods and visual displays can reveal observations that are rare, suspicious, or different from the vast majority of the data. Statistical learning methods are analytical techniques applied to data that allow one to draw conclusions about the data. Specifically, machine learning uses statistical algorithms to predict future outcomes. Machine learning searches for distinguishing patterns that characterize subsets of the data. Typically, a machine-learning analysis partitions the data into a test set and a training set. The training set is used to develop the predictor, and the test set is used to evaluate its performance. We briefly describe each type of machine learning below.

With SML, predictions are made using a defined outcome variable. A typical data set used in SML has many explanatory variables. In a descriptive data set such as the USASpending data, some variables may be used as explanatory variables in one analysis and play the role of outcome variables in another analysis. An SML predictor learns combinations of the explanatory variables that effectively predict the outcome. The goal is to reduce the prediction error to an acceptable level, as measured on the test set. Examples of SML techniques are artificial neural networks, random forests, and support vector machines (James, Witten, Hastie, & Tibshirani, 2013).

With UML, there is no outcome variable. Instead, UML seeks to find clusters or subgroups of the data that may indicate anomalous behavior. In this thesis, we do not use UML techniques. For a discussion of this topic, we refer the reader to Celebi and Aydin (2016).

Random forests is a SML classification or regression technique consisting of a voting system applied to collections of decision trees. Faraway (2016) gives a detailed description of random forests. Whereas a decision tree consists of only one tree in order to predict classification, a random forest builds multiple decision trees and averages classification or regression predictions for a more accurate prediction. Unlike decision trees that search for essential features to create nodal splits, random forests create multiple trees based on bootstrapped samples and use randomly selected features to create the nodal splits.

We use the machine-learning technique known as random forests (Faraway, 2016) to fit classification and regression models. Software for fitting random forest models can be found in many software environments. We use the implementation provided by the randomForest command in the randomForest package (Liaw & Wiener, 2002) that is available in the R software environment (R Core Team, 2014).

Although the randomForest implementation can handle both categorical and continuous predictor variables, it restricts categorical variables to have no more than 32 levels. For our purpose, this limit is too restrictive. For example, in the data that we consider, there are more than 3,131 unique recipients of Army contracts and 258 awarding offices. These variables must be aggregated to have smaller numbers of categories to use this software. We do this by leaving the most frequent categories intact and grouping the remainder. Natural splits in frequencies for recipients and awarding offices produce new categorical variables with 29 levels for awarding office and 19 levels for recipients.

1. Response Variables

Previous work on anomaly detection highlights the importance of variable and model selection to improve accuracy (Rouillard, 2003; Monteiro, 2002). Although anomalies differ depending on data sources and problem sets, we adopt ideas from these frameworks into our approach. The case studies discussed in Chapter II, as well as the agency initiatives to promote best spending practices, motivate the selection of outcome variables to use in random forests. Our response variables pertain to the areas of solicitation, competition, and set aside programs. Our models for categorical outcomes requires the response variable to be binary. The federal action obligation is appropriate for a response variable in our continuous model. In Table 1 we define the following variables, which can be derived from the USASpending data for monitoring awards and spending practices.

Table 1. Response Variables

Response Name	Definition
SetAside	Binary response variable signifying an award is either portioned for a set-aside company or not.
Non-commercial	Binary response variable signifying if the service or product manufactured is available commercially versus not being available commercially
Sole-Source	Binary variable signifying whether the contract was awarded through sole-source methods or not
Multi-year	Binary response variable signifying if the contract duration was longer than one year or not
Non-competed	Binary response variable signifying if the award was competed for under full and open competition or not.
Small	Binary response variable signifying the classification of the recipient as a small corporation or other than a small corporation
Award Amount	The total amount spent on an individual award

2. Trend Modeling

Our final subset of data set contains 73,570 awards. There are 97 different awarding offices and 1,238 unique recipients. We build a data set to examine each response variable separately. Table 2 includes the predictor variables selected along with their definitions.

Table 2. Categorical Random Forest Data Predictor Variables

Predictor Variable	Description
Action Date	A date object signifying the beginning of a contract.
Office	A categorical object consisting of the 29 most prevalent awarding offices as individual categories.
NAICS	The North American Industry Classification System designation code.
LogAmt	The logarithmic total cost of the contract..
Set Aside	A binary response variable signifying an award is portioned as a set-aside.
Non-commercial	A binary response variable signifying if the service or product manufactured is available commercially.
Sole-Source	A binary variable signifying whether the contract was awarded as sole-source.
Multi-year	A binary response variable signifying if the contract duration was longer than one year or not.
Non-competed	A binary response variable signifying if the award was competed for under full and open competition standards.
Small	A binary response variable signifying the classification of the recipient as a small corporation or other than a small corporation.

We build seven random forest models, one for each response variable and plot separate partial-like plots for each response random forest object. The plots consist of the estimated probability of categorization of an award as a response variable as a function of time for a fixed awarding office.

For an outcome variable that has a binary outcome, such as whether or not an award was made on a non-competed solicitation, we examine the awarding offices in two respects: how unusual the number of such awards is relative to other offices and trends over time in making such awards. Both analyses are based on the estimated probabilities of awards belonging to a category (e.g., *Non-Competed*) obtained from a random forest model that we fit to Army contract data obtained from USASpending for the years 2013 to 2018.

For trend analysis, we consider only those awarding offices that contribute at least 50 percent of the awards, as we use the awarding office as an explanatory variable in the model. The 29 most frequently occurring offices, plus an “Other” category for all others combined, constitute a categorical variable with 30 levels. The random forest model predicts a binary outcome variable where “success” implies a contract belongs to a particular attribute category. Other explanatory variables include the date of an award, its NAICS code, and the amount of the award on a logarithmic scale. In order to compare awarding offices under a common set of conditions, we obtain estimated probabilities across all awards that were made holding the awarding office and date fixed, and then average the probabilities. This process is repeated for the awarding office, varying the date over a set of values that span a period of time. The averaged probabilities can be plotted against time to detect trends, along with a plot of probabilities that are averaged across all awarding offices so that the awarding office in question can be compared to the average behavior of its peers.

3. Outlier Modeling

We use the continuous random forest model to find outliers in the spending patterns of awarding offices. We build the data for our continuous model in the same manner as the categorical models, but leave all response variables as predictors and use the logarithmic amount of each award as the outcome variable. Table 3 lists the predictor variables of the continuous random forest object.

Table 3. Random Forest Predictor Variables Used to Model Spending

Predictor Variable	Description
Date	A date object signifying the beginning of a contract.
Office	A categorical object consisting of the 29 most prevalent awarding offices as individual categories, and all others in a group named OTHER.
NAICS	The North American Industry Classification System designation code.
Recipient	A categorical object consisting of the 19 top most prevalent recipients as individual categories, and all others in a group named OTHER.
Set Aside	A binary response variable signifying an award is either portioned for a set aside company or not.
Non-commercial	A binary response variable signifying if the service or product manufactured is available commercially versus not being available commercially.
Sole-Source	A binary variable signifying whether the contract was awarded through sole-source methods or not.
Multi-year	A binary response variable signifying if the contract duration was longer than one year or not.
Non-competed	A binary response variable signifying if the award was competed for under full and open competition standards or not.
Small	A binary response variable signifying the classification of the recipient as a small corporation or other than a small corporation.
Awarding Office	The actual awarding office code corresponding to particular awarding offices

We plot boxplot attributes including the lower quartile, median, and upper quartiles of federal action obligations after sorting them by the median. Since the number of contracts for a given awarding office affects box plot attributes, we only consider awarding offices having 100 or more contracts. For our purpose, an upper outlier is a contract with a measurement that exceeds $Q_3 + 1.5 \times IQR$, where Q_3 is the upper quartile and IQR is the interquartile range. The interquartile range, which is the third quartile minus the first quartile, often is used to measure variability (Devore, 2016). Although boxplot rules also can be used to identify outliers at the lower end, our interest is to identify instances of spending that are unusually high. Along with the plot, we present contracts flagged as outliers by sorting them in descending order by $\frac{X - Q_3}{IQR}$ where X is the logarithm of the contract amount.

4. Analysis of Awarding Offices by Contract Attributes

The trend analysis uses the properties of the random forest model to compare awarding offices to each other under common conditions. Each awarding office, however, varies in the attributes of awards that it makes, which are not reflected in the trend analysis. To examine these awards individually, we use a random forest model that excludes the awarding office as an explanatory variable. This also allows us to examine smaller awarding offices that are combined in the "Other" category in the trend analysis. We do this with respect to each of the binary outcome variables that we use in the trend analyses.

To examine a specific awarding office, let n denote the number of awards that are made by that office over the relevant period, and let X denote the number of those awards that belong to a category of interest, to which we refer as a "success" (e.g., non-competed contracts). An award belonging to the success category is treated as a random event, independent of others, with a probability of success that varies due to the conditions of the contract (e.g., the items or services purchased, size of the award, etc.). Let $\pi_1, \ldots, \pi_n$ denote the probabilities of success for each award. The following expressions for the expected value and variance of X are obtained below:

$$\mu = E(X) = \sum_{i=1}^{n} \pi_i, \ \sigma^2 = Var(X) = \sum_{i=1}^{n} \pi_i (1 - \pi_i) \qquad (1)$$

Suppose that the awarding office makes x awards in the success category: to what extent is that an unusually large (or small) number according to the assumptions that we have adopted? Adopting an hypotheses-testing framework, we measure this by calculating either an upper- or lower-tail p-value as follows:

$$\begin{aligned} P_{\text{High}} &= P(X \geq x) = \left[\sum\nolimits_{k=x}^{n} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \prod\nolimits_{j=1}^{k} \pi_{i_j} / (1 - \pi_{i_j}) \right] \prod\nolimits_{l=1}^{n} \pi_l \\ P_{\text{Low}} &= P(X \leq x) = \left[\sum\nolimits_{k=0}^{x} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \prod\nolimits_{j=1}^{k} \pi_{i_j} / (1 - \pi_{i_j}) \right] \prod\nolimits_{l=1}^{n} \pi_l \end{aligned} \tag{2}$$

Which p-value is used depends on whether it is an unusually large, or small, number of awards in the success category that warrants investigation. A large number of awards made to recipients in set-aside programs, for instance, may increase the costs of goods or services, which suggests that P_{High} is useful for identifying awarding offices that merit scrutiny with respect to that attribute. Similarly, a small number of competed awards is of concern due to higher costs and the opportunities for irregular behavior that non-competed awards may present, which suggests that P_{Low} is an appropriate metric for that attribute.

