3-24-2016

# Mission Dependency Index of Air Force Built Infrastructure: Knowledge Discovery with Machine Learning

Clark W. Smith


# MISSION DEPENDENCY INDEX OF AIR FORCE BUILT INFRASTRUCTURE: KNOWLEDGE DISCOVERY WITH MACHINE LEARNING

THESIS

Clark W. Smith, Captain, USAF

AFIT-ENV-MS-16-M-184

**DEPARTMENT OF THE AIR FORCE**
**AIR UNIVERSITY**

**AIR FORCE INSTITUTE OF TECHNOLOGY**

**Wright-Patterson Air Force Base, Ohio**

AFIT-ENV-MS-16-M-184

# MISSION DEPENDENCY INDEX OF AIR FORCE BUILT INFRASTRUCTURE: KNOWLEDGE DISCOVERY WITH MACHINE LEARNING

THESIS

Presented to the Faculty

Department of Systems Engineering and Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Engineering Management

Clark W. Smith, BS

Captain, USAF

March 2016

AFIT-ENV-MS-16-M-184

# MISSION DEPENDENCY INDEX OF AIR FORCE BUILT INFRASTRUCTURE: KNOWLEDGE DISCOVERY WITH MACHINE LEARNING

Clark W. Smith, BS

Captain, USAF

Committee Membership:

Maj Vhance V. Valencia, PhD
Chair

Brett J. Borghetti, PhD
Member

Lt Col Brent T. Langhals, PhD
Member

## Abstract

Mission Dependency Index (MDI) is a metric developed to capture the relative criticality of infrastructure assets with respect to organizational missions. The USAF adapted the MDI metric from the United States Navy's MDI methodology. Unlike the Navy's MDI data collection process, the USAF adaptation of the MDI metric employs generic facility category codes (CATCODEs) to assign MDI values. This practice introduces uncertainty into the MDI assignment process with respect to specific missions and specific infrastructure assets. The uncertainty associated with USAF MDI values necessitated the MDI adjudication process. The MDI adjudication process provides a mechanism for installation civil engineer personnel to lobby for accurate MDI values for specific infrastructure assets. The MDI adjudication process requires manual review of facilities and MDI discrepancies, justification documentation, and extensive coordination between organizations.

In light of the existing uncertainty with USAF MDI values and the level of effort required for the MDI adjudication process, this research pursues machine learning and the knowledge discovery in databases (KDD) process to identify and understand relationships between real property data and mission critical infrastructure. Furthermore, a decision support tool is developed for the MDI adjudication process. Specifically, supervised learning techniques are employed to develop a classifier that can identify potential MDI discrepancies. This automation effort serves to minimize the manual MDI review process by identifying a subset of facilities for review and potential adjudication.

**Acknowledgments**

I would like to thank my faculty advisor, Maj Vhance Valencia, for encouraging me to step outside of my comfort zone and learn something entirely new. I would also like to thank my committee members, Dr. Brett Borghetti and Lt Col Brent Langhals, for their guidance and feedback throughout the research process. Finally, I would like to thank my amazing wife for her perseverance and support throughout this endeavor. Your partnership means more than words can convey and I look forward to the many adventures ahead.

Clark W. Smith

## Table of Contents

Page
Abstract ..... iv
Table of Contents ..... vi
List of Figures ..... viii
List of Tables ..... ix
I. Introduction ..... 1
Mission Dependency Index Background ..... 4
Problem Statement ..... 5
Machine Learning ..... 6
Research Objective and Investigative Questions ..... 9
Methodology ..... 9
Assumptions/Limitations ..... 10
Overview ..... 12
II. Literature Review ..... 13
Chapter Overview ..... 13
USAF Real Property Portfolio and Requirements ..... 13
Asset Management Background ..... 14
Asset Management Challenges ..... 16
Asset Management within the Federal Government ..... 18
MDI Background ..... 20
NAVFAC MDI Model ..... 22
USAF MDI Implementation ..... 28
USAF MDI Adjudication Process ..... 31
Navy MDI Limitations ..... 35
USAF MDI Limitations ..... 38
Data Facilitates Effective Asset Management ..... 40
Real Property Databases ..... 42
Knowledge Discovery in Databases (KDD) ..... 44
Data Mining Background ..... 46
Data Mining Literature Review ..... 46
Chapter Summary ..... 48
III. Methodology ..... 49
Chapter Overview ..... 49
Knowledge Discovery in Databases (KDD) ..... 49
Step 1: Learn the Application Domain and Establish Goals ..... 50
Step 2: Creating a Target Data Set ..... 52
Step 3: Data Cleaning and Preprocessing ..... 56
Step 4: Data Reduction and Projection ..... 61
Step 5: Choosing the Data Mining Task ..... 64
Chapter Summary ..... 69

IV. Analysis and Results....70

Chapter Overview....70
Steps 6 – 8: Algorithm Selection, Data Mining, Interpretation and Evaluation....71
Step 9: Using Discovered Knowledge....96
Chapter Summary....99

V. Conclusions and Recommendations....100

Chapter Overview....100
Investigative Questions Answered....100
Conclusions of Research....106
Significance of Research....107
Recommendations for Action....107
Recommendations for Future Research....108
Summary....109

Appendix A. Data Mining Algorithms....110

Bibliography....116

## List of Figures

Page
Figure 1. Asset Management Principles (Teicholz et al., 2005) ..... 19
Figure 2. FRPC Data Elements, Performance Measures (Teicholz et al., 2005) ..... 19
Figure 3. NAVFAC Mission Intradependency Matrix (Dempsey, 2006) ..... 25
Figure 4. NAVFAC Mission Interdependency Score Matrix (Dempsey, 2006) ..... 25
Figure 5. Hypothetical Operations Group Intradependencies (Antelman, 2008) ..... 27
Figure 6. Hypothetical Operations Group Interdependencies (Antelman, 2008) ..... 27
Figure 7. MDI Score Distributions at Fairchild AFB (Antelman, 2008) ..... 29
Figure 8. MDI Score Distributions at Langley AFB (Antelman, 2008) ..... 29
Figure 9. MDI Adjudication Status (Current as of Aug 2015) ..... 32
Figure 10. MAJCOM MDI Refinement Histogram (Current as of Aug 2015) ..... 33
Figure 11. MDI Refinement Process: Identify Discrepancies (AFCEC, 2015) ..... 34
Figure 12. MDI Refinement Process: Update Real Property Records (AFCEC, 2015) . 34
Figure 13. USAF Data Set Feature Selection Results ..... 63
Figure 14. Notional ROC Curve ..... 68
Figure 15. Fairchild Classifier Comparison: Subset 1 ROC AUC Values ..... 73
Figure 16. Fairchild Classifier Comparison: Subset 2 ROC AUC Values ..... 73
Figure 17. Fairchild Classifier Comparison: Subset 3 ROC AUC Values ..... 74
Figure 18. Fairchild Classifier Comparison: Subset 4 ROC AUC Values ..... 74
Figure 19. Fairchild Classifier Comparison: Subset 5 ROC AUC Values ..... 75
Figure 20. Fairchild Classifier Comparison: All Features ROC AUC Values ..... 75
Figure 21. Navy Classifier Comparison: ROC AUC Values ..... 83
Figure 22. Navy Random Forests Classifier Tuning Parameter ..... 90
Figure 23. Navy Decision Tree Comparison ..... 91
Figure 24. Training Set Results for Navy C5.0 Classifier with Cost Matrix ..... 93
Figure 25. Test Set Results for Navy C5.0 Classifier with Cost Matrix ..... 93
Figure 26. Variable Importance: C5.0 Classifier with Cost Matrix ..... 95
Figure 27. Facility Class Frequencies for Mission Critical Predictions ..... 98
Figure 28. Category Group Frequencies for Mission Critical Predictions ..... 98

## List of Tables

Page
Table 1. Navy MDI Survey Questions (Antelman, 2008) ..... 23
Table 2. Response Options for Interruptibility (Antelman, 2008)..... 24
Table 3. Response Options for Relocateability and Replaceability (Antelman, 2008) ... 24
Table 4. MARM Categories and Examples (Madaus, 2009)..... 31
Table 5. Facility Attributes for Resource Allocation (Albrice et al., 2014) ..... 41
Table 6. Data Mining Keyword Trends, 2000-2011 (Liao et al., 2012). ..... 47
Table 7. Original MDI Beta Test Data Features..... 53
Table 8. "Fairchild RT_FACILITIES" Original Data Features..... 54
Table 9. "Fairchild RT_REAL_PROPERTY_ASSETS" Original Data Features..... 55
Table 10. Original Navy Data Features ..... 56
Table 11. USAF Data Set Features..... 59
Table 12. Navy Data Set Features..... 60
Table 13. Notional Confusion Matrix..... 66
Table 14. Lasso Model Results for USAF Data Set ..... 80
Table 15. Lasso Model Results for Navy Data Set..... 86
Table 16. Lasso Model Results for Navy Data Set Category Group Feature ..... 88
Table 17. Cost Matrix for C5.0 Algorithm ..... 92
Table 18. AFCENT Real Property Data Features..... 96
Table 19. Classifier Results for AFCENT Installations..... 97

# MISSION DEPENDENCY INDEX OF AIR FORCE BUILT INFRASTRUCTURE: KNOWLEDGE DISCOVERY WITH MACHINE LEARNING

## I. Introduction

In 2004, the United States Air Force (USAF), along with the other federal agencies, received direction via Executive Order (EO) 13327 to implement asset management principles in overseeing real property assets. Asset management can be defined as "a systematic process of maintaining, upgrading, and operating physical assets cost-effectively" (McElroy, 1999). While there are many definitions of asset management, common themes include deliberate processes, data collection, and data analysis employed in managing infrastructure life-cycle-costs. Asset management is especially important within the federal government as taxpayers expect transparency, accountability, and cost effective operations (McElroy, 1999). An analysis conducted approximately one year after the signing of EO 13327 summarized "the EO's primary objective is to promote efficient and economical use of the federal government's real property assets" (Teicholz, Nofrei, & Thomas, 2005). To this end, Major General Del Eulberg, the Air Force Civil Engineer from June 2006 to August 2009, implemented asset management for Air Force civil engineering assets and is arguably the first champion of the USAF transition to asset management. In 2008, Eulberg wrote the following excerpt regarding asset management principles in an issue of the *Air Force Civil Engineer Magazine*: "We can no longer afford to allocate resources according to some fair-share, 'peanut butter spread' method – asset management is all about a proactive, fact-based

approach to analyze data to make the best decisions possible" (Eulberg, 2008). This statement emphasized the deliberate pursuit of data-driven analysis and prioritization of resource allocation. At that time, asset management principles had already been implemented within the Federal Highway Administration (FHWA) as well as other government and business organizations across the globe (Hodkiewicz, 2015; McElroy, 1999). The general's charge was to follow suit in implementing asset management principles in the USAF.

The USAF real property portfolio is vast. The Department of Defense (DOD) Base Structure Report (2013) indicates that the USAF real property portfolio encompasses assets across the globe and has an estimated Plant Replacement Value (PRV) of over $259 billion. The nature of USAF mission sets and support functions necessitate an extensive real property portfolio. The USAF is charged with providing the United States with specific capabilities to enable global vigilance, global reach, and global power. These capabilities are fulfilled through the following proficiencies as indicated in the Air Force 2023 Implementation Plan (USAF, 2013):

**Global Vigilance**

- Space Superiority (Global Space Mission Operations)
- Strategic Warning
- Space Situational Awareness
- Global ISR (includes all domains)
- Defensive Cyberspace Operations
- Theater Missile Warning
- Theater ISR (Airborne and Cyberspace)

**Global Reach**

- Air Refueling (to enable global operations)
- Inter-theater Airlift
- Theater Air Refueling
- Intra-theater Airlift
- Aeromedical Evacuation

**Global Power**

- Nuclear Deterrence Operations
- Global Command and Control (C2)
- Global Precision Attack (includes Offensive Cyberspace Ops)
- Space Superiority (Space Control)
- Theater C2
- Theater Air Superiority
- Theater Precision Attack (Interdiction, Special Ops, Close-Air-Support, Offensive Cyberspace Ops)
- Combat Search and Rescue (Personnel Recovery)

The 20 mission capabilities in the 2023 Implementation Plan represent complex and diverse requirements vital to national defense and combatant command (COCOM) mission requirements. These mission capabilities require significant infrastructure for effective operations; and this infrastructure must be maintained. However, there is a finite budget with which these real property assets can be maintained resulting in a multitude of unfulfilled facility requirements. As such, difficult decisions and tradeoffs must be made with respect to funding mission critical facility requirements first. The

prioritization of enterprise infrastructure requirements is a uniquely complex task for USAF Civil Engineer (CE) personnel.

**Mission Dependency Index Background**

The USAF prioritization model for infrastructure assets has changed multiple times in the past few years. The two most recent models have employed a Mission Dependency Index (MDI) value in resource allocation (Nichols, 2015). The Federal Real Property Council (FRPC) defines MDI as "the value an asset brings to the performance of the mission as determined by the governing agency" (FRPC, 2011). Currently, Air Force MDIs are assigned based on real property category codes (CATCODES). CATCODES are implemented via the Federal Real Property Categorization System (RPCS), which provides a detailed hierarchy of real property uses as directed by the DOD (DOD, 2015).

The MDI process originated from work done by the U.S. Navy and U.S. Coast Guard in order to facilitate real property funding prioritization. The Navy's process is relatively robust, as it includes installation specific interviews with real property stakeholders. These interviews determine the individual input values for an MDI equation. MDI data collection does not come without a cost, however. The United States Army Corps of Engineers (USACE) estimated that the data collection effort may cost anywhere between $40,000 and $75,000 per installation (Michael Grussing et al., 2010).

In 2008, the USAF partnered with Navy MDI experts to conduct a "proof of concept" at two installations, Langley Air Force Base (AFB) and Fairchild AFB (Antelman, 2008). The Navy MDI model implemented in this proof of concept was found to be generally accurate in most cases; but, due to complexity and cost, the Air

Force opted to forego installation specific interviews with stakeholders (AFCEC, 2015). In lieu of data collection, the USAF utilized the asset specific Navy MDI data to derive MDI values for general CATCODES. In order to map MDI values from existing Navy facility data, USAF Civil Engineers created statistical distributions of the MDI values for each of the four-digit Facility Analysis Codes (FACs). FAC categories are common across the entire DOD, thus providing the most equivalent means of comparing Navy and Air Force real property assets. Civil Engineers evaluated the MDI distributions and selected the most appropriate MDI point value for each four-digit FAC. Next, USAF Civil Engineers utilized the four-digit FAC and MDI mapping to further derive MDI values for the more specific six-digit USAF CATCODEs. This meant that each USAF CATCODE was assigned a distinct MDI value, which could then be applied to individual real property assets across the USAF enterprise.

**Problem Statement**

The uncertainty associated with USAF MDI values necessitates extensive review and validation real property mission criticality. USAF Civil Engineers initially developed the existing MDI assignment method as an interim solution as it relied solely on generic facility use categories instead of data collection from the installations. This situation remains with MDI values assigned solely based off of real property CATCODES. The lack of installation and mission specific data, however, leads to inconsistencies with real property MDI values. Nichols (2015) uses the example of a humidity-controlled warehouse, CATCODE 442421, which has an MDI of 59. Nichols conveys that an MDI of 59 may not capture the true mission criticality of such a

warehouse if the facility supports a special operations or cyber warfare mission. In this scenario, the installation real property officer and affected stakeholders would have to lobby for support and provide adequate justification to correct the MDI value.

In 2014, the Air Force Civil Engineer Center (AFCEC) provided guidance to Major Commands (MAJCOMs) and installations on a standardized process to adjudicate real property MDI values not representative of the true mission criticality. The policy concedes the following: "The MDI is currently assigned using a facility's designated CATCODE, which provides an accurate assessment of facility criticality in most cases, but not all" (AFCEC, 2015). The MDI adjudication process requires six-levels of coordination beginning with a base-level engineer review of facilities and MDI values to identify discrepancies. The MDI adjudication process emphasizes the need for base-level input, more sufficient data, and generally, a more effective MDI assignment process that captures the context surrounding infrastructure and missions. In short, the current USAF MDI methodology does not effectively characterize the true relationship between real property assets and mission criticality due to insufficient supporting data. Furthermore, the MDI refinement process drives additional personnel and management workload for installations, MAJCOMs, and AFCEC. Such a process may benefit from advances in computational techniques and automation widely accessible with today's computers.