Unless n is small, the p-values shown in (2) are difficult to calculate exactly. A normal approximation based on (1) can be used but the conditions to justify doing so are difficult to assess and are not likely to be met in many instances. Instead, we use simulation to approximate the p-values. This is done by generating a large number B of collections of n independent Bernoulli random variables with probabilities $\pi_1, \ldots, \pi_n$ and saving their sum as a realization of the random variable X. We then calculate the relative frequency of the event $X \geq x$ or $X \leq x$ as an approximation to the desired p-value. In our investigation, we use $B = 10{,}000$ simulations for each p-value that we approximate.

We make several important observations about the use of the p-values described above for identifying awarding offices that may warrant more detailed investigations:

1. *The use of estimated probabilities for calculating p-values*. In practice the probabilities $\pi_1, \ldots, \pi_n$ are unknown: they are estimated using a random forest model. Using

estimated probabilities imparts additional uncertainty, which we assume is negligible due to a large number of contracts that comprise our sample.

2. *Validity of the independence assumption*. Underlying all of our models is the assumption that the disposition of each contract is a random event that is independent of all others. That this assumption is strictly satisfied is questionable for several reasons. First, the same estimated model is used to produce estimated probabilities if for no other reason that an awarding office may award multiple contracts to the same recipient. The likely effect of this type of dependence is to make the calculated p-values smaller than they should be. If the p-values are used in a strict hypothesis-testing framework a higher than expected rate of false positives may result. For this reason, we recommend that the p-values be interpreted conservatively, with a “close-call positive” not necessarily being treated as a positive result.

3. *Multiple testing and false positives*. Using the p-values to screen a large number of awarding offices increases the chance that at least one of them will be flagged as a false positive. We, therefore, recommend that the p-values be adjusted for multiple testing to manage the rate of false positives. Specifically, we recommend that a Benjamini-Hochberg approach be used (Faraway, 2015) which controls the false-discovery rate. It is less conservative than the Bonferroni or Holm method which control the family-wide Type I error rate (Faraway, 2015). A less-stringent approach is reasonable if the cost of having a controllably small proportion of false positives is not substantial, which we address below.

In any anomaly-detection scenario, the consequences of false positives must be taken seriously. Generating an investigation of actors in the Army’s procurement chain on the basis of statistical evidence can be costly. Moreover, if it turns out that the investigation was not warranted, there is a risk that the investigators will come to distrust statistical evidence. We believe that the proper use of our tools is to point investigators to a subset of awarding offices that warrant a deeper, “second-stage” information gathering that goes beyond what was supplied by the USASpending data, without suggestion that improprieties have been uncovered by the tools. Our recommendation that a Benjamini-Hochberg adjustment of p-values be used concedes that a small number of awarding offices that are flagged using the tools will be revealed not to have warranted scrutiny at the second stage.

Thresholds for making discoveries at the first stage should be set so that the cost of false discoveries is acceptable.

IV. RESULTS AND ANALYSIS

The objective of this thesis is to propose a set of statistical methods to identify DoD contracts or spending practices that may justify further scrutiny. We offer a suite of tools that measure anomalous behavior for DoD awarding offices. These tools facilitate monitoring trends and spending patterns over time, and anomaly detection. Table 1 of Chapter III.B.1 lists the response variables chosen. We build analytical tools using continuous, categorical, and p-value comparisons for each response variable as shown in Table 4.

Table 4. Analytical Tools

Random Forests		P-Value Comparison Categories
Continuous Object Responses	Categorical Object Responses	
LogAmt	Set Aside	Set Aside
	Non-Commercial	Non-Commercial
	Sole-Source	Sole-Source
	Multi-Year	Multi-Year
	Non-Competed	Non-Competed
	Small	Small

Our analysis shows the utility of individual model results by identifying baseline conditions for normal DoD contractual behavior and anomalies or trends present according to our criteria. There may be natural explanations for anomalies or irregular trends shown in our results such as: an awarding office may have unusually large amounts of *sole-source* contracts because they service companies who specialize in the distribution of a product which no other company can make.

A. SPENDING PATTERN ANALYSIS

Recall that the residuals come about from fitting a response model using random forests. We extract the residuals for each awarding office and identify upper outliers only. Lower outliers are not included in this study, as the higher spending contracts correlate with our case study research in chapter II.D.

Spending pattern analysis examines residuals produced by the random forest object for individual awarding offices. Baseline conditions regarding federal action obligations are averages for each awarding office. Figure 5 displays a plot of observations that may warrant further scrutiny with respect to federal action obligation amounts as well as the 25th, 75th, and 50th percent quartiles.

Observations exceeding 1.5 times the IQR are indicated by orange circles increasing in size—dependent on their distance from the IQR.

Figure 5. Random Forest Spending Plot of Boxplot Attributes for Amounts of Awards by Awarding Offices

We sort anomalous observations for individual awards by the distance from the upper inner fence of the observations denoted as: $\{(Q_1 - Q_3) * 1.5\} + Q_3$, with the most extreme at the top of the list, and where Q_1 and Q_3 are the 25th and 75th percentiles, respectively. Our model produces 286 anomalous observations out of 73,570 observations, from 236 awarding offices. The outliers account for less than 0.1% of all of the observations. Table 5 lists the top ten outliers ranked by extremity and contains a select amount of descriptor variables from the USASpending data.

Table 5. Top Ten Contract Awards by Outlier Extremity

Parent Award ID	Award ID PIID	Award Office	Amount (M)	Date
GS10F0333Y	W91QEX14F1B23	W91QEX	$5.96	8/28/2014
GS10F0333Y	W91QEX14F1B23	W91QEX	$6.65	7/10/2015
GS10F0333Y	W91QEX14F1B23	W91QEX	$6.71	8/1/2016
HHSN316201200036W	W91QVN15F0029	W91QVN	$7.02	5/4/2018
W91QVN08D0014	5001	W91QVN	$1.67	10/1/2012
W91QVN13D0059	25	W90VN9	$4.68	12/15/2014
W91QVN13D0059	51	W90VN9	$4.83	12/27/2015
W91QVN13D0059	87	W90VN9	$4.80	12/22/2016
W9127813D0025	DH01	W912DQ	$5.08	9/18/2013
W912DQ11D3003	4	W912DQ	$6.60	9/16/2015

Our model allows for the display of any descriptor variable within the USASpending data such as the awarding office names, recipient names, and financing office names. Anomalies for any actor such as the recipient or funding office are attainable with minor modifications to our model.

B. TREND ANALYSIS

Our models for categorical outcomes identify trends regarding set-aside and sole-source designation and competition metric categories. Trends provide insights that may warrant further scrutiny and depict actor behavior. Table 1 in Chapter III.B.1 lists the response variables for our models, and Table 2 in Chapter III.B.2 lists the explanatory variables for our models. Individual models were chosen based upon their visually identifiable and significant trends. Each output shows the trends for our response variables separately.

We determine baseline conditions for our models from average cases by varying the date and leaving all other variables the same. Each plot shows residual trends in black for each of the 29 most frequently occurring awarding offices as described in Chapter III, and averages across all awarding offices in red. Like the spending pattern analysis, our focus remains on the awarding office. Analysts may want to determine if an awarding office is awarding more set-aside contracts than they have in the past. They also want to analyze the effectiveness of federal agency initiatives obtained by viewing trends regarding the procurement of commercially available items. Our models provide useful insights into the spending practices of individual awarding offices. In this section, we give example outputs of each model with definitions and propose analytical importance.

1. Monitoring Set-Aside Awards

The *Set Aside* model shows frequency trends of set-aside contracts for an awarding office through time and against the average frequency of all awarding offices. Set-aside awards typically cost more for similar services or goods than non-set asides (Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, 2014). The use of these programs, therefore, is a component of an awarding office's spending practices. Figure 6 provides an example of how this information can be used to monitor an awarding office. Estimated probabilities of the awarding office W25G1V to confer a set-aside award as a function of time (black line) are consistently higher than average (red line).

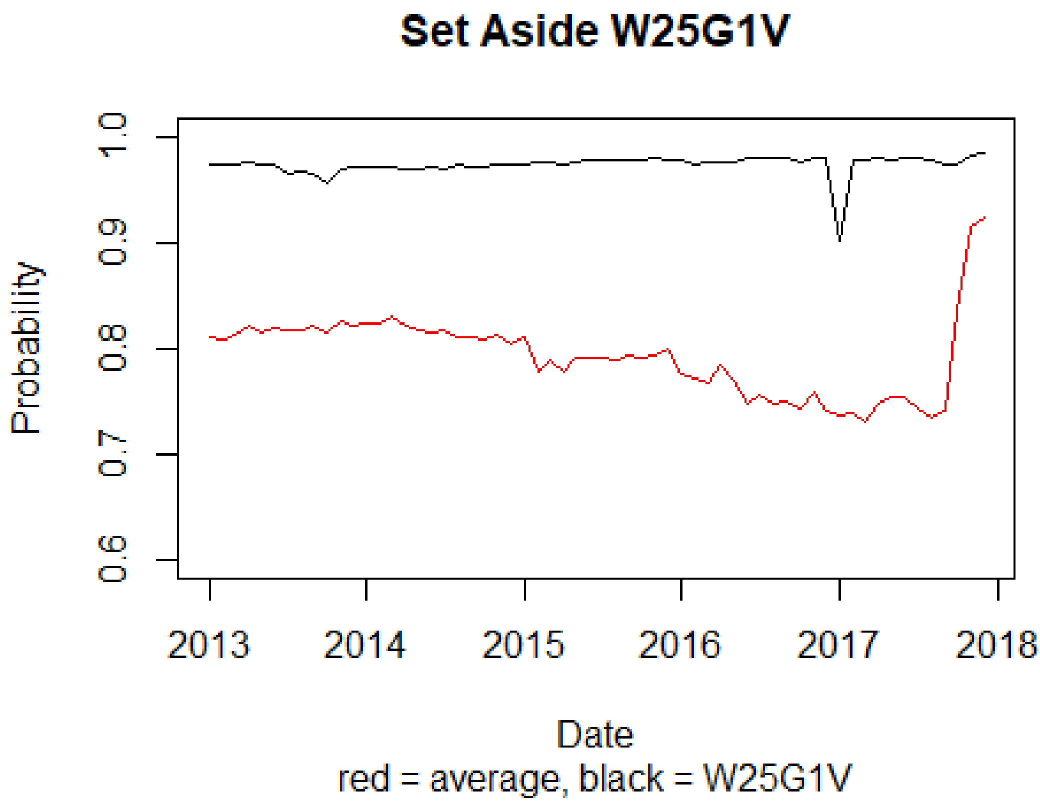

Figure 6. Categorical Random Forest Model Output of the Probability of a Set-Aside Award (Black) for the Awarding Office W25G1V versus the Average Probabilities of all Awarding Offices (Red), across Time

2. Monitoring Multi-year Awards

The *Multi-Year* response variable is a binary outcome for which "success" implies that a contract is active for more than one year. Trends regarding multi-year awards lend additional insight to an awarding office's spending patterns. Analysis of the number of delivery orders against a parent award ID may show a series of small federal action obligation amounts. A parent award lasting longer than a year and containing multiple delivery orders for small federal action obligation amounts may indicate that estimated quantities in the original contract were understated. It also could indicate a deliberate splitting of contract requirements to sustain federal action obligation amounts below certain review or approval levels (Publications Branch, 1993). Figure 7 provides an example of how the *Multi-Year* response variable may be monitored. For awarding office W25G1V, we see a slight downward trend of multi-year contracts (black line) that is slightly less than the average of all awarding offices (red line).

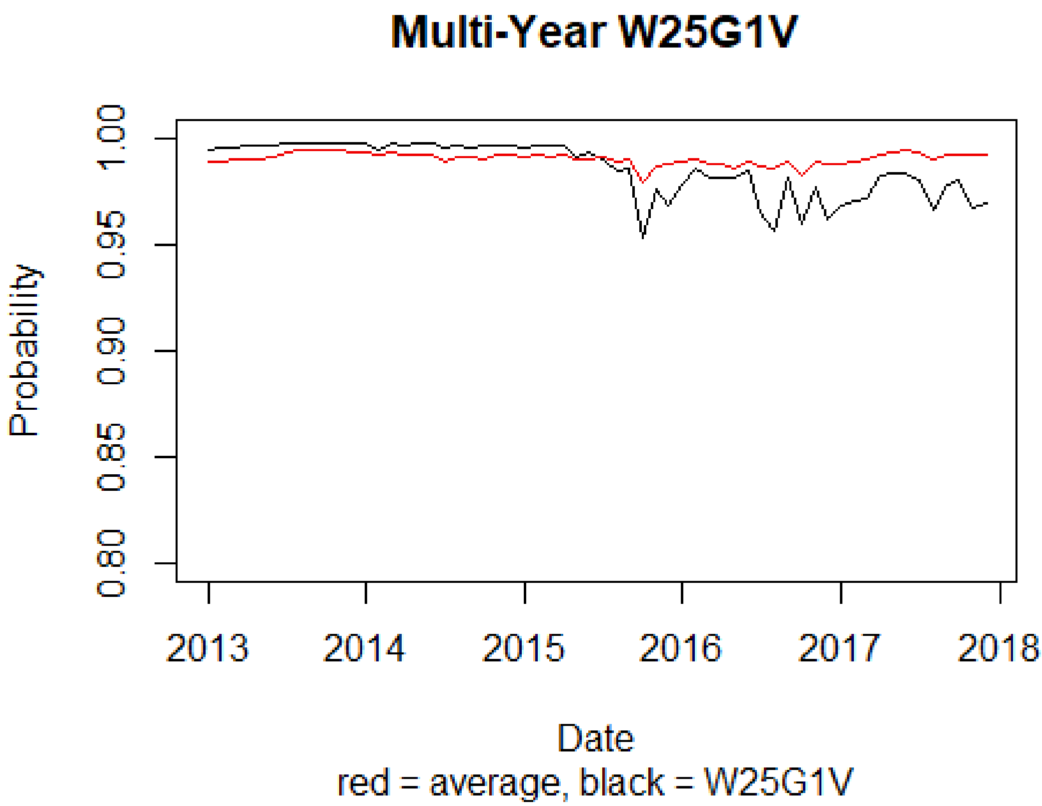

Figure 7. Categorical Random Forest Model Output of the Probability of a Multi-year Award (Black) for Awarding Office W25G1V versus Average Probabilities of All Awarding Offices (Red), across Time

3. Monitoring Non-commercial Awards

A commercial award is one where the item or service purchased is available without modification from the marketplace. The purchase of commercial items is beneficial due to competition in the marketplace leading to lower prices. The *Non-Commercial* response variable is a binary outcome where "success" implies that the item or service purchased is not commercially available. To illustrate the monitoring of this variable, Figure 8 shows estimated probabilities for awarding office W25G1V (black line) being substantially higher than the average (red line) with the distance between the two lines increasing over time.

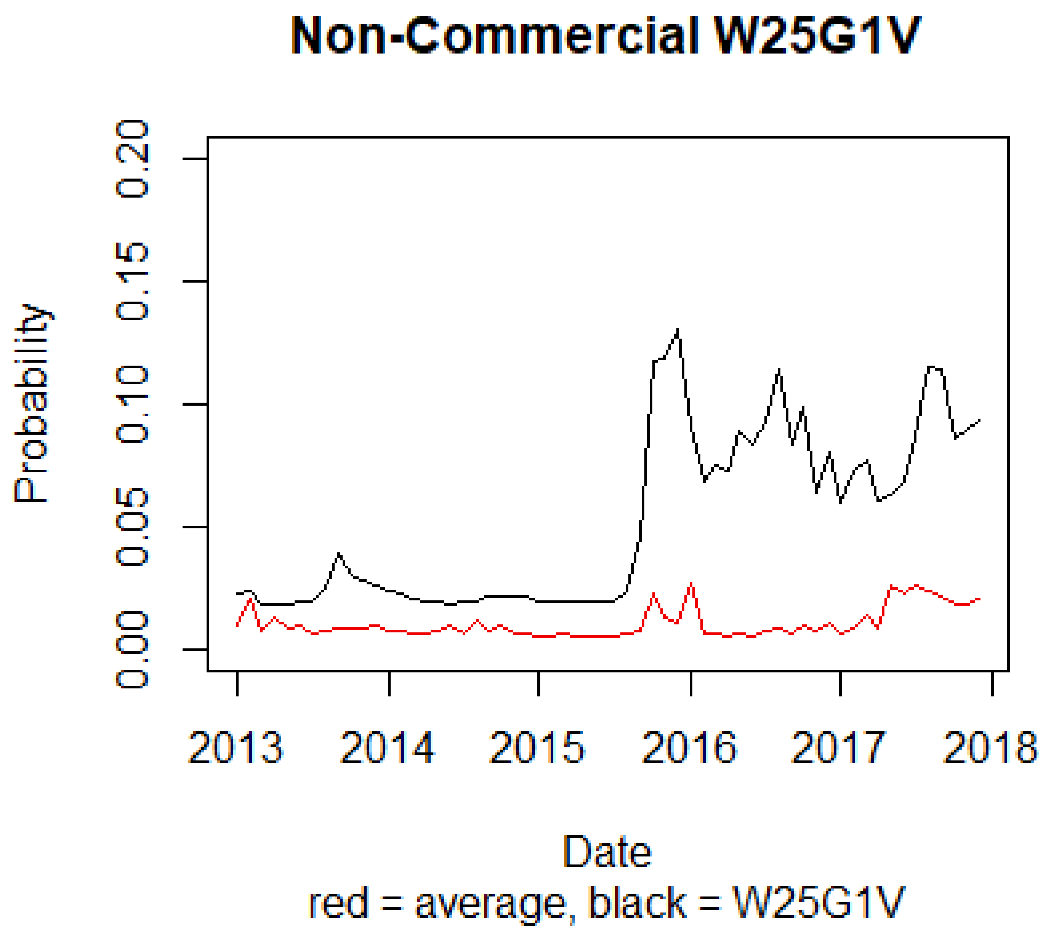

Figure 8. Categorical Random Forest Model Results for the Probability of the Items or Services Purchased Are Commercial (Black) for the Awarding Office W25G1V versus Average Probabilities of All Awarding Offices (Red), across Time

4. Monitoring Sole-Source Awards

A sole-source award is one where the item or service is purchased from a single recipient. Monitoring of sole-source awards directly lends insight to possible insider threats and fraudulent behavior as discussed in Chapter II.D. Awards supplied to a sole-source recipient cost more on average than those fielded in a competitive environment, with average growth rates of cost 57 percent higher than that of regularly competed awards (Wandland & Wickman, 1993). Trends regarding sole-source award frequencies provide insight into spending patterns and competitive environments for awarding offices. Figure

9 shows the estimated probabilities for awarding office W25G1V (black line) being substantially higher than the average (red line) with the gap widening with respect to time.

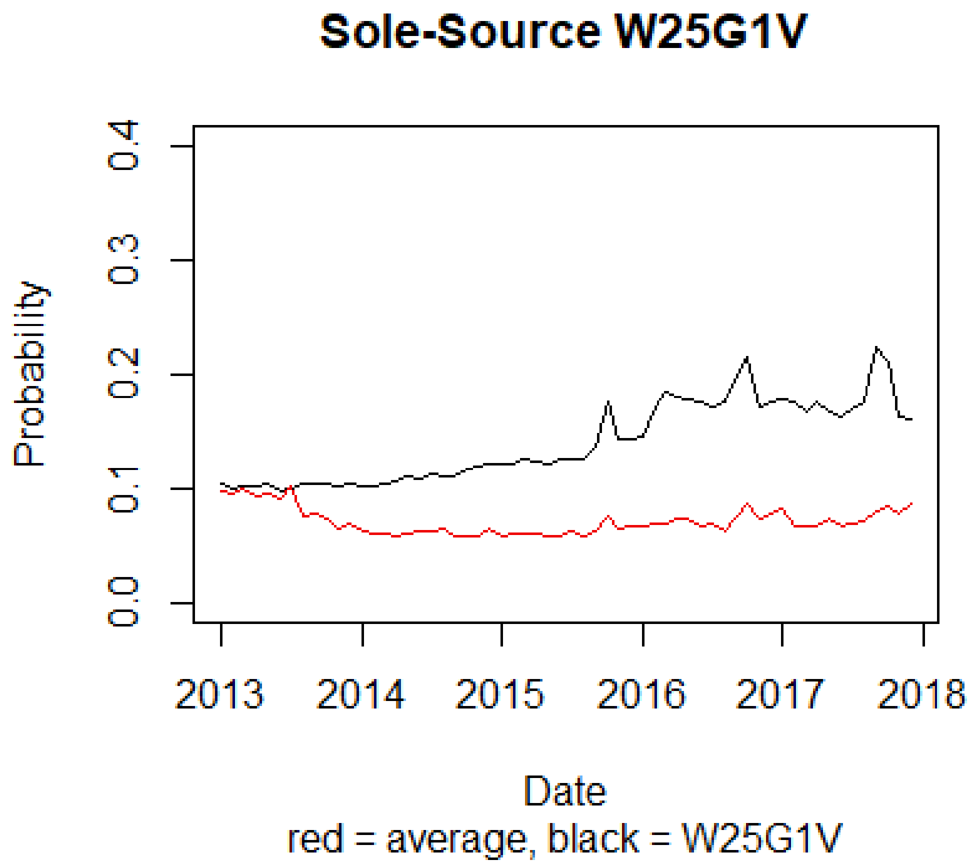

Figure 9. Categorical Random Forest Model Results for the Probability of a Sole-Source Award (Black) for Awarding Office W25G1V versus Average Probabilities of All Awarding Offices (Red), across Time

5. Monitoring Non-competed Awards

The *Non-Competed* response variable is a binary variable indicating "success" as an award that is competed for under full and open competition standards as defined in FAR 25.103. Figure 10 is an illustration that shows the estimated probabilities for awarding office W91ZLK (black line) increasing to a substantially large distance from the average of all awarding offices (red line) through time.