**Machine Learning**

As the world rapidly advances in the information age, reliance on decision support systems (DSS) is driving research for more effective methods of acquiring knowledge from data. Air Force CE priorities are shifting toward data-centric asset

management practices in the wake of academic and private sector developments in this field. This transition is a tremendous paradigm shift from the largely reactive and costly approach to facility sustainment of the past. The Air Force Asset Management Plan (AFAMP) states, "Asset visibility should form a data foundation upon which the Air Force may accurately measure and communicate these risks to defend needed funding" (AFCEC, 2014a). This statement captures several key themes in asset management including measurement, communication, and risk, all of which are undergirded by data.

The intent of the MDI metric is to use data obtained from facility and mission stakeholders to arrive at a quantitative representation of the consequence of failure. This is a highly complex problem and the Navy has invested significant time, money, and personnel in gathering data to implement their MDI methodology. Given that a significant amount of data exists from the Navy's MDI efforts, how can this data be used to better understand the relationships between MDI and real property data? Understanding the relationships between MDI and real property data could lead to beneficial heuristics or rule-based decisions for USAF implementation. Furthermore, this knowledge could eliminate the need for costly data collection or provide a more effective method of adjudicating improperly assigned MDIs. Given that two USAF installations were evaluated using the Navy MDI methodology, can this data be used to create a model that can predict mission critical infrastructure? Machine learning is an area of study that may be able to provide these benefits.

Machine learning is the study of acquiring knowledge from data automatically through efficient computational methods (Langley & Simon, 1995). There are two primary domains within machine learning: supervised learning and unsupervised

learning. Supervised learning utilizes a data set that includes one or more data features where each observation includes the corresponding correct answer, or label. There are many learning paradigms within supervised learning. The analytical objectives and data format typically drive supervised learning paradigm selection. Alternatively, unsupervised learning is the process of looking for structure that exists in the data set without the use of a correct answer, or label. The ever-decreasing cost of data storage and computational processing power catalyzes machine learning techniques and applications. Because of this momentum, machine learning applications are employed in a myriad of fields and provide many options for solving complex problems or simply obtaining a better understanding of relationships in data.

Ultimately, machine learning techniques could be employed to better understand and facilitate real property prioritization based on organizational objectives, facility condition, life cycle cost analysis, real property data features, and mission characteristics. This machine learning application could provide a beneficial decision support tool to aid in the process of managing effective allocation of taxpayer dollars to meet DOD mission objectives.

## Research Objective and Investigative Questions

This research will demonstrate the employment of machine learning for understanding relationships between real property data as well as predicting mission critical real property assets. To facilitate this objective, five research questions were developed to guide the research:

1. How can machine learning techniques, specifically supervised learning, be applied to predict mission critical USAF facilities?
2. What features should be collected for such an algorithm?
3. What is the appropriate architecture for such an algorithm?
4. What are the costs and benefits associated with employing machine learning in Air Force asset management facility prioritization?
5. How can the Knowledge Discovery in Databases (KDD) process be applied to facilitate MDI reviews for AFCENT facilities?

## Methodology

The overarching methodology for this research is the Knowledge Discovery in Databases (KDD) process. This research explores the use of supervised learning algorithms for classifying USAF mission critical infrastructure. Supervised learning has become increasingly common in science and business applications to learn from experience, draw conclusions, and make predictions. In order to utilize supervised learning techniques, the USAF MDI proof of concept data and USN real property data is employed in developing prediction models for USAF real property assets. This is deemed supervised learning because the correct outputs are provided as examples from which the model may learn. Both the Navy real property data and the MDI proof of concept data are separated into two sets: the first is utilized as the training set and the

second serves as the test set for model evaluation. A successful model can be used as a decision support tool to facilitate the USAF MDI adjudication process.

United States Air Forces Central Command (AFCENT) is the sponsor for this research. AFCENT maintains three major air bases in Southwest Asia: Al Udeid Air Base, Qatar; Al Dhafra Air Base, UAE; and Ali Al Salem Air Base, Kuwait. AFCENT has historically utilized Overseas Contingency Operations (OCO) funding to maintain and operate installations. AFCENT is expected to transition away from OCO funding in the future and align with the funding model employed across the USAF. Given this imminent transition, it is in AFCENT's best interest to ensure that infrastructure MDI values accurately reflect mission criticality. AFCENT civil engineer staff provided real property data for each of the three operating locations. The ultimate goal of this research is to support AFCENT civil engineers in identifying mission critical infrastructure and potential opportunities for MDI adjudication to better implement resource allocation.

**Assumptions/Limitations**

The primary assumption with this research is that relationships exist between real property data and mission critical infrastructure. The current USAF methodology supports this assumption as CATCODEs are the primary mechanism for MDI assignments. CATCODEs represent generic functions associated with a given infrastructure asset, which provides some indication of mission criticality. The inconsistencies with CATCODE-assigned MDI values are indicative of the fact that CATCODEs alone are suboptimal for a mission criticality assessment. Additional real property data features employed with generic function codes may offer improved fidelity

for mission criticality assessments. If additional real property features do not contribute to mission criticality, data collection efforts should be pursued for MDI reliability.

There are five limitations with this study. First, one of the most important steps in machine learning algorithm development is determining the appropriate data inputs for the respective data sets. There are many variables that can be considered in determining the consequence of failure for a specific facility. Through employing various machine learning algorithms, it may become evident that the available facility data is not sufficient to explain the underlying mission criticality phenomenon. Second, Navy real property data will be analyzed for relationships between real property data and MDI values. Navy and USAF CATCODES differ at the six-digit level but Facility Analysis Codes (FAC) align at the four-digit level per DOD requirements. This code alignment must be considered in data selection and comparisons between Navy and Air Force data. Third, utilizing CATCODEs requires the assumption that facilities are recorded with the correct CATCODE. Fourth, a general assumption is that real property data, obtained from the respective databases of record, correctly describe the infrastructure assets. Fifth, some facilities serve multiple functions and have multiple CATCODEs, however, USAF facility MDI values are assigned based on the predominant CATCODE. This is indicative of a decrease in fidelity when assigning facility MDI values based solely on CATCODEs.

## Overview

This chapter presented a brief synopsis of the USAF MDI application within the context of the Air Force enterprise real property portfolio and the fiscally constrained environment. Chapter II of this document summarizes the literature reviewed for this research study. Chapter III addresses the KDD process and the steps leading up to the MDI machine learning analysis. Chapter IV presents the KDD analysis and results. Lastly, Chapter V summarizes the conclusions and significance of the research, answers each of the research questions, and recommends additional research areas.

# II. Literature Review

## Chapter Overview

This chapter presents background information on MDI, introduces the KDD process, and discusses data mining concepts and applications. First, the literature review presents the scope of USAF infrastructure management. Second, the literature provides background information on asset management as a field of study as well as implementation within the federal government. Third, the literature review discusses the Navy's MDI model in order to establish baseline knowledge of the original methodology. Fourth, the literature review outlines the USAF adaptation of the Navy's MDI methodology. Fifth, the literature review presents limitations and purported flaws in the MDI methodology. Finally, the literature review introduces the KDD process and the field of machine learning, including techniques and applications, in order to present applicability within the context of the established research problem statement.

## USAF Real Property Portfolio and Requirements

The importance of effective asset management within the government cannot be overstated. The federal government's real property portfolio is immense, and each subordinate agency oversees a conglomeration of facilities that support unique mission sets. USAF real property assets are scattered across the globe and boast an estimated total Plant Replacement Value (PRV) of over $259 billion (DOD, 2013). Additionally, USAF real property asset conditions vary widely in both age and condition.

In 2014, over 4,700 USAF facility projects valued at $3.6 billion were submitted for funding consideration (Maddox, 2014). This is a clear indicator of the high demand

for infrastructure funding and emphasizes the importance of effective asset management practices. Developing a method to effectively allocate the limited resources in pursuit of organizational goals is a complex problem. This prioritization dilemma necessitates an overarching asset management framework and a metric linking facility risk to the USAF mission. Air Force Policy Directive (AFPD) 32-90 (2007), *Real Property Asset Management*, defines real property asset management with the following:

> Air Force real property asset management is the process of accurately accounting for, maintaining[,] and managing real property in the most efficient and economical manner in accordance with Federal Real Property Council guidance, while ensuring that the Air Force has the real property it needs for sustaining current and projected missions.

**Asset Management Background**

Infrastructure asset management is a relatively young area of study and combines aspects of engineering, business practices, and economics in order to effectively management physical assets (McElroy, 1999). Asset management is a holistic approach to managing physical assets and is not specific to any single engineering domain. Because of this wide range of applicability, asset management is often defined within the context of a specific domain such as transport, construction, electricity, and irrigation (Amadi-Echendu et al., 2010). Each of these fields approaches asset management with unique objectives, physical assets, and life-cycle requirements. Despite the unique aspects of different domains, engineering asset management is considered to be a conglomeration of concepts from commerce, business, and engineering (Amadi-Echendu et al., 2010). Amadi-Echendu et al. (2010) provide a synopsis of relevant literature in order to propose a baseline engineering asset management definition:

> The commonalities are focusing on the life-cycle of an asset as a whole, paying attention to economic as well as physical performance and risk measures, appreciating the broader strategic and human dimensions of the asset management environment, with the objective of improving both efficiency and effectiveness of resources.

Along with a common understanding, motivations for asset management practices include the effective use of resources, gaining competitive advantage, increasing profit margin, and ensuring accountability. Asset management practices are advantageous in both the public and private sectors; however, asset management is particularly relevant for government agencies due to "public demands for transparency in government decision-making, greater accountability for those decisions, and greater return-on-investment" (McElroy, 1999). Ultimately, asset management principles facilitate effective decision-making. Vanier (2001) presents the six "whats" of asset management as a means of defining asset management. The six "whats" include: (1) what do you own?, (2) what is it worth?, (3) what is the deferred maintenance?, (4) what is the condition?, (5) what is the remaining service life?, and (6) what do you fix first?. These questions provide an easily understood process for asset management implementation.

Central to Vanier's (2001) conceptual framework for infrastructure asset management are data and information technology. McElroy (1999) states that "the focus on effective asset management is argued to require an asset decision making framework that incorporates organizational structures and information technology aligned with financial and budgetary considerations" (Amadi-Echendu et al., 2010). Information technology is a cornerstone of asset management. Accurate physical asset data fuels the decision-making process in pursuit of organizational goals. Furthermore, Asset management tools and data enable the asset manager to synthesize the dynamic

relationships between organizational goals, budgets, and real property sustainment in tackling the myriad of asset management challenges.

**Asset Management Challenges**

Infrastructure asset management is a complex field with many challenges. These challenges emanate from the inherent intersection of managing physical assets, pursuing organizational goals, integrating information technology, and prioritizing funding under constrained resources all within a political environment. Woodhouse (2001) describes the complexity of asset management as the integration of "sophisticated technical solutions", management processes, and human factors. Amid the inherent complexity, infrastructure asset management is fundamentally data-centric and relies heavily on information technology. When employed effectively, information technology serves as a force multiplier. Vanier (2001) champions data and decision support tools in infrastructure asset management:

> Engineers, technical staff, administrators, and politicians all benefit if decisions about maintenance, repair and renewal are based on reliable data, solid engineering principles and accepted economic values. When reliable data and effective decision-support tools are in place, the costs for maintenance, repair and renewal will be reduced and the services will be timely, with less disruptions. These improvements will all reduce the costs of managing municipal infrastructure.

While data and decision support tools are an obvious catalyst for infrastructure asset management, the lack of quality data and support tools is a resounding message. Vanier (2001) purports that asset managers lack "literature" and "intelligent computer software" to assist in the decision making process. Additionally, Amadi-Echendu et al. (2010) present that data is a primary limitation in employing asset management principles:

> The data requirements for the decision models are very great…ideally, an information system provides continuous data on the physical and financial conditions and changes in condition of a set of assets that is being managed for some purpose.

The benefits of this optimal scenario include value-focused insight and effective decision-making. Unfortunately, "in the vast majority of organizations, the opinion of many engineers is that poor data quality is probably the most significant single factor impeding improvements in engineering asset management" (Woodhouse, 2001). Poor data quality is often the product of incorrectly entered data or simply empty data fields affording limited or totally ineffective engineering asset management support (Amadi-Echendu et al., 2010). Woodhouse (2001) concludes that "the greatest challenges for engineering asset management often do not lie in the technical aspects of implementation …rather they lie in the human element in data collection, entry and analysis." Furthermore, poor data can be attributed to a lack of indoctrination and training for personnel (Amadi-Echendu et al., 2010). Personnel at all levels of an organization must be well trained and educated in the principles and benefits of asset management practices in order to effectively manage infrastructure. Asset management education and implementation are no simple task, especially within the confines and complexities of the federal government. The federal government acknowledges the challenges associated with asset management and identified infrastructure management as a high-risk area as early as 1997 (GAO, 2003).

**Asset Management within the Federal Government**

President George W. Bush signed Executive Order (EO) 13327 on 4 February 2004 laying the foundation for infrastructure asset management principles within the federal government. Section one of the EO outlines the policy vision for asset management as (1) emphasizing efficient and economic use of real property assets and (2) assuring management accountability. Section two defines federal real property as "any real property owned, leased, or otherwise managed by the Federal Government, both within and outside the United States, and improvements on Federal lands" (Executive Order No. 13327, 2004). Section three establishes the requirement for agency Senior Real Property Officers. Senior Real Property Officers are responsible for developing and implementing the asset management planning process for their respective agency. This senior position provides a means of implementing change and assigns a responsible individual, who is accountable for an agency's asset management program (Teicholz et al., 2005). Section four establishes the Federal Real Property Council (FRPC) under the Office of Management and Budget (OMB) "to develop guidance for, and facilitate the success of, each agency's asset management plan" (Executive Order No. 13327, 2004).

EO 13327 established the FRPC as an interagency forum for collaboration in implementing asset management policy directives. The FRPC is comprised of all SRPOs, the Controller of the OMB, the Administrator of General Services, and is chaired by the Deputy Director for Management of the OMB (Executive Order No. 13327, 2004). Initially, the FRPC established four committees to focus on (1) asset management, (2) performance measures, (3) inventory, and (4) systems (Teicholz et al., 2005). These

committees developed key guidance documents including a list of ten guiding asset management principles outlined in Figure 1.


Figure 1. Asset Management Principles (Teicholz et al., 2005)

Additionally, as a means of standardizing federal real property data for asset management, the FRPC established data elements and performance measures, including MDI, as indicated in Figure 2.

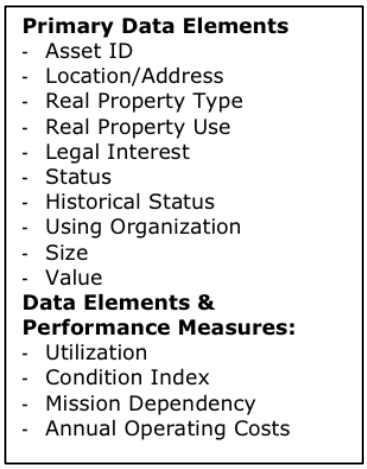



Figure 2. FRPC Data Elements, Performance Measures (Teicholz et al., 2005)

Real property data is a key tenet of asset management that continues to evolve with technology and asset management practices. NAVFAC P-78, *Real Property Inventory (RPI) Procedures Manual*, and AFI 32-9005, *Real Property Accountability and Reporting*, state the following about asset management and data (AF/A7C, 2008; NAVFAC, 2008a):

> Accurate and timely real property asset data is fundamental to effective management of assets. Real property asset data links accountability, regulatory compliance, resource requirements, and decision support. Access to the data is essential across the Defense enterprise, at all levels.

EO 13327 and the standup of the FRPC represent a catalyst for the standardization of real property data across the entire DOD. The FRPC manages recurring Real Property Inventory Reporting (RPIR) in support of the federal government's asset management framework (FRPC, 2015). Ultimately, the requirement for asset management practices within the federal government is driven by the current fiscal environment and accountability for resource allocation decisions.

**MDI Background**

The public sector is unique in that objectives are not tied to profit but the public good (Albrice et al., 2014). Federal agencies are charged with providing specific and unique services that are often hard to compare with motivations found in industry. The USAF *2023 Implementation Plan* (2013) states that the USAF is charged with delivering "decisive global vigilance, global reach, and global power, in and through air space and cyberspace anywhere on the globe at a time and place of our choosing". The plan outlines 20 capabilities to meet these objectives, each of which relies on physical

infrastructure. These mission sets are dynamic and synergistic, which contributes to the complexity of determining risk and value with respect to mission.