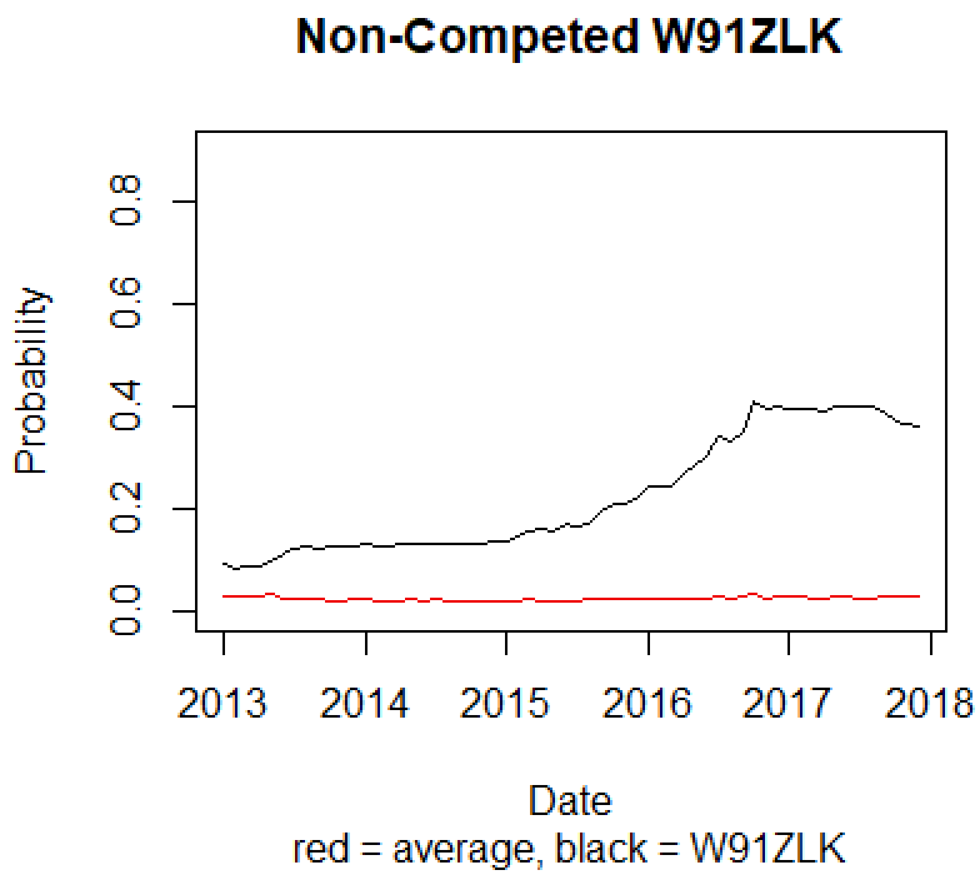

Figure 10. Categorical Random Forest Model Results for the Probability of a Fully and Openly Competed Award (Black) for Awarding Office W91ZLK versus Average Probabilities of All Awarding Offices (Red), across Time

6. Monitoring the Designation of Small Recipients

The outcome variable *Small* indicates whether the recipient is a small business as designated by the contracting officer. It provides a unique perspective on whether or not an awarding office is knowingly awarding contracts to big or small businesses. The U.S. SBA determines, on an annual basis, if a business qualifies as small by considering its number of employees and tax return receipts for goods or services. Individual NAICS categories have unique limits defining small businesses within each market. Businesses may change in size during the year and prior to tax season. The contracting officer's determination of business size ultimately restricts or allows usage of set-aside programs pertaining to small businesses. Figure 11 shows that the estimated probabilities of awarding office W912BU (black line) increase with respect to time between the fiscal years 2016 and 2017 distancing themselves from the average of all awarding offices (red line).

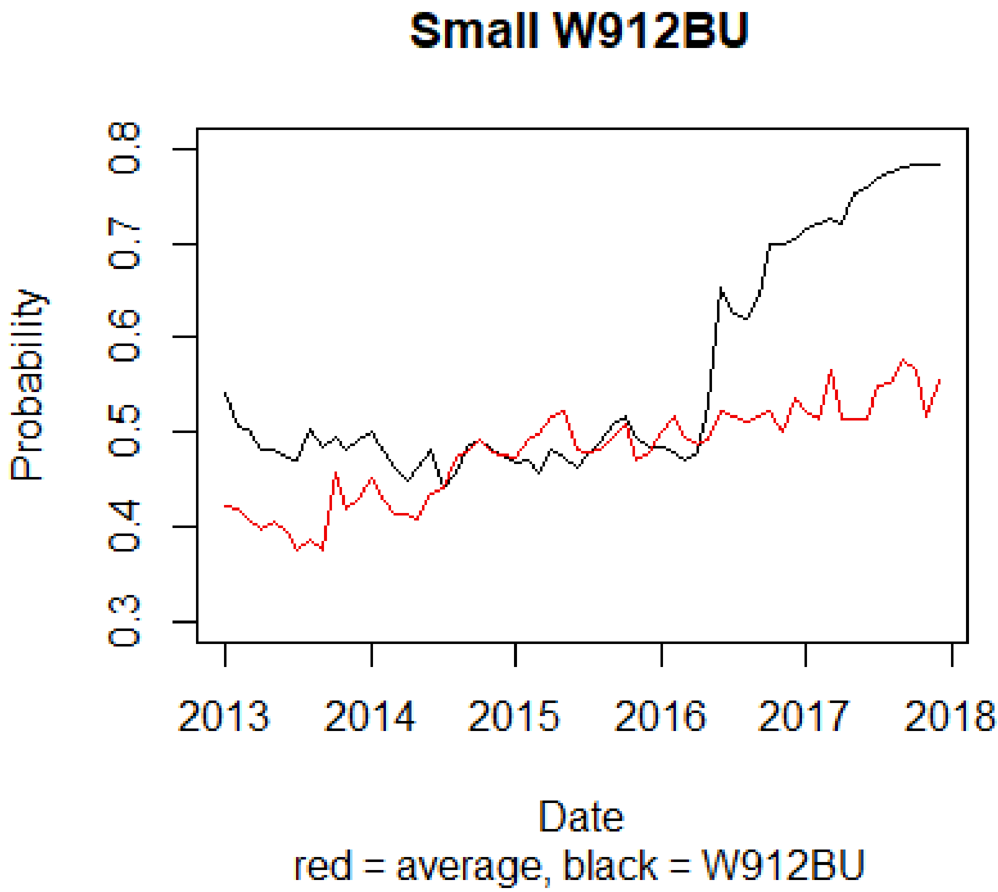

Figure 11. Categorical Random Forest Model Results for the Probability of a Contracting Officer Categorizing a Business as Small (Black) for Awarding Office W912BU versus Average Probabilities of All Awarding Offices (Red), across Time

7. Recipient Focus

The SBA sets standards for the classification of small businesses (e-CFR, Electronic Code of Federal Regulations, 2019). These standards are specific for each NAICS market category. Table 6 displays the NAICS codes that we consider, along with each market's cut-off value regarding the classification of a small business in millions of dollars.

Table 6. NAICS Codes, Descriptions, and the SBA Lower Limit of Consideration for Small Business Designation for FY 2018

NAICS	Description	$ in M limit
541330	Engineering services	15
517110	Wired telecommunications carriers	22
541519	Other computer related services	27.5
541990	All other professional, scientific, and technical services	15
541611	Administrative management and general management consulting services	15
541511	Custom computer programming services	27.5
541512	Computer systems design services	27.5
541690	Other scientific and technical consulting services	15
541614	Process, physical distribution, and logistics consulting services	15
561990	All other support services	11
611420	Computer training	11
811212	Computer and office machine repair and maintenance	27.5
541513	Computer facilities management services	27.5
518210	Data processing, hosting, and related services	32.5
811219	Other electronic and precision equipment repair and maintenance	27.5
561110	Office administrative services	7.5
811412	Appliance repair and maintenance	15
517210	Wireless telecommunications carriers (except satellite)	22
519190	All other information services	27.5
541219	Other accounting services	20.5
517410	Satellite telecommunications	32.5
541618	Other management consulting services	15

If a business's annual revenue falls below this cutoff, then it may apply for set-aside programs associated with small businesses for a corresponding NAICS. Interesting relationships emerge between the contracting officer's determinations of business size and recipients' annual revenue as reported in the USASpending data. Figure 12 shows a sharp downward trend in estimated probabilities for recipient 806026852 (black line) being designated as small, falling below the average of awarding offices (red line) with respect to time.