MDI is an attempt at solving the complex task of assigning value to physical infrastructure based on the mission or missions supported. MDI is a means of describing the "consequence of failure" associated with a real property asset in lieu of a strictly profit-driven decision (AFCEC, 2015). Mission dependency, however, is just one component used in prioritizing Sustainment, Restoration, and Modernization (SRM) funds. MDI is used in conjunction with Facility Condition Index (FCI), which is intended to describe an asset's "probability of failure" based on the asset's condition (AFCEC, 2015). Infrastructure condition is an important aspect of asset management. A Government Accountability Office report (1998) presenting leading practices in capital decision-making emphasizes the utility of condition assessments:

> Routinely assessing the condition of assets and facilities allows managers and other decision makers to evaluate the capabilities of current assets, plan for future asset replacements, and calculate the cost of deferred maintenance.

Together, MDI and FCI support the final phase of asset management decision-making: prioritization of resource allocation.

The primary objective of MDI is to optimize readiness at the lowest possible cost by focusing on critical facilities (high MDI score) that are below acceptable condition (low FCI score) (NAVFAC, 2008b). MDI's link to mission execution is important for public sector organizations as private sector objectives differ despite similar infrastructure challenges (National Research Council, 2008). The Navy was the first service in the DOD to link facilities to missions with the MDI metric (NAVFAC, 2008b):

> MDI is the standard methodology within the Naval shore establishment for determining infrastructure SRM priorities based on mission criticality from a "warfighter", operator or users point of view. It does this by evaluating the impact to the mission if the function provided by the infrastructure is interrupted or relocated. MDI is reported on a scale of 1 to 100, with 100 representing the highest mission importance.

**NAVFAC MDI Model**

Naval Facilities Engineering Service Center (NFESC) first introduced MDI in 2001 (prior to EO 13327) in collaboration with the Coast Guard's Office of Civil Engineering in Washington, D.C (Dempsey, 2006). MDI is an operational risk management metric that seeks to link facilities to mission execution (Dempsey, 2006). The original intent of the MDI metric was to provide actionable information for maintenance, repair, sustainment, resource allocation, divestiture, and physical security (Dempsey, n.d.). The Navy's MDI metric facilitates these efforts by assessing interruptibility, relocateability, and replaceability of real property assets as viewed by senior level decision makers responsible for operational and facility decisions. NAVFAC completed MDI assessments at all major navy bases by August of 2007 with the intent to update on a three-year cycle (NAVFAC, 2008b). Additionally, the Navy established a process for updating or revising MDI values if required prior to the standard three-year cycle in order to maintain currency for facility decision making.

Because the Navy MDI is based on deliberate communication with mission and facility stakeholders, it facilitates the capturing of tacit knowledge through survey questions. Data collectors pose four questions to stakeholders in order to assess mission criticality. Two questions assess mission criticality with respect to mission-intradependency (dependencies within a mission) and two questions on mission-

interdependency (dependencies between missions). Intradependency seeks to capture the dependencies *within* a given functional area while interdependency captures dependencies *between* functional areas. The two primary concerns with interdependency and intradependency are maximum interruption durations and the degree of difficulty associated with relocation or replaceability. The four MDI survey questions are presented in Table 1. The four MDI survey questions are answered qualitatively in the context of time and difficulty. Table 2 and Table 3 present the definitions and possible responses for the interruptibility and relocateability and replaceability survey questions, respectively.

Table 1. Navy MDI Survey Questions (Antelman, 2008)

| Primary Topic | Measure | Metric | Question # | Question Verbiage |
|---|---|---|---|---|
| Intradependency | Interruptibility | Duration (Time) | 1 | How long could the "functions" supported by the (facility, structure, or utility) be stopped without adverse impact to the mission? |
| | Relocateability | Difficulty | 2 | If your (facility, structure, or utility) was not functional, could you continue performing your mission by using another (facility, structure, or utility), or by setting up temporary facilities? |
| Interdependency | Interruptibility | Duration (Time) | 3 | How long could the services provided by (named functional Area) be interrupted before impacting your mission readiness? |
| | Replaceability | Difficulty | 4 | How difficult would it be to replace or replicate the services provided by (named functional Area) with another provider from any source? |

Table 2. Response Options for Interruptibility (Antelman, 2008)

| Interruptibility Responses (Time) | |
|---|---|
| **Response** | **Definition** |
| None (N) | The functions performed within the facility must be maintained continuously (24/7) |
| Urgent (U) | Minutes not to exceed 30 minutes |
| Brief (B) | Minutes or hours not to exceed 24 hours |
| Short (S) | Days not to exceed 7 days |
| Prolonged (P) | More than a week |

Table 3. Response Options for Relocateability and Replaceability (Antelman, 2008)

| Relocateability and Replaceability Responses (Difficulty) | |
|---|---|
| **Response** | **Definition** |
| Impossible (I) | There are no known redundancies or excess/surge capacities available, or there are no viable commercial alternatives – only this site/command can provide these services |
| Extremely Difficult (X) | (there are minimally acceptable redundancies or excess/surge capacities available, or there are viable commercial alternatives, but no readily available contract mechanism in place to replace the services) |
| Difficult (D) | Services exist and are available, but the form of delivery is ill defined or will require a measurable and unbudgeted level of effort to obtain (money/man-hours), but mission readiness capabilities would not be compromised in the process |
| Possible (P) | Services exist, are available, and are well defined |

With the survey responses collected from facility stakeholders, risk matrices are utilized to the obtain intradependency ($MD_w$) and interdependency ($MD_b$) values as indicated in Figure 3 and Figure 4, respectively. These $MD_w$ and $MD_b$ values as well as the total number of interdependencies, "n", are then inserted into the MDI formula, presented in Equation 1. The product of the MDI equation is a quantitative score (index) between 0-100 where a higher score equates to more severe consequences. Ultimately,

this consequence of failure score (MDI) is assigned to individual assets within the Navy real property portfolio.

| MISSION INTRA-DEPENDENCY SCORE | | | | | |
|---|---|---|---|---|---|
| $MD_W$ | | Q1: Interruptability | | | |
| | | Immediate (24/7) | Brief (min/hrs) | Short (<7days) | Prolonged (>7days) |
| Q2: Relocateability | Impossible | 4.0 | 3.6 | 3.2 | 2.8 |
| | Extremely Difficult | 3.4 | 3.0 | 2.6 | 2.2 |
| | Difficult | 2.8 | 2.4 | 2.0 | 1.6 |
| | Possible | 2.2 | 1.8 | 1.4 | 1.0 |
| $MD_W$ = Mission Dependency Within a Command's AoR | | | | | |

Figure 3. NAVFAC Mission Intradependency Matrix (Dempsey, 2006)

| MISSION INTER-DEPENDENCY SCORE | | | | | |
|---|---|---|---|---|---|
| $MD_B$ | | Q3: Interruptability | | | |
| | | Immediate (24/7) | Brief (min/hrs) | Short (<7days) | Prolonged (>7days) |
| Q4: Replaceability | Impossible | 4.0 | 3.6 | 3.2 | 2.8 |
| | Extremely Difficult | 3.4 | 3.0 | 2.6 | 2.2 |
| | Difficult | 2.8 | 2.4 | 2.0 | 1.6 |
| | Possible | 2.2 | 1.8 | 1.4 | 1.0 |
| $MD_B$ = Mission Dependency Between Commands | | | | | |

Figure 4. NAVFAC Mission Interdependency Score Matrix (Dempsey, 2006)

$$MDI = 26.54 \times \left[ MD_w + 0.125 \times \frac{1}{n} \sum_{i=1}^{n} MD_{bi} + 0.1 \times \ln(n) \right] \quad (1)$$

As stated previously, the $MD_w$ and $MD_b$ components represent the mission intradependency and mission interdependency values, respectively. Figure 5 and Figure 6 present graphical depictions of a hypothetical scenario of intradependencies (within a functional area) and interdependencies (between functional areas) for a generic USAF functional area, the operations group. For example, the air traffic control tower in Figure 5 received an interruptibility response of "urgent" (not to exceed 30 minutes) and a relocateability/replaceability response of "extremely difficult". These two responses make up the intradependency score for the asset using matrix in Figure 3. Next, the interdependency scores for all applicable interdependencies displayed in Figure 6 are averaged in Equation 1 yielding a single $MD_b$ value. Lastly, the n-component in the MDI equation (Equation 1) denotes the number of mission interdependencies, which are represented by the "links" between functional areas in Figure 6. The natural log function provides a scoring scale for the total number of interdependencies identified between functional areas. More specifically, as the number of interdependencies increases, the n value in the equation is constrained.

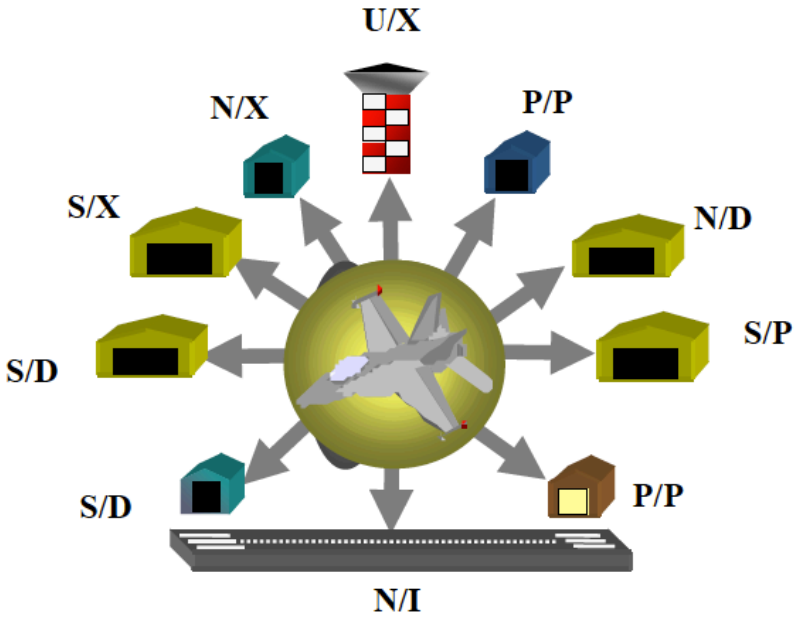



Figure 5. Hypothetical Operations Group Intradependencies (Antelman, 2008)

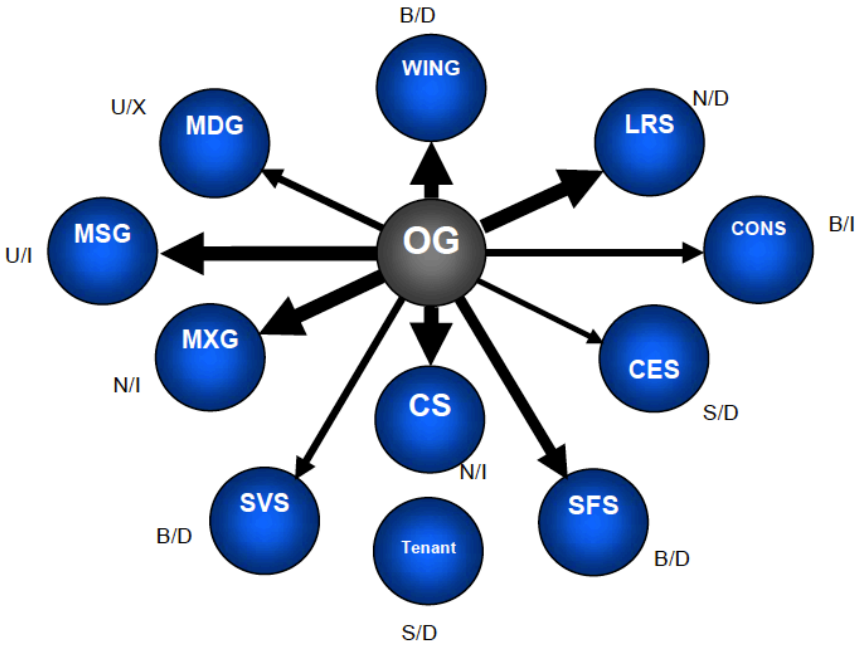



Figure 6. Hypothetical Operations Group Interdependencies (Antelman, 2008)

**USAF MDI Implementation**

The USAF began investigating the MDI metric in 2008 using the Navy methodology as the primary reference guide (AFCEC, 2015). At that time, there was no clear government or industry standard to calculate MDI (Madaus, 2009). The transition to asset management principles through EO 13327 served as the catalyst for incorporating the MDI metric into USAF business practices. The goal of MDI implementation and the "Asset Optimization Concept" was to move toward a common approach that would enable an Air Force-wide analysis of real property requirements (Madaus, 2009).

At that time, the USAF determined that the Navy had a proven MDI methodology but that it was "complex and expensive" (Madaus, 2009). Because the Navy methodology requires in-person interviews, there are significant costs associated with the data collection effort. The United States Army Corps of Engineers (USACE) estimated that MDI data collection would cost between $40,000 to $75,000 per installation depending on the number of mission sub-elements (Michael Grussing et al., 2010). With 185 installations world-wide, the estimated initial cost for data collection is $7.4 to $13.9 million and the estimated annual cost is $2.5 to $6.9 million for recurring assessments (Nichols, 2015). Antelman (2008) purports that both the Navy and NASA have employed internet-based surveys for data collection yielding up to 50 percent cost savings.

Before making a decision on enterprise-wide MDI implementation, the USAF worked with Naval Facilities Engineering Service Center to execute two "proof of concept" evaluations at Langley AFB and Fairchild AFB (Antelman, 2008). Antelman (2008) cites the following as motivation for the USAF proof of concept:

> The current process used by the Air Force lacks a disciplined driven asset strategy and metrics that link assets to its missions, thereby making it difficult to make prudent, long-term funding decisions.

The *MDI Refinement Playbook* (2015) states that these beta tests were found to be generally accurate in most cases. The results of the MDI proof of concept for Fairchild AFB and Langley AFB are presented in Figure 7 and Figure 8, respectively.

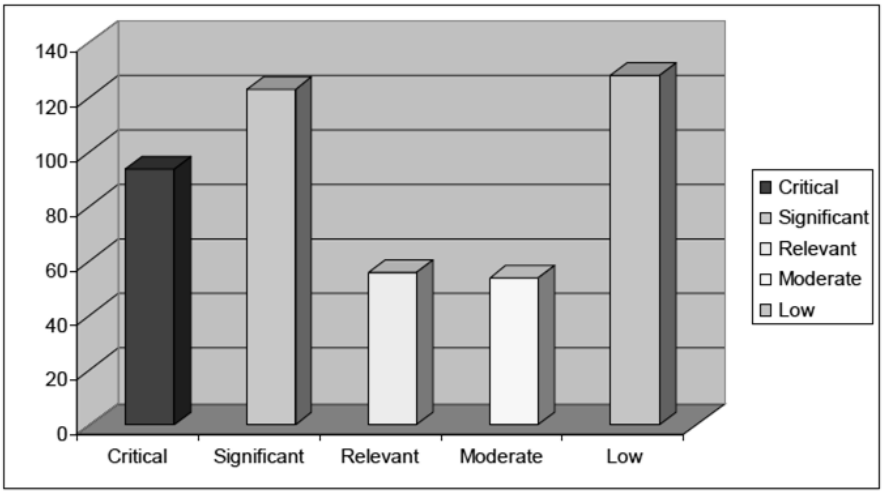



Figure 7. MDI Score Distributions at Fairchild AFB (Antelman, 2008)

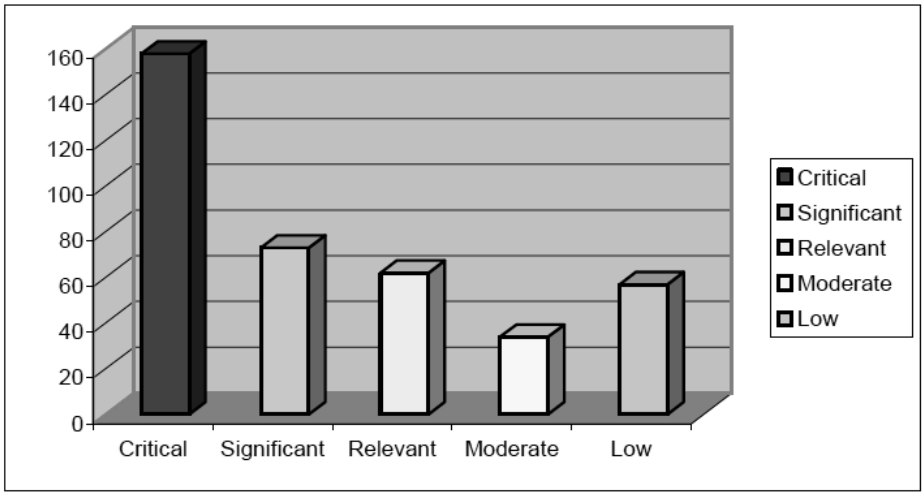



Figure 8. MDI Score Distributions at Langley AFB (Antelman, 2008)

Ultimately, the USAF opted to implement an MDI methodology that combined aspects of the Navy and the National Park Service (NPS) MDI methodologies (AFCEC, 2015; Madaus, 2009). The new USAF MDI implementation employed existing real property Category Codes (CATCODES) to assign MDI values to facilities. At that time, the NPS was also assigning MDI values based on CATCODES. This method negated the requirement for the extensive data collection process in use by the Navy.