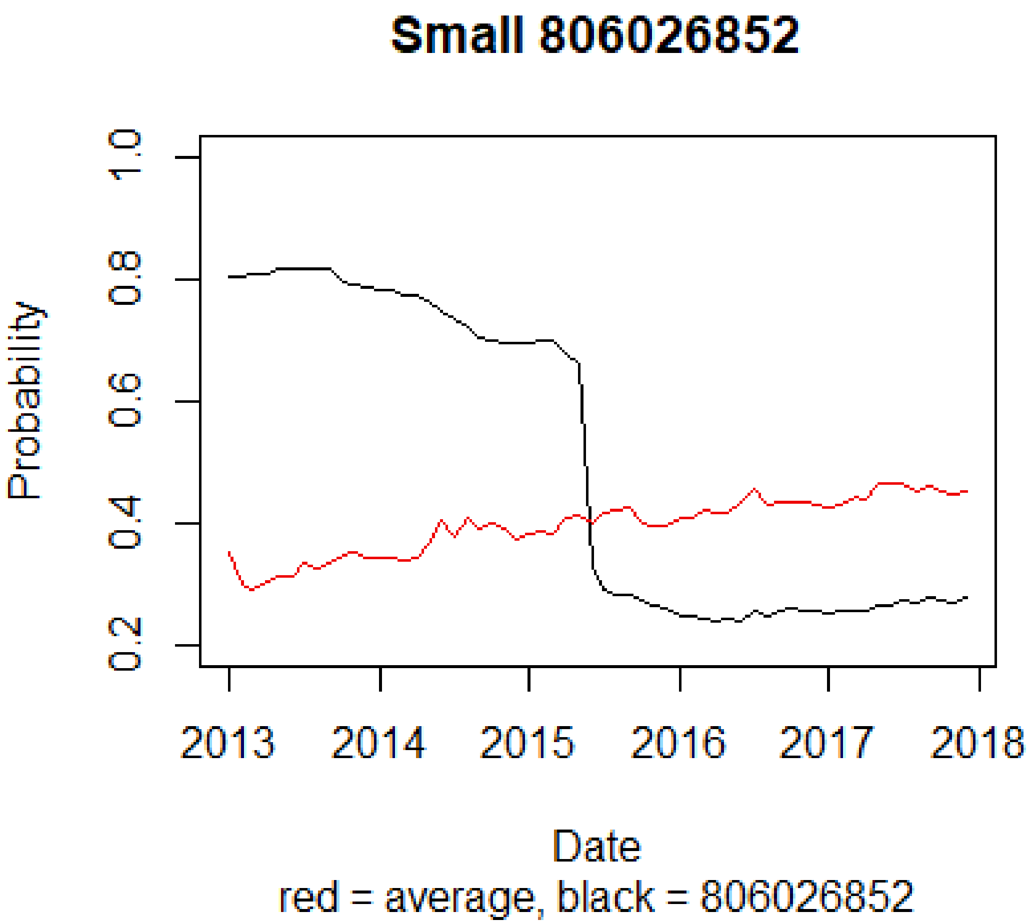

Figure 12. Categorical Random Forest Predicted Residuals of a Recipient's Determination of Small Business (Black) Plotted against All Other Recipient Averages (Red)

Figure 13 shows the annual revenue for recipient 80602685 through time.

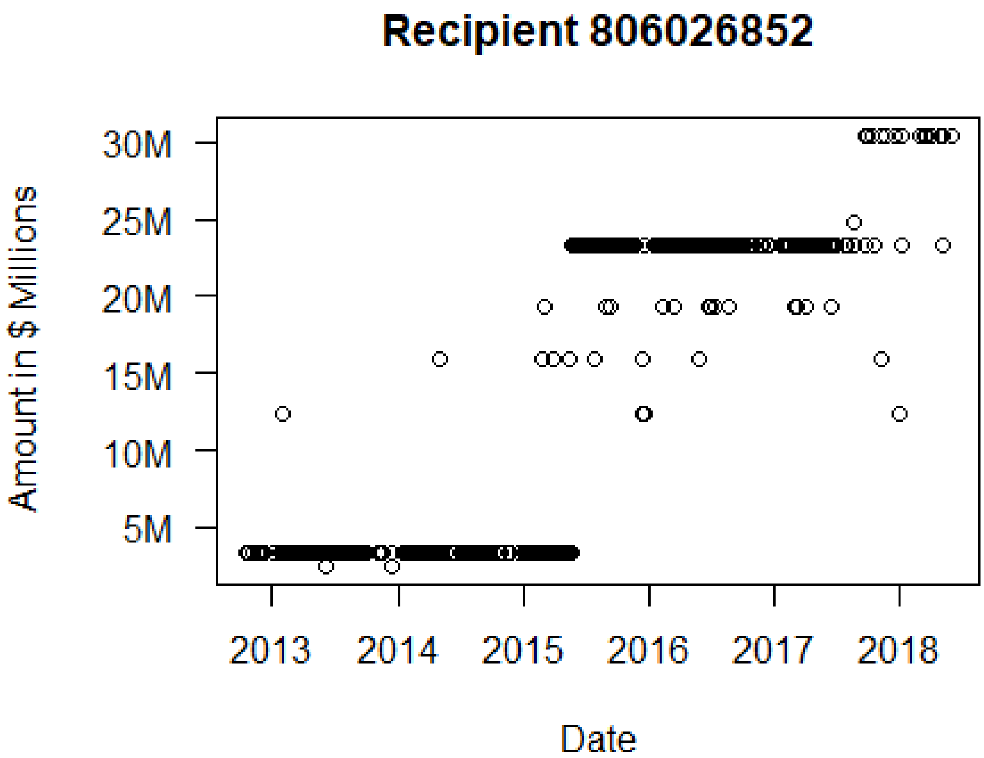

Figure 13. Amount of Annual Revenue for Recipient 806026852 in Millions of Dollars, through Time

Comparison of Figure 12 with Figure 13 explains the reasoning behind the sudden shift in designation of the recipient as a small business. During FY 2015, a large gap in annual revenue occurs shifting from under $5 M to approximately $25M regarding annual revenue. This sudden growth in revenue prompted contracting officers to remove the designation of a small business for this recipient.

C. ANALYSIS OF P-VALUES

Our trend analysis uses the properties of the random forest model to compare awarding offices to each other under common conditions. Our trend analysis models are fit to the categorical responses as shown in Table 1 of Chapter III.B.1. Plots of these trends for each awarding office provide a means of visual identification of anomalous behavior, but due to the volume of results, it is difficult to compare all the offices together in a single instance. To examine awards individually, we use a random forest model that excludes the awarding office as an explanatory variable as described in Chapter III.B.4. This allows us to examine awarding offices that are combined in the "Other" category and provides a means to numerically, rather than visually, compare awarding office trends. Table 7 shows the first ten awarding offices sorted in descending order of their relative difference (Rel Diff) calculated by $\left(\frac{SetAside-Expected}{Expected}\right) \times 100$, where N is the number of contracts for the awarding office, SetAside is the number of contracts categorized as a *SetAside*, Expected is the number of contracts predicted based on the random forest model. The P.value is the probability of an awarding office having the number of *SetAside* contracts for the Expected amount. The P.BH value is the upwardly adjusted P.value amount for multiple testing as described in Chapter III.B.4.

Table 7. Top Ten Awarding Offices by Extremity of the Number if Set-Aside Awards Made

AwardingOffice	N	SetAside	Expected	Diff	Rel Diff	P.value	P.BH
W56KGU	58	56	30.24	25.76	85.20%	< 0.0001	< 0.0001
W9124V	137	119	82.67	36.33	43.90%	< 0.0001	< 0.0001
W9125F	38	38	28.15	9.85	35.00%	< 0.0001	< 0.0001
W912P7	174	159	125.7	33.30	26.50%	< 0.0001	< 0.0001
W912DS	451	341	282.54	58.46	20.70%	< 0.0001	< 0.0001
W91B4N	19	18	14.94	3.06	20.50%	< 0.001	< 0.01
W91WRZ	37	36	30.62	5.38	17.60%	< .0001	< 0.001
W9124A	46	44	37.65	6.35	16.90%	< 0.001	< 0.001
W912DW	256	172	147.92	24.08	16.30%	< 0.0001	< 0.0001
W91243	90	84	72.26	11.74	16.20%	< 0.0001	< 0.0001

D. COMBINED APPROACH

Combining our tools provides detailed analysis regarding the anomalies of an awarding office. We suggest two methods for using our tools illustrating different deliberate methods to obtain results based on an analyst's interest. Interests may vary amongst analysts; some may be interested in only examining commercial item acquisition trends. Others may want to explore if correlation occurs between the *sole-source* and *non-competed* variables. All of our previous examples in chapter IV relay the individual value of each output. However, combining our outputs may lend relevancy in distinguishing related causality between outputs, and provide a more detailed analysis of individual actors.

1. Random Selection

In the first method, analysts choose a response variable to inspect at random. Identification of an awarding office with relatively high probabilities respective to a response average is easily identifiable, as visual representations of these metrics are available from the plots. We chose to inspect *SetAside* trends. The estimated probabilities for awarding office W25G1V (black line) remain substantially higher than the average of all awarding offices (red line) as shown in Figure 14.

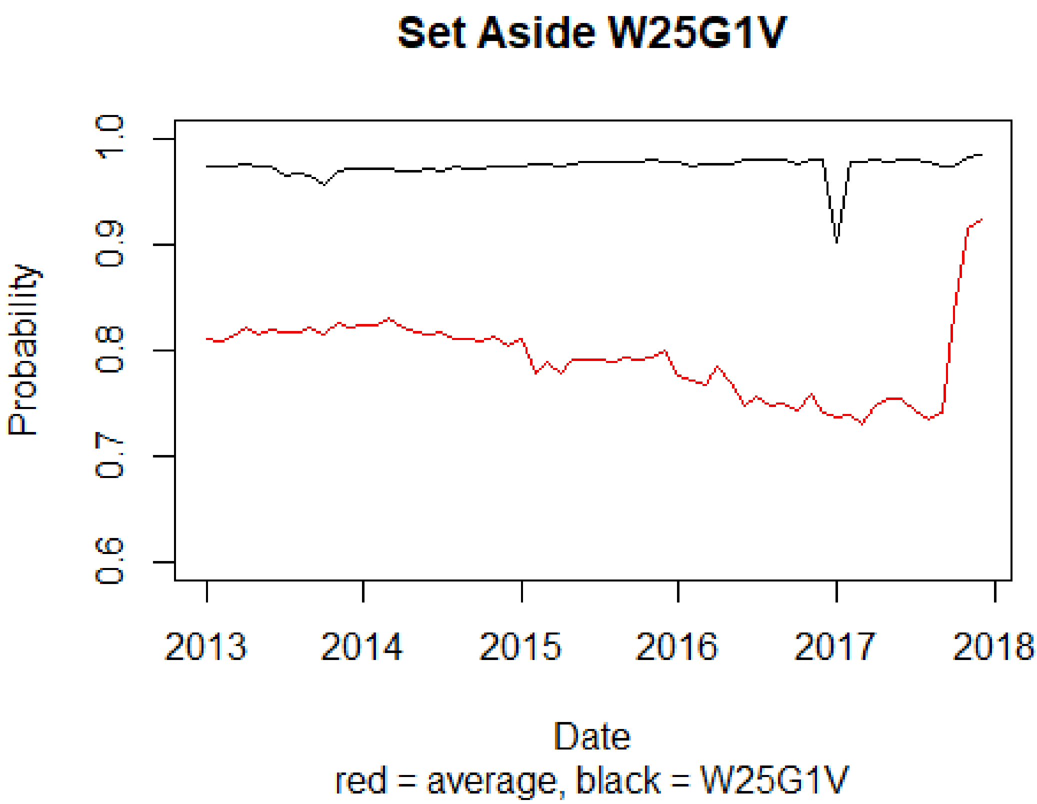

Figure 14. Categorical Random Forest Model for the Probability of a Set-Aside Award (Black) For the Awarding Office W25G1V versus the Average Probabilities of All Awarding Offices (Red), over Time

A trend of high amounts of awards designated as *SetAside* can occur naturally for awarding offices. However, our motivation remains on identifying anomalous behavior respective to normal DoD spending patterns, and the substantially large difference between the average of all awarding offices and awarding office W25G1V may warrant further investigation prompting us to identify if awarding office W25G1V is in our spending pattern analysis output. Awarding office W2G1V has 16 federal action anomalies in our spending pattern analysis output shown in Table 8.

Table 8. Anomalies for Awarding Office W25G1V Expressed in Thousands of Dollars

Parent ID	Award ID	Award Office	Total
W91QUZ07D0010	ZSD5	W25G1V	2,834.6
W25G1V07A0321	1265	W25G1V	3,087.6
W91QUZ07D0010	ZSA4	W25G1V	1,478.3
W91QUZ07D0006	ZS3X	W25G1V	1,431.2
W91QUZ07D0010	ZSH3	W25G1V	2,049.8
W91QUZ07D0010	ZSJ2	W25G1V	2,423.3
W25G1V07A0321	1207	W25G1V	87.3
W25G1V07A0321	2608	W25G1V	96.8
W25G1V07A0321	2462	W25G1V	96,.8
W25G1V07A0321	2597	W25G1V	96.8
W25G1V07A0321	1249	W25G1V	77.3
W91QUZ07D0010	ZSK6	W25G1V	5,732.9
W91QUZ07D0010	ZSD5	W25G1V	2,834.6
W25G1V07A0321	1265	W25G1V	308.7
W91QUZ07D0010	ZSA4	W25G1V	1,478.3
W91QUZ07D0006	ZS3X	W25G1V	1,431.2

Two of our outputs provide irregularities found in both a trend output and the anomaly output. Analysts may desire the ability to examine other trends regarding the same actor. Our application allows this congruent examination to occur, and provides tools that consolidate anomalies for desired actors into a table.

2. Compelling Trends

Changes in our response trends provide insight into awarding office behavior. These changes become increasingly important depending on the intensity of the change over time. The difference between the point of origin of estimated probabilities and their terminal point provide insight. Figure 15 shows a substantial incline of estimated probabilities regarding *Non-Competed* awards for awarding office W91ZLK (black line) beginning in FY 2015 compared to the average of all awarding offices (red line).

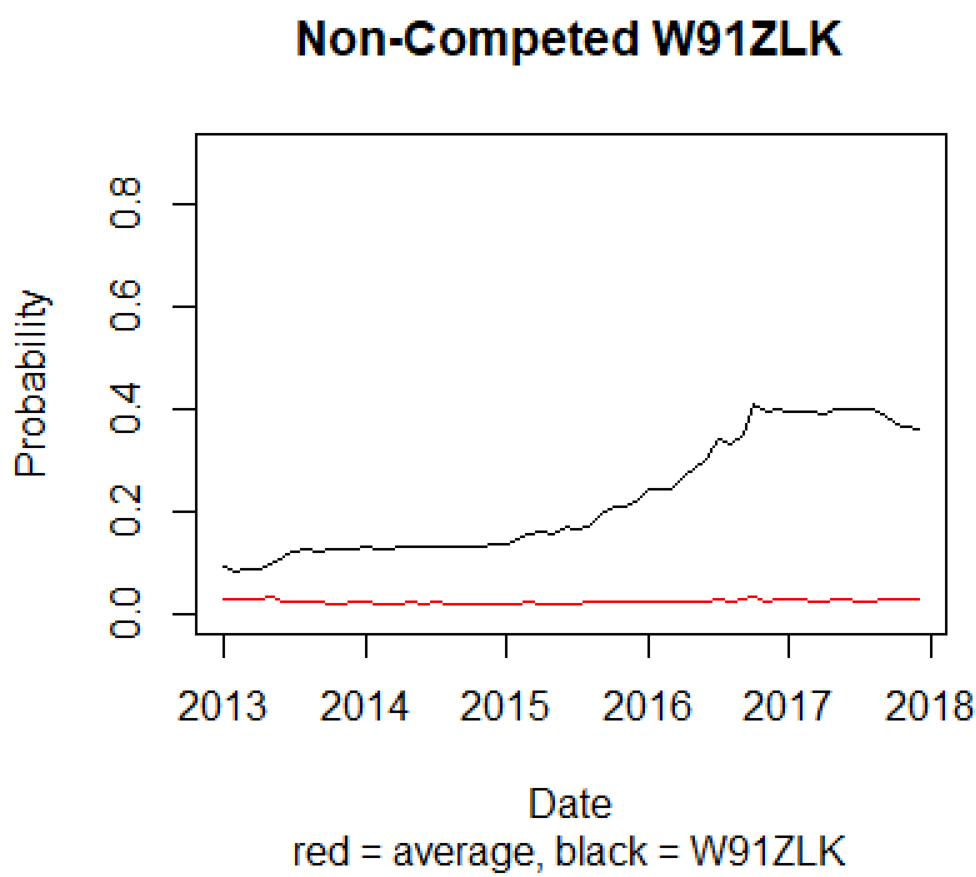

Figure 15. Categorical Random Forest Plot of Predicted Residual Probabilities of Awards Categorized as Fully and Openly Competed (Black) for Awarding Office W91ZLK, versus Averages (Red) of All Awarding Offices

As mentioned in chapter II.E, the federal government has instituted a multitude of studies surrounding best practices for contractual award processes. In December of 2014, the Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics published guidelines for creating and maintaining a competitive environment (Office of the Secretary of Defense for Acquisition, Technology, and Logistics, 2014). These guidelines emphasize the importance of competition, focusing on how a competitive environment leads to secure lower prices, raise innovation, invoke higher quality, and raise performance standards. The trend shown in Figure 14 may prompt an investigator to inquire what changes caused a significant increase in non-competed awards beginning in FY 2015.

Figure 16 also shows an increasing trend of awards categorized as *Sole-Source* for the same awarding office as Figure 15.

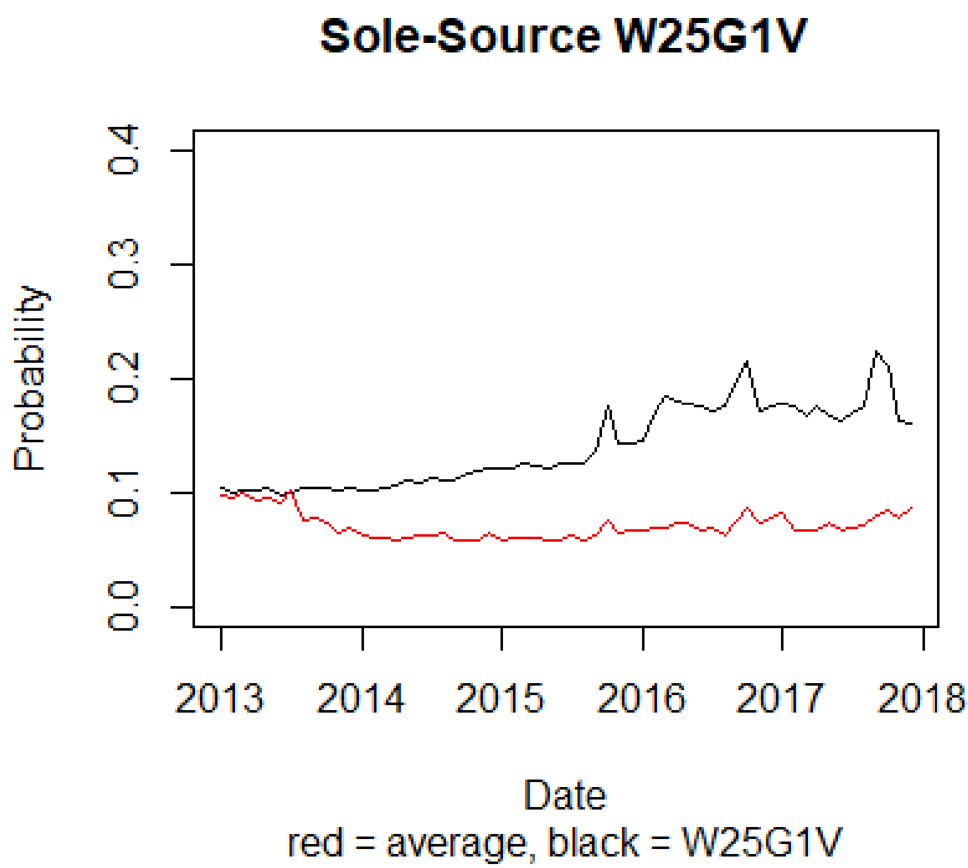

Figure 16. Categorical Random Forest Plot of Predicted Residual Probabilities of Awards Categorized as Sole-Source (Black) for Awarding Office W25G1V, versus Averages (Red) of All Awarding Offices

The trends in Figures 15 and 16 begin to ascend in FY 2015 suggestive of a possible correlation between a less competitive environment and the awarding of sole-source contracts for awarding office W25G1V. This correlation provides investigators with more knowledge surrounding possible reasons for concerning award trends and equips them with more depth to their analysis and questions they may have.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS, RECOMMENDATIONS, AND FUTURE WORK

Our study of the USASpending data demonstrates the combination of identification of important variables related to best spending practices to identify trends and contracts that stand out as unusual. We show anomalies in multiple ways to include (1) visual comparison of averages; (2) outlier identification regarding spending patterns, (3) secondary testing for values regarding our metrics, (4) processes to combine each standalone technique together to gain a more in-depth analysis of an awarding office, and (5) an application to view multiple results at once.

Utilizing our approach provides contractual auditors with information like that of Table 8. Combining Table 8, along with results from other models gives auditors multiple avenues of approach to question deviations from baseline conditions and affords them the chance to acquire details regarding observations of interest not contained within the USASpending data. This effort may improve money management, and best spending practices.

A. FUTURE WORK

With the suite of tools we provide, analysts have many options for exploring the USASpending data. We approach our research questions focusing on the awarding office. However, minor modifications to our models allow for the examination of any actor in the contractual pathway. There are over three-hundred and sixty permutations for our nine models, not including variations for separate actors. Therefore, it is impossible for us to represent all model combinations.