In lieu of data collection, the USAF related Navy real property CATCODEs and USAF real property CATCODES using DOD four-digit Facility Analysis Codes (FACs) (AFCEC, 2015; Madaus, 2009). The four-digit FACs are equivalent across each of the DOD services whereas CATCODES are unique to each service (DOD, 2013). Through this process, Navy facilities slated for demolition or disposal with an MDI less than 25 were ignored. For a given USAF CATCODE, the USAF selected the average MDI value when the standard deviation was less than 10. In situations where the standard deviation was greater than 10, USAF personnel performed a manual review of the data and selected the most appropriate MDI based on subject matter expert judgment, the MDI beta test results, and Mission Area Rating Matrix (MARM) groupings and priorities. At that time, MARM groupings were used to develop inputs for the USAF facility investment strategy and program objective memorandum (POM) (Sharp, 2002). The MARM categories include Primary Mission, Mission Support, Base Support, and Community Support. Table 4 presents examples of facilities included in each of the MARM categories. The USAF MDI adaptation was implemented in February of 2009 as an “interim” method with the understanding that the MDI results would be “less granular” than collecting field data (Madaus, 2009).

Table 4. MARM Categories and Examples (Madaus, 2009)

| MARM Category | Facility Examples |
|---|---|
| Primary Mission | Airfield pavements, navigational aids, airfield electrical distribution, operational squadron operations centers, missile alert facilities, academic facilities at AETC and USAFA, base operations center, research laboratories, depot maintenance shops at AFMC bases |
| Mission Support | Primary emergency response facilities (immediate life support and rescue facilities such as central security control and fire department), aircraft maintenance facilities, test stands, fire stations, base communications center, medical functions, primary water and electrical distribution systems |
| Base Support | Admin facilities, chapels, headquarters buildings, supply warehouse, civil engineering shops, photo lab, fitness center, essential feeding facilities, dormitories, billeting |
| Community Support | Housing, lodging facilities, theaters, youth centers and child development centers, credit unions, aero club, exchange facilities, recreation site lodging, consolidated clubs, museums. |

## USAF MDI Adjudication Process

During the initial MDI investigation period (2008-2009), the MDI project team understood that the interim MDI methodology would produce inferior results as compared to collecting data from the field (Madaus, 2009). This statement proved true as installations and MAJCOMs began to identify facilities with inaccurate MDI values that had the potential to negatively impact funding allocation. Both USAF Civil Engineers and mission operators identified discrepancies with facility MDIs. In July 2013, the CE Board cited large-scale improvement of the MDI as a priority (AFCEC, 2015). Additionally, the FY 15-21 Air Force Activity Management Plan (AFAMP)/Air Force Comprehensive Asset Management Plan (AFCAMP) Business Rules identified 39 CATCODES with inconsistent MDI values and allowed for reviews of specific assets (AFCEC, 2015).

In response to the feedback from facility stakeholders and the widespread need to correct certain facility MDIs, the Air Force Civil Engineer Center (AFCEC) published the *MDI Refinement Playbook* in 2014. This playbook established a standard process for MDI adjudication (AFCEC, 2015). To date, over 1,000 facilities have been submitted for adjudication as seen in Figure 9. Furthermore, the adjudication data collected thus far potentially reveals that some Major Commands (MAJCOMs), including Air Force Global Strike Command (AFGSC), have been affected more significantly than others. The AFCEC MDI adjudication data is presented in Figure 10 and indicates the frequency of MDI adjudication requests by the respective MAJCOMs.

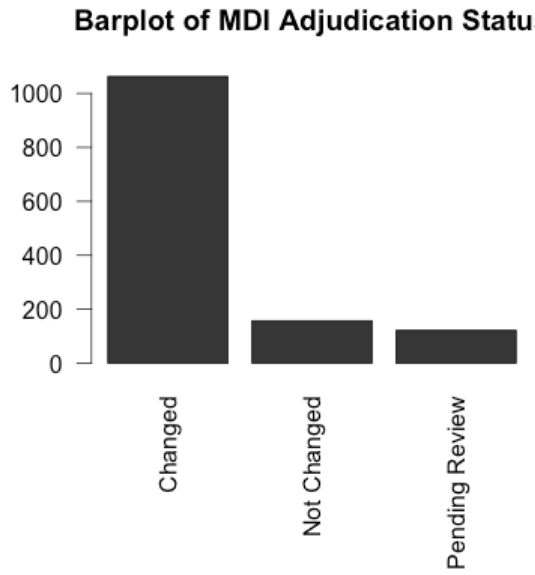



Figure 9. MDI Adjudication Status (Current as of Aug 2015)

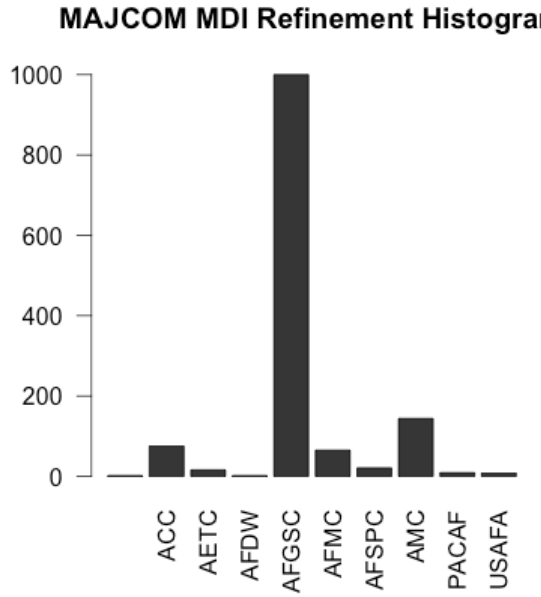



Figure 10. MAJCOM MDI Refinement Histogram (Current as of Aug 2015)

The MDI refinement process has three primary steps as outlined in the *MDI Refinement Playbook* (2015). Six distinct parties contribute to the overall process: (1) installation CE personnel, (2) installation functional experts, (3) the installation commander, (4) MAJCOM CE personnel, (5) MAJCOM functional experts, and (6) AFCEC/CPA. The first step is to identify MDI discrepancies, which is presented in Figure 11. In the second step, AFCEC/CP adjudicates the changes (Figure 12). The last step is to update approved MDI changes in the real property records (Figure 12).

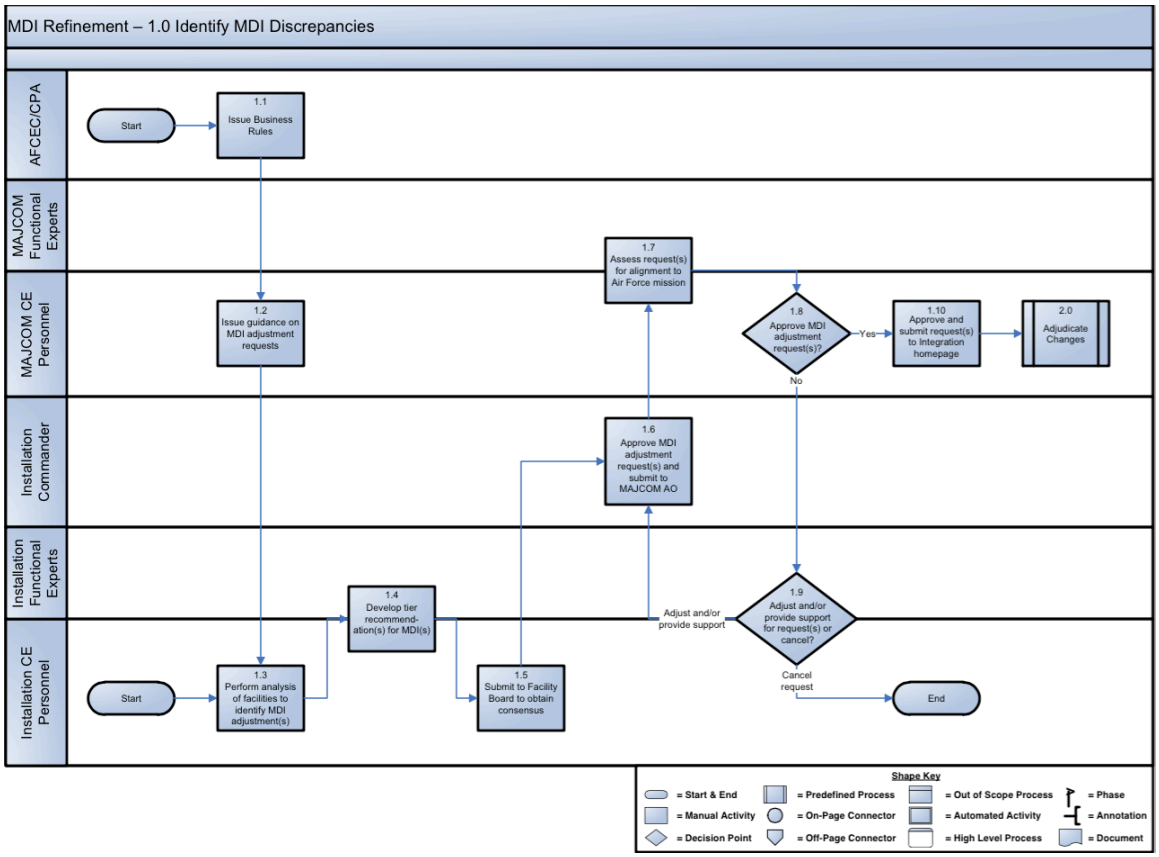



Figure 11. MDI Refinement Process: Identify Discrepancies (AFCEC, 2015)

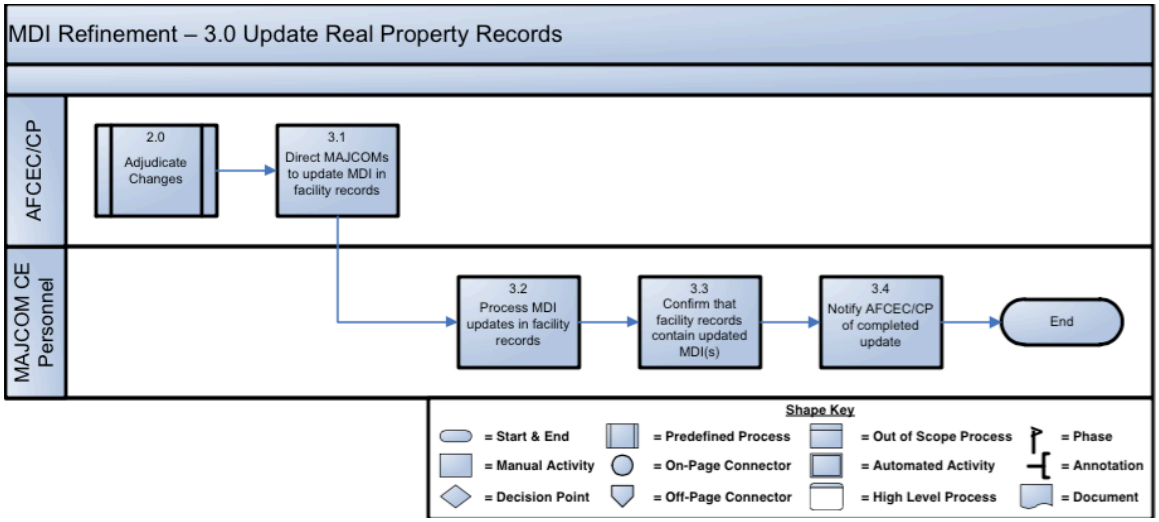



Figure 12. MDI Refinement Process: Update Real Property Records (AFCEC, 2015)

The MDI refinement process provides installations with a means of rectifying incorrectly assigned MDI values. This additional process illuminates the fact that the MDI methodology, implemented in 2009, does not adequately describe the value that each facility brings to the performance of the mission due to a lack of underlying data and over-generalization.

**Navy MDI Limitations**

While numerous limitations have been identified pertaining to the application of the MDI methodology within the DOD, four primary issues dispute the MDI methodology's effectiveness:

1. Questionable value of risk matrices.
2. Inconsistent with Operational Risk Management (ORM) practices.
3. Mathematical flaws in calculating MDI values.
4. Lack of analytical support for the MDI equation.

First, the Navy MDI methodology asserts that "MDI uses Operational Risk Management techniques of probability and severity and applies them to facilities in terms of interruptibility, relocateability, and replaceability" (Antelman, 2008). ORM principles are not a new concept within the Department of Defense. ORM is incorporated into many facets of military operations and each service maintains some form of ORM instruction or manual presenting the step-by-step process for conducting risk management analyses. The Navy instruction for ORM is OPNAVINST 3500.39C (USN, 2010), *Operational Risk Management*. This instruction presents an iterative five-step process for risk management: (1) identify the hazards, (2) assess the hazards, (3) make risk decisions, (4) implement controls, and (5) supervise. This ORM process employs

risk matrices for a qualitative assessment of probability and severity for a given hazard. Cox (2008) acknowledges the popularity of risk matrices (in both public and private sector applications) but identifies four significant concerns with their use:

1. Risk matrices provide poor resolution.
2. Risk matrices can mistakenly assign higher qualitative ratings to quantitatively smaller risks.
3. Risk matrices produce suboptimal resource allocation.
4. Risk matrices are the product of ambiguous inputs and outputs that require subjective interpretation.

Cox's most relevant assertion with respect to MDI is that "calculating optimal risk management resource allocations requires quantitative information beyond what a risk matrix provides" (Cox, 2008). Here, Cox is clearly stating that quantitative data as a requirement for risk management resource allocations and risk matrices are not the optimal tool.

Second, Kujawski and Miller (2009) identify that the MDI methodology deviates from the Navy ORM instruction with respect to assessing hazards with probability and severity. The ORM instruction explicitly defines probability and severity qualitatively using letters and roman numerals (based on an ordinal scale) whereas the MDI methodology employs real numbers. Probability and severity are represented as qualitative values in risk management for the express purpose of avoiding enumeration, a "risk assessment pitfall" (USN, 2010). Additionally, Kujawski and Miller (2009) contend that the Navy's MDI method "makes no attempt to quantify probability and includes no discussion of mishap likelihood" in accordance with the ORM process.

Third, the MDI methodology employs mathematics with ordinal numbers, via the MDI equation, to arrive at the final MDI score. Kujawski and Miller (2009) point out

that performing mathematics on ordinal numbers to arrive at MDI values is not an acceptable practice and produces ambiguous results (Kujawski & Miller, 2009). Similarly, Hubbard (2014) purports that popular risk management methods that employ scores actually introduce error. "Scores are merely ordinal, but many users add error when they treat these ordinal scores as a real quantity…a higher ordinal score means 'more' but doesn't say how much more" (Hubbard, 2014). While presenting mission dependency on a 100-point scale is an attractive option, the approach utilized in obtaining these values may actually introduce additional error.

The fourth limitation is the lack of analytical support in the development of the actual MDI equation. Kujawski and Miller (2009), drawing from Edward Tufte's (2006) book *Beautiful Evidence*, argue this point based on the assertion that the MDI methodology is the product of field-testing instead of documented analytical methods:

> The method for validation via field-testing is not described. Any analysis involving validating fitted polynomial curves of quantitative data requires, at a minimum, the number of samples collected, the raw data matrix, equations of the fitted models along with plotted curves and plotted raw data, quality of the fit of the curves and substantive meaning of the estimated models.