1. Application

Our suite of tools offer a large number of output results. In the trend analysis, there are six response variable plots for each of the 30 awarding offices resulting in 180 plots. Combined with the anomaly detection plot and another 180 p-value comparison charts equates to 361 outputs. Interactive displays offering the ability to investigate trends and spending patterns for awarding offices, organized in a tabular fashion is a natural fit for our tools. We organize our application in this manner, separating trend analysis, spending analysis, and p-value analysis results. We also expand the scope of our research by including the analysis of funding offices side by side with the awarding offices thus doubling the output stated above to 722 outputs.

2. Relationships

Our p-value and spending pattern analysis outputs allow for printing of any variable from the USASpending data. We chose to focus on the awarding office, but the recipients, funding agencies and sub-agencies are viewable within our results as well. We recommend exploring the relationships between occurrences of anomalous behaviors for the main actors, possibly eluding to interesting trends amongst actors.

3. Data

We recommend adding a variable to the USASpending data regarding adjudication of fraudulent activity occurring in the past. The Office of Justice Programs (n.d.) collects records of defense-related fraudulent activity. Merging this data with the USASpending data would allow for the enhancement of UML techniques to identify probabilities of individual contract fraud.

We look forward to the possible additions of detailed spending records describing the type of product or service procured, as well as amounts, to the USASpending data. Itemized lists of goods or services would allow for precise methods in measuring trends respective to pricing the awarding office agrees to, and average prices for the items located in the same geographic area in the same season. Subcontractor data would also prove useful in the analysis of relationship trends regarding actors—many of the fraudulent cases cited in chapter II.D involved subcontractor schemes.

APPENDIX A. SIX CATEGORIES OF CONTRACT FRAUD

Category	Definition
Defective Product/Product Substitution	Refers to instances in which the contractors knowingly deliver products to the Government that does not meet contract requirements, yet, the Government is not informed of the defect by the contractor.
Defective Testing	Defective testing refers to instances in which the contractor fails to perform contractually required tests or when they do not test the product in a manner agreed upon within the terms of the contracts.
Bid Rigging	Bid rigging can be accomplished on both the government side and the private side or with some illegal cooperation between the two. It involves an agreement to limit competitive sources to the Government. When bid rigging occurs, the Government usually pays a much higher price for the product or service because of a lack of true competition. When found to exist, the damages awarded for bi rigging are the difference between what the Government actually paid on the fraudulent claim and what it would have paid if there had been fair and open competitive bidding.
Bribery and Public Corruption	This involves the breach of Government employee's duty of loyalty to the taxpayer for personal gain.
Defective Pricing	Defective pricing refers to when a contractor has verified that their anticipated costs are current, accurate and complete as per the specific terms under the Truth in Negotiations Act. When their costs are not within their certified specifications, they are guilty of defective pricing.
False Invoices	False invoices occur when contractors knowingly submit invoices for products or services not delivered, or when the full invoice is knowingly submitted after the terms of the contract have not been fulfilled.

Adapted from Gayton (2004).

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. NAICS CODES AND DESCRIPTIONS

NAICS	Description
541330	Engineering services
517110	Wired telecommunications carriers
541519	Other computer related services
541990	All other professional, scientific, and technical services
541611	Administrative management and general management consulting services
541511	Custom computer programming services
541512	Computer systems design services
541690	Other scientific and technical consulting services
541614	Process, physical distribution, and logistics consulting services
561990	All other support services
611420	Computer training
811212	Computer and office machine repair and maintenance
541513	Computer facilities management services
518210	Data processing, hosting, and related services
811219	Other electronic and precision equipment repair and maintenance
561110	Office administrative services
811412	Appliance repair and maintenance
517210	Wireless telecommunications carriers (except satellite)
519190	All other information services
541219	Other accounting services
517410	Satellite telecommunications
541618	Other management consulting services

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. GENERAL FRAUD INDICATORS

General Indicators
Management override of key controls
Inadequate or weak internal controls
No written policies and procedures
Overly complex organizational structure
Key employee never taking leave or vacation
High turnover rate, reassignment, the firing of key personnel
Missing electronic or hard copy documents that materialize later in the review
Lost or destroyed electronic or hard copy records
Photocopied documents instead of originals. Copies are poor quality or illegible
“Unofficial” electronic files or records instead of “archived” or “official” files or records
Revisions to electronic or hard copy documents with no explanation or support
Use of means of alteration to data files
Computer-generated dates for modifications to electronic files that do not fit the appropriate timeline for surrounding their creation
Missing signatures of approval or discrepancies in signature/handwriting
Computer report totals not supported by source documentation
Lengthy unexplained delays in producing requested documentation

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. MANAGEMENT FRAUD INDICATORS

Indicator Description
Failure to display and communicate an appropriate attitude regarding the importance of internal control, including a lack of internal control policies and procedures; ethics program; codes of conduct; self-governance activities; and oversight of significant controls.
Displaying through words or actions that senior management is subject to less stringent rules, regulations, or internal controls than other employees.
Significant portion of compensation being incentive-driven based on accomplishment of aggressive target goals linked to budgetary or program accomplishments or stock prices.
High turnover of senior executives or managers.
Hostile relationship between management and internal and/or external auditors. This would include domineering behavior towards the auditor, failure to provide information, and limiting access to employees of the organization.
Failure to establish procedures to ensure compliance with laws and regulations and prevention of illegal acts.
Indications that key personnel are not competent in the performance of their assigned responsibilities.
Adverse publicity concerning an organization's activities or those of senior executives.
Lack of, or failure to adhere to, policies and procedures requiring thorough background checks before hiring key management, accounting, or operating personnel.
Inadequate resources to assist personnel in performing their duties, including personal computers, access to information, and temporary personnel.
Failure to effectively follow-up on recommendations resulting from external reviews or questions about financial results.
Nondisclosure to the appropriate Government officials of known noncompliance with laws, regulations, or significant contract or grant provisions.
Directing subordinates to perform tasks that override management or internal controls.
Undue interest or micromanagement of issues or projects that most knowledgeable individuals would identify with a substantially lower level manager.
A manager that claims disinterest or having no knowledge about a sensitive or high profile issue in which you would expect management involvement.
Constant over usage or inappropriate use of cautionary markings on management or organizational documents such as "Attorney Client Privilege/Attorney Work Product," "For Official Use Only," or other markings indicating an item is business sensitive or has a higher security classification than is appropriate.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E. DATA COLUMN NAMES

	column name
1	award_id_piid
2	modification_number
3	transaction_number
4	parent_award_agency_id
5	parent_award_agency_name
6	parent_award_id
7	parent_award_modification_number
8	federal_action_obligation
9	base_and_exercised_options_value
10	current_total_value_of_award
11	base_and_all_options_value
12	potential_total_value_of_award
13	action_date
14	period_of_performance_start_date
15	period_of_performance_current_end_date
16	period_of_performance_potential_end_date
17	ordering_period_end_date
18	awarding_agency_code
19	awarding_agency_name
20	awarding_sub_agency_code
21	awarding_sub_agency_name
22	awarding_office_code
23	awarding_office_name
24	funding_agency_code
25	funding_agency_name
26	funding_sub_agency_code
27	funding_sub_agency_name
28	funding_office_code
29	funding_office_name
30	foreign_funding
31	foreign_funding_description
32	sam_exception
33	sam_exception_description
34	recipient_duns
35	recipient_name
36	recipient_doing_business_as_name
37	cage_code
38	recipient_parent_name
39	recipient_parent_duns

40	recipient_country_code
41	recipient_country_name
42	recipient_address_line_1
43	recipient_address_line_2
44	recipient_city_name
45	recipient_state_code
46	recipient_state_name
47	recipient_zip_4_code
48	recipient_congressional_district
49	recipient_phone_number
50	recipient_fax_number
51	primary_place_of_performance_country_code
52	primary_place_of_performance_country_name
53	primary_place_of_performance_city_name
54	primary_place_of_performance_county_name
55	primary_place_of_performance_state_code
56	primary_place_of_performance_state_name
57	primary_place_of_performance_zip_4
58	primary_place_of_performance_congressional_district
59	award_or_idv_flag
60	award_type_code
61	award_type
62	idv_type_code
63	idv_type
64	multiple_or_single_award_idv_code
65	multiple_or_single_award_idv
66	type_of_idc_code
67	type_of_idc
68	type_of_contract_pricing_code
69	type_of_contract_pricing
70	award_description
71	action_type_code
72	action_type
73	solicitation_identifier
74	number_of_actions
75	product_or_service_code
76	product_or_service_code_description
77	contract_bundling_code
78	contract_bundling
79	dod_claimant_program_code
80	dod_claimant_program_description
81	naics_code
82	naics_description

83	recovered_materials_sustainability_code
84	recovered_materials_sustainability
85	domestic_or_foreign_entity_code
86	domestic_or_foreign_entity
87	dod_acquisition_program_code
88	dod_acquisition_program_description
89	information_technology_commercial_item_category_code
90	information_technology_commercial_item_category
91	epa_designated_product_code
92	epa_designated_product
93	country_of_product_or_service_origin_code
94	country_of_product_or_service_origin
95	place_of_manufacture_code
96	place_of_manufacture
97	subcontracting_plan_code
98	subcontracting_plan
99	extent_competed_code
100	extent_competed
101	solicitation_procedures_code
102	solicitation_procedures
103	type_of_set_aside_code
104	type_of_set_aside
105	evaluated_preference_code
106	evaluated_preference
107	research_code
108	research
109	fair_opportunity_limited_sources_code
110	fair_opportunity_limited_sources
111	other_than_full_and_open_competition_code
112	other_than_full_and_open_competition
113	number_of_offers_received
114	commercial_item_acquisition_procedures_code
115	commercial_item_acquisition_procedures
116	small_business_competitiveness_demonstration_program
117	commercial_item_test_program_code
118	commercial_item_test_program
119	a76_fair_act_action_code
120	a76_fair_act_action
121	fed_biz_opps_code
122	fed_biz_opps
123	local_area_set_aside_code
124	local_area_set_aside
125	price_evaluation_adjustment_preference_percent_difference