There is limited documentation on the development of the MDI equation. The MDI equation seems to take on the form of a "black box" that produces what appear to be reasonable values. Unfortunately, the justification behind the weighting of the coefficients and the supporting evidence for the underlying phenomenon remains elusive.

**USAF MDI Limitations**

Because USAF MDI scores were derived from Navy MDI scores (mapped via FAC codes), the limitations associated with the Navy MDI model also apply to the USAF implementation. Given the nature of the USAF MDI implementation without data collection, additional issues were introduced compounding the inconsistencies and further deviating from the original intent of the MDI metric. When the USAF first implemented MDI, the methodology was employed as an interim approach. Today, it is generally understood across the USAF civil engineer career field that "MDI isn't perfect" (Maddox, 2014). Since the USAF's implementation in 2008, specific limitations include the necessary MDI adjudication process, MDI inflation, discord across MAJCOM priorities, and biased assumptions.

The first limitation is the fact that the interim MDI methodology, adapted from the Navy, produced arguably lower fidelity than what data collection from the field could have provided. This interim solution led to the creation of the MDI adjudication process to correct MDI values. This MDI adjudication process is, in a way, data collection from the field that requires resources in the way of time and manpower. Furthermore, the MDI adjudication process is primarily focused on making sure that mission critical infrastructure ("tier 1") has the appropriate MDI value (AFCEC, 2015). Given an incorrectly assigned MDI value for mission critical infrastructure asset, base personnel would be motivated to navigate the MDI adjudication process to increase the MDI score. However, given an incorrectly assigned non-critical infrastructure asset, base personnel may not be motivated to invest the time and resources to decrease an asset's MDI score. This leads to the second limitation, MDI inflation.

The second limitation is MDI inflation. MDI inflation threatens to limit the metric's contribution to decision support. MDI values are intended to represent an index with a range of values between 0 and 100. Nichols (2015) identifies that the range of MDI values within the USAF real property portfolio is actually between 32 and 99, which diminishes decision support value by compressing the real property assets into a smaller range of values.

The third limitation is the natural discord between MAJCOMs perspectives on mission critical infrastructure. Nichols (2015) presents that USAF MAJCOMs with unique mission sets do not necessarily fare as well as strictly operational MAJCOMs when MDIs are assigned based on CATCODE alone. A specific example of this scenario is Air Education and Training Command (AETC) whose mission is to "recruit, train and educate Airmen to deliver airpower for America" (AETC, 2015). The AETC mission requires facilities such as dorms, classrooms, and training facilities that may not have a CATCODE matched with a high MDI value. For this reason, AETC facility MDI values may not accurately capture unique MAJCOM mission sets and support requirements. AFGSC is also an example of a MAJCOM with a unique mission set. AFGSC's mission is to "develop and provide combat-ready forces for nuclear deterrence and global strike operations…" (AFGSC, 2015). AFGSC is responsible for Intercontinental Ballistic Missile (ICBM) operations with facilities that are spread out across large land areas. As of August 2015, AFGSC had identified over 1,000 facilities with MDI values that did not adequately convey the mission criticality. This is indicative of the inherent limitation with CATCODE-assigned MDI values as well as an unintended consequence with respect to the interim MDI methodology and specific MAJCOM mission sets. This

situation was a significant driver in establishing the additional MDI adjudication process to correct MDI values (AFCEC, 2015).

The final MDI limitation is that of biased assumptions inherent in the MDI methodology. A general assumption, given the USAF adaptation of the Navy's MDI method, is that the Navy and USAF consider the same types of infrastructure mission critical. Given the nature of the Navy's mission, this assumption may not be true in all cases. Another potential bias with the USAF MDI application is the assumption that all airfields are equally important. This was codified by assigning an MDI of 99 to the CATCODE for airfield runways. This rule has been in place since the implementation of the USAF MDI methodology. Generally speaking, this rule is in direct agreement with USAF flying mission sets, however, not all missions identified in the USAF 2023 plan are tied to airfield runways (USAF, 2013). Some installations with an extant airfield currently do not support missions that require an active runway. This MDI rule is inconsistent with mission critical infrastructure priorities of MAJCOMS with non-flying missions. The limitations associated with the USAF application of the MDI metric stem from a lack of data and MDI values that originated from an outside source.

**Data Facilitates Effective Asset Management**

Data is a proven force multiplier for effective management of real property portfolios. Albrice, Branch, and Lee (2014) employ physical and financial attributes to (1) draw correlations between data sets, (2) identify patterns in the data, (3) classify and organize data into groups, (4) benchmark individual assets or facilities against Key Performance Indicators (KPIs), (5) and to establish prioritization schemes. More

specifically, a business case for resource allocation decisions can be developed using correlations, patterns, and multivariate analyses with 18 pertinent facility attributes. The 18 facility attributes employed in resource allocation are listed in Table 5.

Table 5. Facility Attributes for Resource Allocation (Albrice et al., 2014)

| | **Attribute** |
|---|---|
| 1 | Age of the Facility |
| 2 | Size of the Facility |
| 3 | Reproduction Value (CRN) |
| 4 | Mission Dependency Index |
| 5 | Backlog of Deferred Maintenance (FCI) |
| 6 | Capital Load over Tactical Horizon (5 years) |
| 7 | Capital Load over Strategic Horizon (30 years) |
| 8 | Adequate Replacement Reserves |
| 9 | Ownership Structure (Freehold or Leasehold) |
| 10 | Function |
| 11 | Primary, Secondary and Tertiary Uses |
| 12 | Number of Systems and Assets in the Facility |
| 13 | Date of Last Condition Assessment |
| 14 | Post-disaster Designation (PD) |
| 15 | Revenue generating capacity and lease income |
| 16 | Energy Use Intensity (EUI) and Efficiency (BEPI) |
| 17 | Geographical Location and Bundled Co-locations |
| 18 | Functional Obsolescence (FNI) |

Albrice et al. (2014) provide four examples of complex decisions that asset managers encounter. The first question is continued investing in facility sustainment versus rebuilding with a new facility. The second question deals with "adaptive renewal opportunities" and whether or not a specific infrastructure component should be replaced with a more efficient option. The third question deals with running a component, asset, or facility to beyond its intended life span or, ultimately, running to failure. The fourth question deals with determining an ideal ownership to leasehold ratio. Additionally, the authors reveal that resource allocation decisions benefit from correlations identified between the variables (Albrice et al., 2014). With these specific questions in mind, the

authors utilized the 18 pertinent data features and developed a nine-step multivariate decision support tool to build a business case for resource allocations. This emphasizes the fact that databases are indispensable for effective asset management.

**Real Property Databases**

The DOD is accountable to the executive branch for implementing asset management principles as outlined in EO 13327. One of the key tenets of asset management is the identification of real property inventory for accurate reporting (Vanier, 2001). Communicating data across large organizations is a challenge and necessitates strict real property accountability and inventory reporting guidance. The guiding documents for Real Property Inventory (RPI) reporting are EO 13327 (2004), Federal Real Property Asset Management, Department of Defense Instruction (DODI) 4165.70 (2005), Real Property Management, and DODI 4165.14 (2014), Real Property Inventory and Forecasting. Additionally, the FRPC publishes annual guidance for federal agency real property inventory and reporting. Execution of these real property inventory and reporting requirements necessitates agency-specific information systems and databases for real property records. Each agency's real property database represents a catalog of facts about specific infrastructure assets. With existing data mining techniques, real property databases may serve as a potential source of un-tapped knowledge for infrastructure mission criticality.

**Navy Real Property Data**

The Navy's real property inventory is governed by document P-78, Real Property Inventory (RPI) Procedures Manual (NAVFAC, 2008a). P-78 was developed in accordance with requirements under EO 13327 and describes the Navy's official real property database known as the internet Navy Facilities Asset Data Store (iNFADS). The iNFADS database is "the system which provides the means by which data on Navy and Marine Corp property is collected, processed, stored and displayed for its facilities" (NAVFAC, 2008a).

**USAF Real Property Data**

Following the implementation of asset management within the federal government via EO 13327, the USAF incorporated the major changes and vision of asset management in Air Force Policy Directive (AFPD) 32-90, *Real Property Asset Management* (USAF, 2007). The AFPD outlines the asset management vision and inventory reporting requirements with the following (USAF, 2007):

> The Air Force will maintain an accurate inventory of its real property in accordance with Federal Real Property Council, DOD, and Air Force instructions. Real property used by the Air Force will be reported on the annual Air Force Financial Statement. The Air Force will record fiscal, physical, legal, environmental, and geospatial information on real property assets to which the Air Force has legal interest. Data from real property inventories and accountability will serve as the basis for current sustainment and future capital investments.

AFI 32-9005, *Real Property Accountability and Reporting* (2008), is the implementation document that fulfills directives from AFPD 32-90, DODI 4165.70, and DODI 4165.14. Additionally, AFCEC developed the *Real Property Accountability and*

*Inventory Playbook* (2014b) to provide installation Real Property Officers (RPO) with support and guidance on the tasks associated with USAF real property accountability.

Similar to the Navy's iNFADS system, the USAF employs the Automated Civil Engineer System, which is defined with the following excerpt from AFI 32-9005.

> ACES (Automated Civil Engineer System)—The current system used by civil engineering personnel to account for and manage AF assets. ACES is the original 'book of entry' for financial accounting in terms of original acquisition cost and cost of any major improvements over the statutory threshold under the Chief Financial Officers Act of 1990.

The ACES Real Property module, ACES-RP, serves as the USAF's official real property database of record (USAF, 2008).

**Knowledge Discovery in Databases (KDD)**

In the 1990s, the booming data paradigm drove an "urgent need" for "tools to assist humans in extracting useful information from the rapidly growing volumes of digital data" (U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996a). An entire field known as Knowledge Discovery in Databases (KDD) was born out of the new data analysis limitations as manual data analysis across a myriad of fields was quickly becoming unrealistic.

Fayyad, Piatetsky-Shapiro, and Smyth authored an article in 1996 that provides an overview of the field of KDD: *Knowledge Discovery in Databases for Extracting Useful Knowledge from Volumes of Data* (U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996b). The authors purport that large databases are not inherently valuable and label them as a "dormant potential resource" (U. Fayyad et al., 1996b). KDD aims to rectify this database dormancy by discovering useful knowledge through a deliberate process.

The authors define the KDD process as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (U. Fayyad et al., 1996b). The KDD process includes the following nine steps: (1) learning the application domain, (2) creating a target data set, (3) data cleaning and preprocessing, (4) data reduction and projection, (5) choosing the function of data mining, (6) choosing the data mining algorithms, (7) data mining, (8) interpretation, and (9) using discovered knowledge (U. Fayyad et al., 1996b). To emphasize the importance of a process-centered approach, the authors maintain that applying data mining techniques without implementing the other KDD steps is “dangerous” and can lead to identifying “meaningless patterns” (U. Fayyad et al., 1996b).

Of the nine KDD steps, the data mining step is the true crux of knowledge discovery. There are two primary goals in data mining, (1) prediction and (2) description or inference (U. Fayyad et al., 1996b). These goals cater to different applications and data mining techniques. Additionally, the authors point out some noteworthy challenges encountered in the data mining realm at the time of publication. These challenges included: (1) massive databases and high dimensionality, (2) user interaction and prior knowledge, (3) overfitting and assessing statistical significance, (4) missing data, (5) understandability of patterns, (6) managing changing data and knowledge, (7) integration with other systems, and (8) nonstandard, multi-media, and object oriented data (U. Fayyad et al., 1996b). Because this article was published in 1996, the momentum in data mining research has fostered progress in these areas of concern.

## Data Mining Background

Computer system technologies have greatly increased over the past half-century serving as a catalyst for innovative solutions to complex problems. Driven by this technology wave, the tools of the trade in the data science world are continually evolving. "Traditionally, it was the responsibility of business analysts, who generally use statistical techniques" (Bose & Mahapatra, 2001).

Furthermore, data science exploded following the introduction of the internet to general users in 2000 (Liao, Chu, & Hsiao, 2012). This paradigm shift in the world of information necessitated more effective and efficient methods of knowledge management technologies. IBM reports that 2.5 quintillion bytes of data are created every day and that "90% of the data in the world today has been created in the last two years alone" (IBM, 2015). This proliferation of data and databases necessitates what are coined as "data mining" techniques in order to use "information and knowledge intelligently" (Liao et al., 2012).

## Data Mining Literature Review

Liao, Chu & Hsiao (2012) conducted a literature review using five journal databases and a keyword search for "data mining technique" ultimately identifying 14,972 articles authored between 2000 and 2011. This team then narrowed down the pool to 216 articles from 169 journals, all of which related specifically to "data mining application." Of these articles, the authors identified nine categories of data mining applications: (1) neural networks, (2) algorithm architecture, (3) dynamic prediction-based, (4) analysis of systems architecture, (5) intelligence agent systems, (6) modeling,

(7) knowledge-based systems, (8) system optimization, and (9) information systems (Liao et al., 2012). The time period analyzed in this literature is significant due to dynamic nature of progress in the field of data mining.

Advances in computer technology and the proliferation of data bases promote and necessitate data mining techniques in order to use "information and knowledge intelligently" (Liao et al., 2012). Data mining techniques are broken into numerous methods including generalization, characterization, classification, clustering, association, data visualization, among others. Additionally, there are multiple types of databases these techniques can be applied to: relational, transactional, object oriented, spatial and active databases, and global information systems (Liao et al., 2012). A noteworthy milestone in the history of data mining was the introduction of the Internet to general users, which drastically increased the availability of information and communication technology (Liao et al., 2012). In their literature review, Liao et al. (2012) identified key trends in data mining techniques based on a keyword search from a selection of journal articles. Table 6 presents the data mining trends.

Table 6. Data Mining Keyword Trends, 2000-2011 (Liao et al., 2012).

| Keyword | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data mining | 1 | 0 | 6 | 5 | 8 | 8 | 12 | 16 | 10 | 16 | 16 | 16 | 114 |
| Decision tree | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 2 | 4 | 3 | 4 | 17 |
| Artificial neural network | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 0 | 2 | 4 | 2 | 3 | 22 |
| Clustering | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 1 | 0 | 2 | 1 | 1 | 9 |
| Association rule | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 3 | 0 | 1 | 8 |
| Artificial intelligence | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 1 | 6 |
| Bioinformatics | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 4 |
| Customer relationship management | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 4 |
| Fuzzy logic | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 4 |
| Total | 2 | 1 | 10 | 9 | 11 | 15 | 15 | 24 | 16 | 32 | 27 | 26 | 188 |

These keyword trends provide an indication of both the progression of data mining techniques for the selected time period. Data mining, decision tree, and artificial neural

networks claim the top three positions with respect to total usage. Furthermore, the keyword frequencies indicate a noticeable increase over the 10-plus year time period.

Over time, data mining has proven its value and applicability across a myriad of applications. The articles analyzed in the Liao et al. (2012) literature review span many disciplines including engineering, biology, medicine, finance, social sciences and business. Liao et al. (2012) predict that going forward, data mining techniques will continue to progress and become "more expertise-oriented" and "problem-centered." Given that real property databases exist for both the Navy and Air Force, there are many data mining algorithms that could be applied to the MDI problem.

## Chapter Summary

This literature review provided a history of the federal government's shift to asset management principles, outlined federal agency MDI methodologies, presented limitations and purported fallacies associated with MDI and, lastly, presented KDD and machine learning as a mechanism for further understanding of the relationships between MDI and real property data. Linking real property assets to mission criticality is a highly complex task. As such, there is no agreed upon methodology for assigning mission criticality to real property assets within the public or private sector. Further investigation is warranted and currently available real property data may provide a better understanding of existing prioritization methods. The next chapter presents the methodology for investigating the connection between real property data and mission dependency.

# III. Methodology

## Chapter Overview

This chapter introduces the knowledge discovery in databases (KDD) process as the overarching framework employed in answering the five research questions. Next, the chapter discusses the specific procedures and rationale behind steps one through five of the KDD process leading up to the analysis and data mining specific steps. The subsequent chapter presents the data mining analysis and results from steps six through nine of the KDD process.

## Knowledge Discovery in Databases (KDD)

The explosive growth of technology and data over recent decades serves as the motivation for a codified knowledge discovery process specific to databases (U. Fayyad et al., 1996b; Frawley, Piatetsky-Shapiro, & Matheus, 1992). KDD is a holistic approach aimed at discovering knowledge from data (U. Fayyad et al., 1996b). Maimon and Rokach (2005) describe KDD as "an automatic, exploratory analysis and modeling of large data repositories." While KDD is sometimes considered synonymous with data mining, KDD encompasses numerous fields of study including machine learning, pattern recognition, statistics, artificial intelligence, data visualization, and information retrieval (Frawley et al., 1992). Furthermore, Fayyad et al. (1996a) purport that KDD is distinguishable from other fields because KDD is focused on the overarching process necessary for knowledge discovery.

The KDD process includes nine steps: (1) learning the application domain and establishing goals, (2) creating a target data set, (3) data cleaning and preprocessing, (4) data reduction and projection, (5) choosing the function of data mining, (6) choosing the data mining algorithms, (7) data mining, (8) interpretation and evaluation, and (9) using the discovered knowledge (U. Fayyad et al., 1996b). Prior to delving into the KDD process, pertinent definitions are necessary to understand the methodology. Fayyad et al. (1996a) present the following definitions associated with KDD:

> *KDD* is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.
>
> *Data* are defined as a set of facts.
>
> *Patterns* are defined as an expression in some language describing a subset of the data or a model applicable to the subset.
>
> *Process* implies that KDD comprises many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations.
>
> *Nontrivial* refers to the fact that some search or inference is involved; that is, it is not a straightforward computation of predefined quantities like computing the average value of a set of numbers.

**Step 1: Learn the Application Domain and Establish Goals**

The first step of the KDD process is to learn the application domain and define the knowledge discovery goals. In this research, the application domain is the USAF's current implementation of the MDI metric as a derivative of the Navy's MDI methodology. Chapter two provides a comprehensive literature review for the application domain. The literature review covers the background of asset management principles, the MDI methodologies implemented in the Navy and USAF, and existing limitations with the USAF MDI implementation. The literature review serves as the

foundation for the overarching knowledge discovery process. Next, knowledge discovery goals are necessary to guide the overarching KDD process. There are two goals for this MDI knowledge discovery research:

1. Infer relationships between real property data and mission critical infrastructure.
2. Predict mission critical infrastructure using real property data.

The first goal is inference-based and seeks to use the data and supervised learning techniques to better understand relationships between real property data and mission critical infrastructure. Because military services are required to maintain accurate and current real property records, a plethora of real property data is available. Currently, relationships between real property data elements and mission critical infrastructure are not codified or well understood. Newfound knowledge of potential relationships between real property data and mission critical infrastructure could enable the use of rules or heuristics to facilitate identification of mission critical infrastructure and generally improve the existing body of knowledge.

The second goal is to predict mission critical infrastructure by training a supervised learning model using real property data. Mission critical infrastructure is defined as infrastructure with an MDI value greater than or equal to 85 (AFCEC, 2015). A predictive model that can reliably identify mission critical infrastructure has the potential to facilitate the MDI adjudication process and overall validation of the vast USAF real property inventory. Accomplishing this objective will benefit the civil engineer career field by significantly reducing person-hours currently required for manual MDI reviews.

## Step 2: Creating a Target Data Set

The second step of the KDD process is to create a target data set. This MDI knowledge discovery research employs two target data sets. The two data sets consist of Navy and Air Force real property from the respective real property databases of record. Real property data is comprised of facts pertaining to specific facilities and infrastructure. Because real property inventory reporting is required within the federal government, every branch of the DOD maintains a database with current real property data. The two target data sets consist of a matrix with facilities in rows and real property features in columns. The labels corresponding to each infrastructure asset are the MDI values obtained via the Navy MDI methodology for stakeholder input. The existing real property data features represent the potential to predict mission critical infrastructure and reveal relationships with mission critical infrastructure.

### Air Force Data Set

The USAF MDI beta test represents the only data collection effort undertaken to date with the express objective of identifying MDI values for specific infrastructure assets through subject matter expert insight. As such, this data constitutes the most thorough assessment of USAF facility mission criticality available. The MDI beta test included both Fairchild AFB and Langley AFB. Multiple attempts to obtain the Langley AFB data proved unsuccessful. Personnel at the USAF Civil Engineer School at Wright-Patterson AFB provided the Fairchild AFB MDI beta test data in a CSV file. The original Fairchild data set contains 571 observations and 17 data features from the MDI survey data collection effort. The data from the 2008 beta test does not include real property data. Table 7 displays the original MDI beta test data features.

Table 7. Original MDI Beta Test Data Features

| Feature Name | Data Type |
|---|---|
| MDI | Integer |
| HOST INSTALLATION CODE | Text |
| INSTALLATION | Text |
| TENANT | Logical |
| FUNCTIONAL AREA | Text |
| FACILITY NUMBER | Integer |
| FACILITY NAME | Text |
| MDI QUESTION 1 | Letter |
| MDI QUESTION 2 | Letter |
| MDb AVERAGE | Numeric |
| n | Integer |
| Surveyor | Text |
| Surveyor Group | Number |
| Group ID | Number |
| Interview Date | Date |

In order to evaluate the Fairchild infrastructure MDI values against the corresponding real property data, the beta test data is merged with the most current (FY15) real property data from ACES-RP. ACES support personnel located at Maxwell-Gunter Annex provided FY15 real property data for Fairchild AFB. The two CSV files provided from ACES-RP contain standardized real property data by facility and are titled "Fairchild RT_FACILITIES" and "Fairchild RT_REAL_PROPERTY_ASSETS", respectively. The original "Fairchild RT_Facilities" file contains 812 observations and 35 features. The original "Fairchild_RT_REAL_PROPERTY_ASSETS" file contains 935 observations and 34 features. The original ACES-RP data sets contain redundant features, unique identifiers, textual information, and many missing entries. Significant preprocessing is necessary to prepare the USAF data set for analysis. The original data features for the two Fairchild AFB real property data sets are listed in Table 8 and Table 9, respectively.

Table 8. “Fairchild RT_FACILITIES” Original Data Features

| Feature Name | Data Type |
|---|---|
| RPUID | Integer |
| ACES_INSTALLATION_CD | Text |
| ACES_FACILITY_NBR | Integer |
| FACILITY_NBR | Integer |
| CIP_PHASE_YN | Logical |
| CONSTRUCT_MATERIAL_CD | Text |
| CONSTRUCT_TYPE_CD | Text |
| ADA_COMPLIANCE_CD | Text |
| BOOK_VALUE | Integer |
| BUILT_DT | Date |
| CURRENT_PERIOD_DEP_AMT | Integer |
| EST_USE_LIFE_ADJ_QTY | Integer |
| EST_USE_LIFE_QTY | Integer |
| HEIGHT_QTY | Integer |
| HEIGHT_UOM | Text |
| HOUSING_ATTRIBUTE_CD | Text |
| LENGTH_QTY | Integer |
| LENGTH_UOM | Text |
| MODULE_QTY | Integer |
| PLANT_REPLACEMENT_VALUE | Integer |
| REPLACEMENT_DEPT_REG_CD | Integer |
| REPLACEMENT_FUND_CD | Integer |
| REPLACEMENT_SUB_ACCT_CD | Integer |
| REPLACEMENT_ORG_CD | Integer, Text |
| RESTORE_MOD_DEPT_REG_CD | Integer |
| RESTORE_MOD_FUND_CD | Integer |
| RESTORE_MOD_SUB_ACCT_CD | Integer |
| RESTORE_MOD_ORG_CD | Integer, Text |
| TOT_ACCUM_DEP_AMT | Integer |
| TOT_CAPITAL_IMP_COST | Integer |
| WIDTH_QTY | Integer |
| WIDTH_UOM | Text |
| FLOOR_ABOVE_GROUND_QTY | Integer |
| FLOOR_BELOW_GROUND_QTY | Integer |
| PHYSICAL_QUALITY_RATE | Integer |

Table 9. "Fairchild RT_REAL_PROPERTY_ASSETS" Original Data Features

| Feature Name | Data Type |
|---|---|
| RPUID | Integer |
| ACES_INSTALLATION_CD | Text |
| ACES_FACILITY_NBR | Integer |
| SITE_UID | Integer |
| ANNUAL_OPERATING_COST | Integer |
| COMMAND_CLAIMANT_CD | Integer, Text |
| CONSTRUCT_AGENT_CD | Text |
| CURRENT_USE_FUNC_CAP_CD | Text, Integer |
| DEPTH_QTY | Integer |
| DEPTH_UOM | Text |
| DESCRIPTION | Text |
| FINANCIAL_REPORTING_ORG_CD | Integer, Text |
| HISTORIC_STATUS_CD | Text |
| HISTORIC_STATUS_DT | Date |
| INTEREST_TYPE_CD | Text |
| MISSION_DEPENDENCY_CD | Text |
| NEIGHBORHOOD_NAME | Text |
| OPERATIONAL_STATUS_CD | Text |
| PRED_CURRENT_USE_CAT_CD | Integer |
| PRED_CURRENT_USE_FAC_CD | Integer |
| PRED_DESIGN_USE_CAT_CD | Integer |
| PRED_DESIGN_USE_FAC_CD | Integer |
| PREPONDERANT_USING_ORG_CD | Integer, Text |
| RPA_NAME | Text |
| RPA_TYPE_CD | Text |
| SALVAGE_VALUE_AMT | Integer |
| SALVAGE_VALUE_REASON_CD | Text |
| SERVICE_DT | Date |
| TOTAL_UOM | Text |
| TOTAL_UOM_QTY | Integer |
| UTILIZATION_RATE | Integer |
| RPA_SUSTAINABILITY_CD | Integer |
| COST_SHARING_PARTNERS | Text |
| TARGET_ASSET_OWNER_ORG_CD | Integer, Text |

**Navy Data Set**

Navy facilities are evaluated via the Navy's MDI methodology, which means that the current Navy real property records reflect MDI values derived from stakeholder input. Navy real property personnel at the Pentagon provided a CSV file with the entire real property inventory for the Navy. While the data set consisted of a high number of

observations (infrastructure assets), the real property data features were limited to a subset of features. The original Navy data set contained 119,275 observations and 10 real property data features (including MDI). The Navy data set contains far fewer features than the USAF real property data set. Despite containing fewer features, preprocessing of the Navy data set is necessary in step three of the KDD process to prepare the data for analysis. The original Navy real property features are listed in Table 10.

Table 10. Original Navy Data Features

| Feature Name | Data Type |
|---|---|
| REGION | Letters |
| UIC | Numbers and Letters |
| INSTALLATION_NAME | Text |
| FAC | Integers |
| CATEGORY_CODE | Integers |
| UM | Letters |
| MEASUREMENT | Integers |
| PRV | Integers |
| MDI | Integers |
| FACILITY NAME | Text |

**Step 3: Data Cleaning and Preprocessing**

The third step of the KDD process is data cleaning and preprocessing. The importance of data cleaning and preprocessing cannot be overstated as this step enhances the reliability of the data (Maimon & Rokach, 2005). In research literature, the data mining step of the KDD process garners much of the attention, however, Fayyad et al. (1996a) purport that the data preprocessing step is equally important. For useful results, data mining requires clean and accurate data (Maletic & Marcus, 2010). The data cleansing process includes defining and determining error types, searching and identifying error instances, and correcting any uncovered errors (Maimon & Rokach,

2005). Common concerns in data cleaning and preprocessing include dealing with outliers, noisy and missing data, and data types (U. Fayyad et al., 1996b). Additionally, data cleaning and preprocessing is a necessary precursor to data mining because errors are common in large data sets (Maletic & Marcus, 2010). Fayyad et al. (2003) convey that 40 percent of all collected data contain errors. As such, extracting and manipulating data is often where the majority of time is spent in the KDD process (U. M. Fayyad et al., 2003). The data cleaning and preparation process is iterative and typically occurs throughout the KDD process as new findings are identified and different techniques are applied. A commonly shared statistic is that approximately 80 percent of data analysis is spent on cleaning and preparing the data (Wickham, 2014).

**USAF Data Set**

Because the USAF target data set is composed of three distinct data sets, significant preprocessing is required to merge and investigate potential issues. The target data set requires distinct infrastructure assets as observations with corresponding real property data features. Correctly merging the features from the three data sets for specific facilities is a major concern for the USAF target data set. The two ACES-RP data sets contain real property unique identifiers (RPUIDs), which offer a distinct link for merging the two data frames. The MDI beta test data, however, does not contain RPUIDs and the data set must be merged with the ACES-RP data by facility number. Some data features were removed altogether due to missing data and some features were altered to better suit the data mining application. One example is the “age” feature, which is derived from the service date by calculating the difference in the service year and the current year. As another example, the facility class and category group features

are not standalone features in the RP data and must be derived from four-digit FAC codes. The facility class is the first digit of the FAC code and the category group is the first two digits of the FAC code. As such, the facility class and category group features were created by parsing the first digit and the first two digits from the FAC code feature, respectively. As a third example, since the total unit of measure quantity is reported against different units of measure for different infrastructure asset types, the numeric values for the measurement quantity are standardized to have a mean of zero and a standard deviation of one. This processing step rectifies the issue of inconsistent units. After the data cleaning and preprocessing phase, the USAF target data set contains 304 observations and 45 features. Table 11 displays the USAF data set features and class types after preprocessing; "MC" is the mission critical response variable where "MC" indicates an MDI of 85 or higher and "nonMC" indicates MDI below 85.

Table 11. USAF Data Set Features

| Feature Name | Class Type |
|---|---|
| MDI | Integer |
| MC | Factor, 2-levels |
| CONSTRUCT_MATERIAL_CD | Factor, 14-levels |
| CONSTRUCT_TYPE_CD | Factor, 3-levels |
| ADA_COMPLIANCE_CD | Factor, 2-levels |
| BOOK_VALUE | Numeric |
| CURRENT_PERIOD_DEP_AMT | Numeric |
| EST_USE_LIFE_QTY | Factor, 2-levels |
| HEIGHT_QTY | Numeric |
| LENGTH_QTY | Numeric |
| WIDTH_QTY | Numeric |
| DEPTH_QTY | Integer |
| PLANT_REPLACEMENT_VALUE | Numeric |
| REPLACEMENT_DEPT_REG_CD | Factor, 3-levels |
| REPLACEMENT_FUND_CD | Factor, 6-levels |
| REPLACEMENT_SUB_ACCT_CD | Factor, 2-levels |
| RESTORE_MOD_DEPT_REG_CD | Factor, 3-levels |
| RESTORE_MOD_FUND_CD | Factor, 8-levels |
| RESTORE_MOD_SUB_ACCT_CD | Factor, 2-levels |
| RESTORE_MOD_ORG_CD | Factor, 9-levels |
| TOT_ACCUM_DEP_AMT | Numeric |
| FLOOR_ABOVE_GROUND_QTY | Factor, 6-levels |
| FLOOR_BELOW_GROUND_QTY | Factor, 2-levels |
| PHYSICAL_QUALITY_RATE | Integer |
| BOOK_VALUE_ZERO | Factor, 2-levels |
| Age | Numeric |
| age.over45 | Factor, 2-levels |
| ANNUAL_OPERATING_COST | Numeric |
| COMMAND_CLAIMANT_CD | Factor, 2-levels |
| CONSTRUCT_AGENT_CD | Factor, 2-levels |
| FINANCIAL_REPORTING_ORG_CD | Factor, 4-levels |
| HISTORIC_STATUS_CD | Factor, 4-levels |
| OPERATIONAL_STATUS_CD | Factor, 4-levels |
| PREPONDERANT_USING_ORG_CD | Factor, 10-levels |
| RPA_TYPE_CD | Factor, 3-levels |
| TOTAL_UOM | Factor, 10-levels |
| TOTAL_UOM_QTY | Numeric |
| UTILIZATION_RATE | Integer |
| RPA_SUSTAINABILITY_CD | Factor, 2-levels |
| ANNUAL_OPERATING_COST_ZERO | Factor, 2-levels |
| CONSTRUCT_AGENT_CD_USACE | Factor, 2-levels |
| facilityClass | Factor, 7-levels |
| categoryGroup | Factor, 25-levels |
| utilization | Factor, 3-levels |
| costShare | Factor, 2-levels |

The number of observations in the USAF data set is low for a machine learning application. This presents a potential limitation as there may not be enough examples from which the algorithm can "learn" and generalize to unseen data. Additionally, there are many features in the USAF RP data. As such, deliberate actions are necessary to minimize the number of features through feature selection techniques.