126 clinger_cohen_act_planning_code
127 clinger_cohen_act_planning
128 materials_supplies_articles_equipment_code
129 materials_supplies_articles_equipment
130 labor_standards_code
131 labor_standards
132 construction_wage_rate_requirements_code
133 construction_wage_rate_requirements
134 interagency_contracting_authority_code
135 interagency_contracting_authority
136 other_statutory_authority
137 program_acronym
138 parent_award_type_code
139 parent_award_type
140 parent_award_single_or_multiple_code
141 parent_award_single_or_multiple
142 major_program
143 national_interest_action_code
144 national_interest_action
145 cost_or_pricing_data_code
146 cost_or_pricing_data
147 cost_accounting_standards_clause_code
148 cost_accounting_standards_clause
149 gfe_gfp_code
150 gfe_gfp
151 sea_transportation_code
152 sea_transportation
153 undefinitized_action_code
154 undefinitized_action
155 consolidated_contract_code
156 consolidated_contract
157 performance_based_service_acquisition_code
158 performance_based_service_acquisition
159 multi_year_contract_code
160 multi_year_contract
161 contract_financing_code
162 contract_financing
163 purchase_card_as_payment_method_code
164 purchase_card_as_payment_method
165 contingency_humanitarian_or_peacekeeping_operation_code
166 contingency_humanitarian_or_peacekeeping_operation
167 alaskan_native_owned_corporation_or_firm
168 american_indian_owned_business

169 indian_tribe_federally_recognized
170 native_hawaiian_owned_business
171 tribally_owned_business
172 veteran_owned_business
173 service_disabled_veteran_owned_business
174 woman_owned_business
175 women_owned_small_business
176 economically_disadvantaged_women_owned_small_business
177 joint_venture_women_owned_small_business
178 joint_venture_economic_disadvantaged_women_owned_small_bus
179 minority_owned_business
180 subcontinent_asian_asian_indian_american_owned_business
181 asian_pacific_american_owned_business
182 black_american_owned_business
183 hispanic_american_owned_business
184 native_american_owned_business
185 other_minority_owned_business
186 contracting_officers_determination_of_business_size
187 contracting_officers_determination_of_business_size_code
188 emerging_small_business
189 community_developed_corporation_owned_firm
190 labor_surplus_area_firm
191 us_federal_government
192 federally_funded_research_and_development_corp
193 federal_agency
194 us_state_government
195 us_local_government
196 city_local_government
197 county_local_government
198 inter_municipal_local_government
199 local_government_owned
200 municipality_local_government
201 school_district_local_government
202 township_local_government
203 us_tribal_government
204 foreign_government
205 organizational_type
206 corporate_entity_not_tax_exempt
207 corporate_entity_tax_exempt
208 partnership_or_limited_liability_partnership
209 sole_proprietorship
210 small_agricultural_cooperative
211 international_organization

212 us_government_entity
213 community_development_corporation
214 domestic_shelter
215 educational_institution
216 foundation
217 hospital_flag
218 manufacturer_of_goods
219 veterinary_hospital
220 hispanic_servicing_institution
221 receives_contracts
222 receives_grants
223 receives_contracts_and_grants
224 airport_authority
225 council_of_governments
226 housing_authorities_public_tribal
227 interstate_entity
228 planning_commission
229 port_authority
230 transit_authority
231 subchapter_scorporation
232 limited_liability_corporation
233 foreign_owned_and_located
234 for_profit_organization
235 nonprofit_organization
236 other_not_for_profit_organization
237 the_ability_one_program
238 number_of_employees
239 annual_revenue
240 private_university_or_college
241 state_controlled_institution_of_higher_learning
242 X1862_land_grant_college
243 X1890_land_grant_college
244 X1994_land_grant_college
245 minority_institution
246 historically_black_college
247 tribal_college
248 alaskan_native_servicing_institution
249 native_hawaiian_servicing_institution
250 school_of_forestry
251 veterinary_college
252 dot_certified_disadvantage
253 self_certified_small_disadvantaged_business
254 small_disadvantaged_business

255 c8a_program_participant
256 historically_underutilized_business_zone_hubzone_firm
257 sba_certified_8a_joint_venture
258 last_modified_date

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Amadeo, K. (2018). U.S. federal budget breakdown: The budget components and impact on the U.S. economy. *The Balance*, 1. Retrieved from https://www.thebalance.com/u-s-federal-budget-breakdown-3305789

Anders. (2015). The fraud triangle: Three conditions that increase the risk of fraud. Retrieved from http://anderscpa.com/the-fraud-triangle-three-conditions-that-increase-the-risk-of-fraud/

Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud analytics: Using descriptive, predictive, and social network techniques* (1st ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Buttrey, S., & Whitaker, L. (2018). *A data scientist's guide to acquiring, cleaning, and managing data in R* (1st ed.). Hoboken NJ: John Wiley & Sons, Inc. doi:9781119080022

Celebi, E. M., & Aydin, K. (2016). Unsupervised learning algorithms. New York, NY: Springer. doi:10.1007/978-3-319-24211-8

Cheng, L. (2013). *Unsupervised topic discovery by anomaly detection* (Master's thesis). Retrieved from https://calhoun.nps.edu/handle/10945/37599

Code of ethics for government services of 1980, Pub. L No. 96–303, Stat 94. (1980). Retrieved from https://www.govinfo.gov/content/pkg/STATUTE-94/pdf/STATUTE-94-Pg855.pdf

Costa, D. (2017). Insider threat blog: Real-world work combating insider threats. Retrieved from https://insights.sei.cmu.edu/insider-threat/2017/03/cert-definition-of-insider-threat---updated.html

Defense Pricing and Contracting. (2018). Federal Organizational Quarterly Competition Report Card. Retrieved from https://www.acq.osd.mil/dpap/cpic/cp/docs/Publication_of_DoD_Competition_Reporting_-_1st_Quarter_FY18.pdf

Devore J. (2016). *Probability and statistics: for engineering and the sciences* (9th ed.). Boston MA: Cengage Learning. doi:9781305251809

e-CFR Electronic code of federal regulations. (2019). Retrieved from https://www.ecfr.gov/cgi-bin/text-idx?SID=0ff5f0839abff4eec707b4478ed733c6&mc=true&node=pt13.1.121&rgn=div5

Faraway, J. (2015). *Linear models with R* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC

Faraway, J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (2nd ed.). Boca Raton FL: CRC Press .

Friedman, J. (2018). Here are 10 genius quotes from warren buffet. Retrieved from https://www.forbes.com/sites/zackfriedman/2018/10/04/warren-buffett-best-quotes/#4553e0174261

García-Teodoroa, P., Díaz-Verdejoa, J., Maciá-Fernándeza, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, *28*(1-2), 18–28.

Gayton, J.R.S. (2004). *Procurement fraud case studies* (Master's report). Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a424697.pdf

GSA federal procurement data system-next generation (FPDS-NG) data element dictionary. (2018). Retrieved from https://www.fpds.gov/downloads/Version_1.5_specs/FPDSNG_DataDictionary_V1.5.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (1st ed.). New York, NY: Springer. doi:10.1007/978-1-4614-7138-7

Kramer, M. (2012). The most common procurement fraud schemes and their primary red flags. Retrieved from https://iacrc.org/procurement-fraud/the-most-common-procurement-fraud-schemes-and-their-primary-red-flags/

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News 2*(3), 18--22. Retrieved from https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Mcallister, T. (2017). 2 San Diego contractors sentenced in department of defense scam. Retrieved from https://timesofsandiego.com/crime/2017/12/14/2-san-diego-contractors-sentenced-in-department-of-defense-scam/

Mihm, C., & Holman, B. (1998). *Competitive contracting: Information related to the redrafts of the freedom from government competition act* (No. B-279861). Washington, DC: United States General Accounting Office. Retrieved from https://www.gao.gov/assets/90/87613.pdf

Monteiro, A. S. (2002). *Multiple additive regression trees—A methodology for predictive data mining for fraud detection* (Master's thesis). Retrieved from http://hdl.handle.net/10945/4812

Murray, K. (2016). SBA's 8(a) certification program explained. Retrieved from https://www.sba.gov/blogs/sbas-8a-certification-program-explained

Office of Justice Programs, Bureau of Justice Assistance, U.S. Department of Justice, Defense Procurement Fraud Debarment (DPFD) Clearinghouse. (n.d.). Retrieved March 22, 2018, from https://www.bja.gov/ProgramDetails.aspx?Program_ID=56

Publications Branch. (1993). *Handbook on fraud indicators for contract auditors*. Arlington, VA: Inspector General of Department of Defense. Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a340512.pdf

R Core Team. (2014). R: A language and environment for statistical computing [Computer program]. Retrieved from http://www.R-project.org/.

Rouillard, G. W. (2003). *An improved unsupervised modeling methodology for detecting fraud in vendor payment transactions* (Master's thesis). Retrieved from https://calhoun.nps.edu/bitstream/handle/10945/916/03Jun_Rouillard.pdf?sequence=1&isAllowed=y

Sellers, D. P. (1996). *Predicting fraudulent behavior: An examination of characteristics commonly exhibited by fraudulent DoD contractors* (Master's thesis). Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a319528.pdf

Shane, L. (2011). Report: U.S. wasted $60 billion in contracting fraud, abuse. *Stars and Stripes*, 1. Retrieved from https://www.stripes.com/news/report-u-s-wasted-60-billion-in-contracting-fraud-abuse-1.153787

Tan, L. (2013). *An analysis of internal controls and procurement fraud deterrence* (Master's thesis). Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a620549.pdf

Teefy, J. (2018). *Tracking federal awards: USAspending.gov and other data sources* (R44027 Version 14). Washington, DC: Congressional Research Service. Retrieved from https://fas.org/sgp/crs/misc/R44027.pdf

United States Attorney's Office District of Kansas. (2017, Dec. 1). Indictment: Veteran charged in $40 Million Government Contract Fraud Scheme. Retrieved from https://www.justice.gov/usao-ks/pr/indictment-veteran-charged-40-million-government-contract-fraud-scheme

United States Attorney's Office District of Columbia. (2013, Jul. 11). Former U.S Army Corps of Engineers manager sentenced to more than 19 years in prison in $30 million bribery and kickback scheme. Retrieved from https://archives.fbi.gov/archives/washingtondc/press-releases/2013/former-u.s.-army-corps-of-engineers-manager-sentenced-to-more-than-19-years-in-prison-in-30-million-bribery-and-kickback-scheme

United States Attorney's Office District of Maryland. (2017, Jun. 27). Defense contractor sentenced to 30 months in federal prison for $53 million procurement fraud and illegal gratuities scheme. Retrieved from https://www.justice.gov/usao-md/pr/defense-contractor-sentenced-30-months-federal-prison-53-million-procurement-fraud-and

USASpending. (2019). Award data archive. Retrieved from https://www.usaspending.gov/#/download_center/award_data_archive

USASpending Data Lab. (2019). Retrieved from https://datalab.usaspending.gov/contract-explorer.html

Wandland, K. W., & Wickman, G. P. (1993). *Analysis of cost and schedule growth on sole source and competitive air force contracts* (Master's thesis). Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a273920.pdf

World Bank. (n.d.). GNI per capita. Retrieved March 22, 2019, from https://data.worldbank.org/indicator/ny.gnp.pcap.pp.cd

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California