**Navy Data Set**

The cleaning and preprocessing for the Navy data set is far less complex than the USAF data due to the single data frame and far fewer features. The primary data preprocessing actions included parsing the facility class and category group features from the FAC codes and standardizing the measurement quantity by the corresponding unit of measure. After the data cleaning and preprocessing phase, the Navy target data set contains 81,224 observations in rows and six features in columns. The number of observations is promising for model training; however the low number of real property features presents a possible limitation for both prediction and inference if the features are not significantly associated with the response. Table 12 displays the data features and class types; "MC" is the response variable indicating whether or not the asset is mission critical.

Table 12. Navy Data Set Features

| Feature Name | Class Type |
|---|---|
| MC | Factor, 2-levels |
| UM | Factor, 19-levels |
| MEASUREMENT | Numeric |
| PRV | Integer |
| facilityClass | Factor, 8-levels |
| categoryGroup | Factor, 39-levels |

### Step 4: Data Reduction and Projection

The data reduction and projection step is focused on determining useful features in the dataset and using dimensionality reduction or transformation methods to minimize the number of data features (U. Fayyad et al., 1996b). Transforming the data may also be required in this step depending on the data, task, and methods employed. The data reduction and projection step is an iterative process and varies based on the application (Maimon & Rokach, 2005).

#### USAF Data Set

Given the high number of features and low number of observations in the USAF data set, dimension reduction is an important consideration. This research pursues the use of three filter feature selection methods in order to minimize the number of features employed in modeling. The filter methods include RELIEF-F, Correlation Based Feature Selection (CFS), and information gain. Each of the filter method functions are employed using the "FSelector" package in the R programming language (Romanski & Kotthoff, 2014). Also, a wrapper method, known as recursive feature selection (RFE), is employed with the random forests algorithm to identify a subset of optimal features. RFE is employed through the caret data mining package also in R (Kuhn, 2012). Feature selection techniques are employed with the training data.

The RELIEF-F feature selection method produces weights corresponding to each feature in the data set. The features are stratified by their weightings where higher values are indicative of better features. This feature selection method does not automatically identify the number of features to use in modeling so a cutoff must be selected. Two

subsets are selected using the RELIEF-F feature selection method. The first RELIEF-F subset, referred to as subset one, includes the five highest weighted features: (1) Category Group, (2) Facility Class, (3) Preponderant Using Organization Code, (4) Replacement Organization Code, and (5) Utilization Rate. The second RELIEF-F subset, referred to as subset two, is based on the significant difference between feature weights and includes the top three features: (1) category group, (2) facility class, and (3) preponderant using organization code. Based on the RELIEF-F filter method, the top three features are categorical features that correspond to the facility function and the organization associated with the facility, respectively.

The CFS feature selection method uses entropy and correlation measures to select an optimum feature subset. The CFS filter yielded a subset of seven categorical and numeric features, referred to as subset three, and includes the following attributes: (1) book value, (2) height quantity, (3) length quantity, (4) total accumulated depreciation amount, (5) preponderant using organization code, (6) total unit of measure quantity, and (7) cost sharing.

The information gain filter method produces weights for all features where the highest weight is the most important feature. Again, a cutoff of five features is employed to select the optimum features. The information gain feature subset, referred to as subset four, includes (1) category group, (2) height quantity, (3) replacement organization code, (4) preponderant using organization code, and (5) length quantity. The common themes among the filter methods include the facility function, the using organization, and physical attributes.

The RFE method with random forest provides a fairly large subset with 15 features. The subset includes both categorical and numeric features. The RFE feature subset, referred to as subset five, includes the following attributes: (1) category group, (2) length, (3) height, (4) plant replacement value, (5) total accumulated depreciation amount, (6) book value, (7) total unit of measure quantity, (8) replacement organization code, (9) width, (10), current period depreciation amount, (11) cost sharing, (12) total unit of measure, (13) facility class, (14) floor above ground quantity, and (15) construction material code.

Five themes emerged from the four feature selection methods employed. The themes include infrastructure function, physical attributes, financial characteristics, organizational characteristics, and infrastructure utilization rate. Figure 13 displays the feature selection results and corresponding themes across the four feature selection methods.

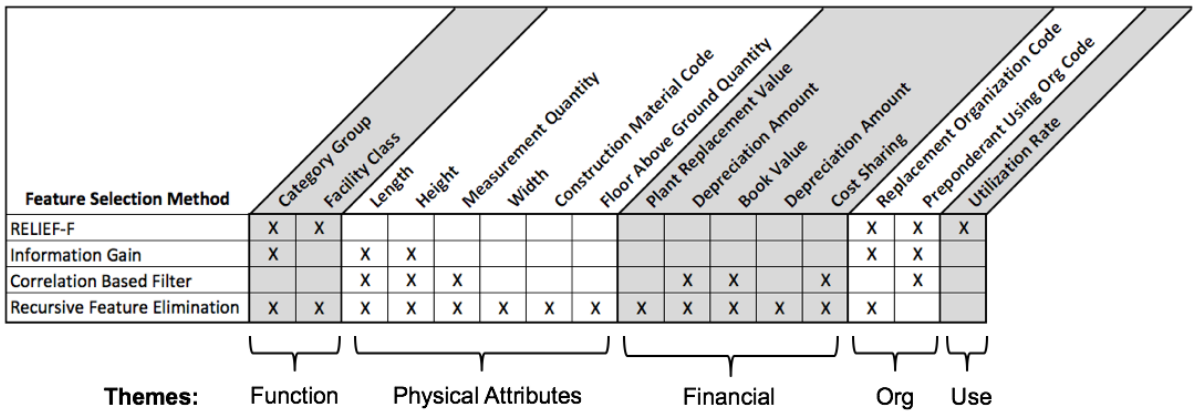



Figure 13. USAF Data Set Feature Selection Results

### Navy Data Set

The target Navy data set mirrors the requirements for the USAF data set where infrastructure assets are in rows and real property features in columns. Because the Navy data set contains far fewer features than the USAF data set, data reduction and projection is not a significant of a concern. The four primary features suitable for analysis include measurement, plant replacement value, facility group (first digit of FAC code), and category group (first two digits of FAC code). The facility group and category group are correlated as the first digit represents the same facility "group" for both features. The category group has a relatively high number of factor levels at 39, which increases the computational complexity. Using further break outs of the function codes (e.g. four-digit FAC codes) is computationally prohibitive due to the high number of factor levels. Given the small number of features with the Navy data set, no feature selection algorithms are employed.

## Step 5: Choosing the Data Mining Task

There are two distinct data mining tasks in this research. The first data mining objective is to identify and describe relationships in the real property data with respect to mission critical infrastructure. The second data mining objective is to predict mission critical infrastructure using real property data. These objectives are pursued as a supervised learning classification task. AFCEC defines "mission critical" as an MDI greater than or equal to 85 (AFCEC, 2015). This definition facilitates discretization of MDI values to two levels, "mission critical" (85 or higher) or "non-mission critical" (84

or lower). With these labels, the problem qualifies as a binary classification task where the positive class is “mission critical” and the negative class is “non-mission critical”.

Classifiers are unique in that high accuracy does not necessarily guarantee that the intended objectives are met. As such, accuracy is not the optimum measure of success. For prediction of mission critical infrastructure, emphasis is placed on the sensitivity associated with classifying the positive class, “mission critical”. Therefore, the objective is to train a classifier with minimum sensitivity and specificity values of 0.8 on test data. More concretely, a successful classifier should have a minimum 80-percent true positive and true negative rate. A classifier meeting these specifications has the potential to serve as a decision support tool to facilitate identification of MDI discrepancies, thereby minimizing the number of facilities for adjudication review.

A secondary objective with the classification task is inference. Inference pursues a deeper understanding of the relationships between the predictors and the response variable. Inference provides valuable insight into the “why” associated with the model and underlying phenomenon. Alternatively, a “black box” classifier with high predictive accuracy offers limited knowledge discovery value. Understanding how a model is employing predictors to classify mission critical infrastructure is central to true knowledge discovery. By identifying important features and the relationships between the features and the mission critical classification, civil engineers can better understand the intuitions associated with critical infrastructure.

### Evaluating Classifiers

Classification algorithms are often evaluated by the accuracy on the test set, however, additional metrics are useful in evaluating classifiers. Confusion matrices are the standard tool utilized in evaluating classifier effectiveness. In binary classification, a confusion matrix identifies the classifications based on positive and negative class assignments. In the case of MDI, mission critical ("MC") is the positive class and non-mission critical ("nonMC") is the negative class. The four possible classification outcomes are (1) true positive (TP), (2) true negative (TN), (3) false positive (FP), and (4) false negative (FN). True positives occur when the classifier correctly classifies an observation as positive and true negatives occur when a classifier correctly classifies an observation as negative. Alternatively, false positives occur when the classifier incorrectly classifies an observation as positive and false negatives occur when the classifier incorrectly classifies an observation as negative. Table 13 presents the four classifier outcomes in a notional confusion matrix.

Table 13. Notional Confusion Matrix

| | | **Predicted Class** | | |
|---|---|---|---|---|
| | | **Negative** | **Positive** | **Total** |
| **True Class** | **Negative** | True Negative (TN) | False Positive (FP) | N |
| | **Positive** | False Negative (FN) | True Positive (TP) | P |
| | **Total** | N | P | |

Classifier diagnostic metrics are derived from the four outcomes presented in the confusion matrix in Table 13. For example, accuracy is calculated by adding the total number of true positives and true negatives and dividing by the total number of observations. Next, specificity and sensitivity present class-specific performance (James et al., 2013). Sensitivity is known as the true positive rate, which equates to the true positives divided by the sum of the true positives and false negatives. Specificity is known as the true negative rate, which is the number of true negatives divided by the sum of the true negatives and false positives. Furthermore, precision is the positive predictive value, which is calculated by dividing the number of true positives by the sum of the true positives and false positives. Alternatively, the negative predictive value is the number of true negatives divided by the sum of the true negatives and false negatives. In classification tasks, the accuracy may not be the best method of determining the costs and benefits associated with a given classifier due to the lack of information about the false positive and false negative predictions.

**Receiver Operating Characteristics Curve**

The receiver operating characteristics (ROC) curve is an effective means of presenting the overall performance of a classifier (James et al., 2013). A ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis. A notional ROC curve is presented in Figure 14.

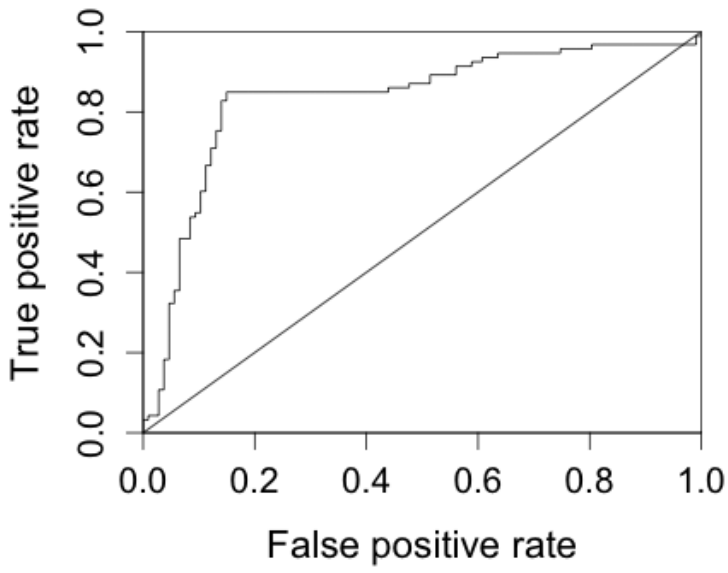



Figure 14. Notional ROC Curve

The metric associated with ROC curves is known as the area under the curve (AUC). "ROC curves are useful for comparing different classifiers, since they take into account all possible thresholds" (James et al., 2013). Furthermore, Huang and Ling (2005) conclude that AUC is superior to accuracy for classifier comparison. A good classifier will have a curve close to the top left corner, which is indicative of a high true positive rate and low false positive rate. The diagonal line across the plot indicates an AUC value of 0.5, which is considered a minimally effective classifier comparable to a coin toss. An AUC value of 1.0 is indicative of a perfect classifier.

## Chapter Summary

This chapter presented steps one through five of the KDD process with specific actions taken as precursors to the data mining analysis. The USAF and Navy data sets are preprocessed to include infrastructure assets in rows and real property features in columns. The binary response variable indicates that an observation is "mission critical" or "non-mission critical" based on the MDI mission critical threshold of 85. With the target data sets prepared in tidy data frames, the remaining steps of the KDD process are executed; step six encompasses algorithm selection, step seven is implementation, step eight is interpretation and evaluation, and step nine is using the knowledge. The next chapter presents the analysis and results of the data mining-specific steps pursuant to the two KDD goals, inference and prediction.

# IV. Analysis and Results

## Chapter Overview

This chapter presents the analysis and results of the data mining-specific steps in the KDD process. The goals of the KDD process are to (1) infer relationships between real property data and mission critical infrastructure and (2) to classify mission critical infrastructure using real property data. Specific results are addressed for both the Air Force and Navy data sets within the context of the two KDD goals.

Step six of the KDD process entails data mining algorithm selection, step seven is the data mining implementation, step eight is interpretation and evaluation, and step nine is using the knowledge. Steps six through eight of the KDD process are combined due to the integrated nature of the procedures. The data mining analysis first employs numerous classification paradigms with differing strengths and weaknesses. Specifically, tradeoffs between the different classifiers include flexibility, interpretability, and computational complexity. The classification models investigated include logistic regression, linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbors, generalized additive models, and multiple classification tree algorithms. The classification tree models span bagging, boosting, random forests, and C5.0. General descriptions of the learning algorithms are presented in Appendix A. Resampling techniques are employed in classifier training for test error estimation and model selection. The classifier performance is compared using the area under the curve (AUC) for the respective receiver operating characteristics (ROC) curves. The best algorithm(s) are then selected for the respective goals. The chapter culminates with step nine, using the knowledge, by

applying the best classification model as a decision support tool for AFCENT infrastructure MDI adjudication.

### Steps 6 – 8: Algorithm Selection, Data Mining, Interpretation and Evaluation

This section combines steps six, seven, and eight of the KDD process, which encompass the data mining implementation. Step six is algorithm selection and builds on step five, choosing the data mining task, by further investigating specific classification models. Step seven entails the actual data mining implementation including training and tuning the respective models. Step eight includes interpretation of the data mining results within the context of the KDD objectives. The analysis and results are presented separately for the USAF and Navy data sets.

First, the most suitable data mining algorithms must be identified. Data mining algorithm selection is contingent upon the goals established early in the KDD process. The first KDD objective favors model interpretability while the second necessitates high classification performance. These objectives can represent competing priorities in data mining algorithm selection. For classification performance, higher flexibility tends to yield higher accuracy at the cost of interpretability. Inference, however, is generally best implemented with relatively inflexible models that allow for increased interpretability. Given these objectives, this research pursues an algorithm that can fulfill both of the objectives. This is heavily dependent upon the data, however, and the algorithm with the best prediction performance may not be the most interpretable model. Therefore, numerous algorithms are pursued and the costs and benefits are compared in the context of the KDD goals.

Data mining algorithm selection is typically a function of the available data and the specified data mining task. In this research, the data sets contain both numeric and categorical predictors and the first data mining task is prediction via classification. There are many model types available for classification and tradeoffs abound. Given the computational power available with standard computers, numerous data mining algorithms are employed in order to compare classifier performance. All classifiers are employed with cross-validation and a range of tuning parameters in order to identify the best tuned models. Furthermore, the area under the ROC curve is employed in comparing the models for data mining algorithm selection.

### USAF Data Set

For the Fairchild data set, the feature selection process in step four yielded five feature subsets for investigation. In order to select the most appropriate classification algorithm, each of the five feature subsets is evaluated against the potential algorithms. Additionally, a sixth iteration is included with all features from the Fairchild data set. The Fairchild data set is partitioned into a training and test set. The training set is comprised of two-thirds of the data and the test set is comprised of one-third of the data. K-fold cross validation resampling is employed with each of the six feature subsets. Specifically, the training data is partitioned into five-folds for cross validation, repeated 20 times, to estimate error on the test data set. Five-folds are selected over ten-folds due to the limited number of observations in the training set. For an initial comparison of classifier performance, the 95 percent confidence intervals for the area under the ROC curve (AUC) are plotted together. The classifier comparison plots for each of the feature subsets are presented in Figure 15 through Figure 20.

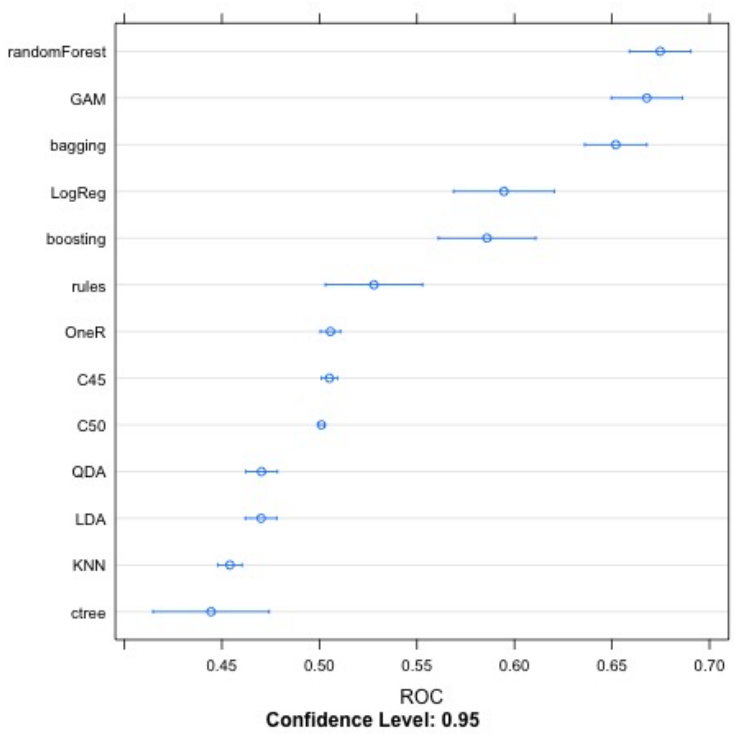



Figure 15. Fairchild Classifier Comparison: Subset 1 ROC AUC Values

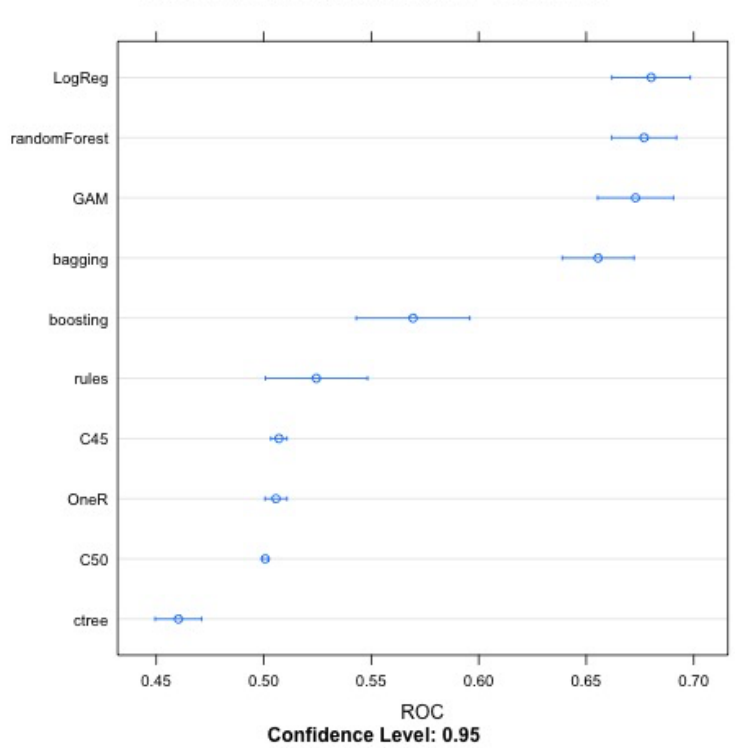



Figure 16. Fairchild Classifier Comparison: Subset 2 ROC AUC Values

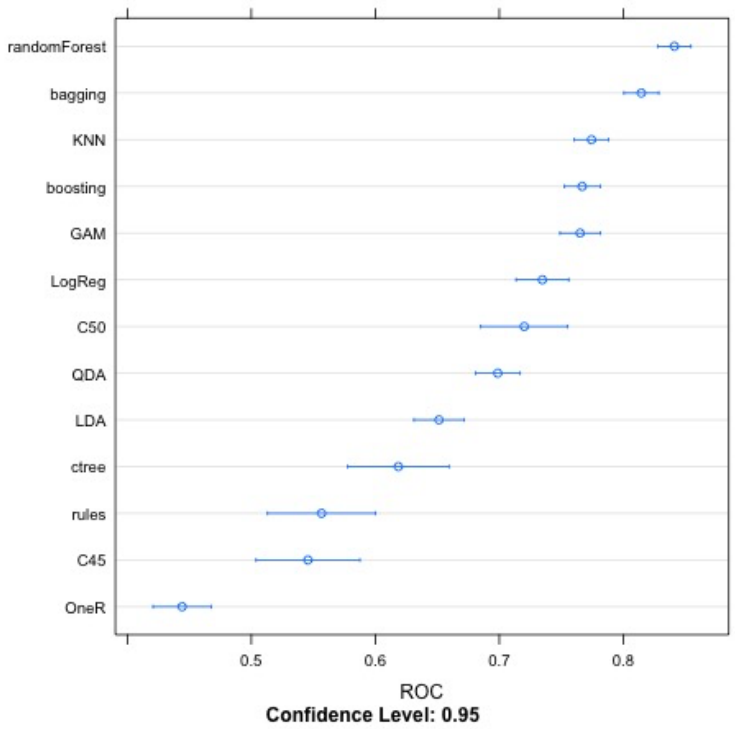



Figure 17. Fairchild Classifier Comparison: Subset 3 ROC AUC Values

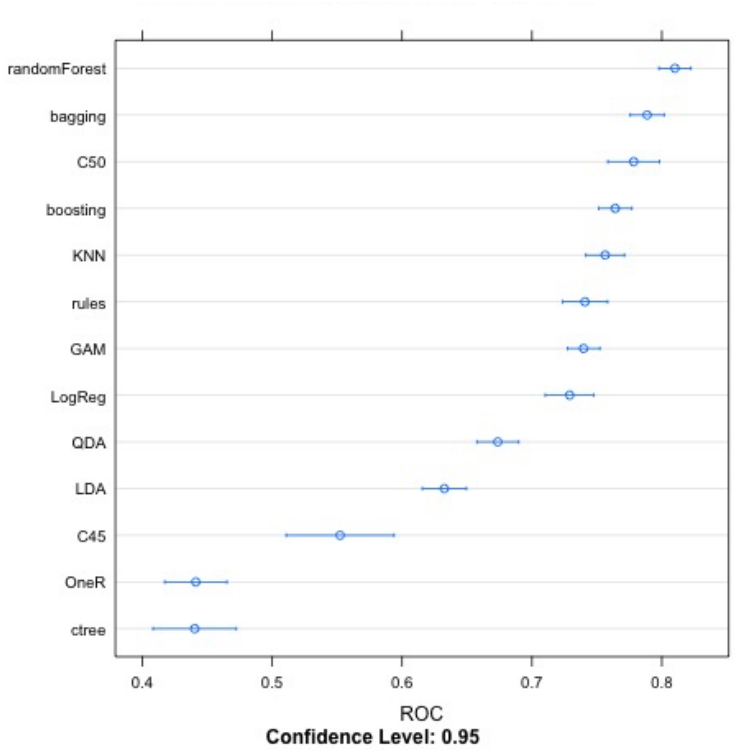



Figure 18. Fairchild Classifier Comparison: Subset 4 ROC AUC Values

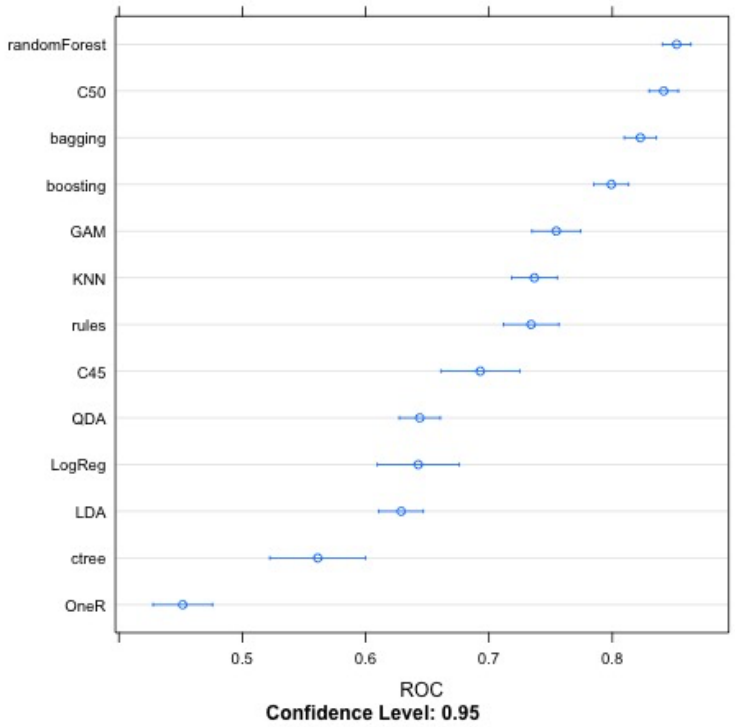



Figure 19. Fairchild Classifier Comparison: Subset 5 ROC AUC Values

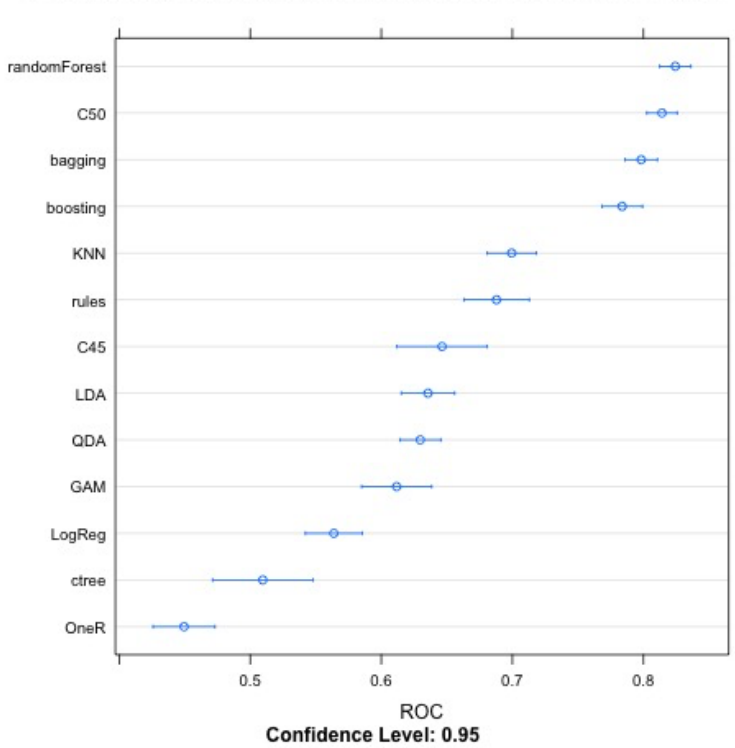



Figure 20. Fairchild Classifier Comparison: All Features ROC AUC Values

Figure 15 through Figure 20 shows that decision tree classifiers outperform the other classifiers across all combinations of feature subsets. The ROC AUC values suggest that the random forests algorithm is the best classification algorithm evaluated in this study. Specifically, the random forest classifiers yield ROC AUC values of 0.8 or higher on eight out of the twelve comparisons. ROC AUC values between 0.8 and 0.9 are generally considered good classifiers and values between 0.9 and 1.0 are considered excellent. The random forests algorithm employed with subsets three and five produce the highest ROC AUC values. Alternatively, the parametric models never attain a ROC AUC value over 0.8. Logistic regression paired with the features in subset three yields the best parametric model performance with a ROC AUC value between 0.7 and 0.8. As such, logistic regression appears to provide the best platform for the inference objective.

The ROC AUC values across the feature subsets also make evident the tradeoffs associated with the different feature combinations. ROC AUC values are generally lower when fewer features are employed in model training. Specifically, subsets one and two achieve ROC AUC values less than 0.70 across all classifiers. ROC AUC values between 0.60 and 0.70 are generally considered to have poor classifier performance. Furthermore, all classifiers utilizing fewer than five features produce ROC AUC values below 0.80. This suggests that modeling of the underlying phenomenon improves when more real property features are available and that the best features in the subset lack the information necessary to accurately classify mission critical infrastructure.

## Inference

The USAF MDI beta test data from Fairchild AFB presents an opportunity to learn about the relationships between real property data and mission critical infrastructure as identified by USAF stakeholders. The small sample size does minimize the benefits associated with supervised learning; however, the inference pursuit may yield insight previously unknown. Two specific inference questions for the MDI problem are as follows:

1. What real property features contribute to classifying USAF mission critical infrastructure?
2. What are the relationships between USAF real property features and mission critical infrastructure?

The inference pursuit requires deliberate consideration in model selection. Parametric models, such as logistic regression, linear discriminant analysis and quadratic discriminant analysis, generally provide superior inference capability over more flexible models. Specifically, parametric models afford direct insight into the contributions of specific features through the respective coefficient values. In this study, logistic regression performed the best out of the parametric models as identified in the initial classifier comparison. As such, logistic regression is selected for the inference objective.

Given that there are many features in the USAF data set, specific techniques are available to further increase the inference capability of a logistic regression model. For the inference objective, the "glmnet" package in R is employed to take advantage of the Least Absolute Shrinkage and Selection Operator ("lasso") and regularization tuning parameters, alpha and lambda, in order to zero in on the important features (Hastie & Qian, 2014). These methods utilize penalized maximum likelihood for linear models,

which effectively minimize the coefficient values for less significant features. The ridge regression model is fit when the alpha tuning parameter is set to zero and the lasso model is fit when the alpha parameter is set to one (James et al., 2013). In both cases, the lambda value serves as the tuning parameter. The ridge regression method minimizes coefficient values for less important predictors while the lasso method shrinks less important predictor coefficients to zero. For prediction accuracy, the choice between lasso and ridge regression comes down to the bias variance tradeoff and the characteristics of the data. Ridge regression tends to have lower variance than the lasso method whereas the lasso method tends to experience higher bias (James et al., 2013). For inference, the lasso method increases the interpretability of a logistic regression model as the predictors that are not associated with the response shrink to exactly zero, which leaves a subset of important features. The lasso method is employed with the Fairchild data to identify important real property features associated with mission critical infrastructure.

The lasso model is employed using all 46 real property features with the categorical features decomposed into dummy variables for their respective factor levels. Five-fold cross validation is repeated 20 times in order to identify the best lambda value for the area under the ROC curve and to estimate test performance. The logistic regression model ultimately provides the probability that a specific observation is "mission critical". The continuous features are all preprocessed to have a mean of zero and standard deviation of one. The final lasso model retains 38 significant predictors. Each coefficient value equates to the respective change in the log odds for the respective feature, assuming all other feature values remain constant. Generally, coefficient

estimates with a value of zero is indicative of no association, positive values indicate an increased likelihood of mission critical, and negative values indicate a decreased likelihood of mission critical. Furthermore, taking the exponent of the coefficient value yields the odds ratio value for the feature, which equates to the odds change for a one-unit change in the feature value. The lasso model results indicate three general categories of real property data that are significant in distinguishing between mission critical and non-mission critical infrastructure. The three categories include (1) infrastructure characteristics, (2) financial characteristics, and (3) organization codes. These feature categories serve to shape the intuitions surrounding mission critical infrastructure at Fairchild AFB. Table 14 presents the significant predictors with their respective coefficients and odds ratios.

Table 14. Lasso Model Results for USAF Data Set

| Feature Name | Coefficient | Odds Ratio |
|---|---|---|
| (Intercept) | -1.1526 | |
| REPLACEMENT_ORG_CD1L | 0.5701 | 1.77 |
| costShareTRUE | 0.4152 | 1.51 |
| TOTAL_UOMLF | 0.3536 | 1.42 |
| TOT_ACCUM_DEP_AMT | 0.1692 | 1.18 |
| REPLACEMENT_FUND_CD3830 | 0.1649 | 1.18 |
| PLANT_REPLACEMENT_VALUE | 0.1395 | 1.15 |
| FLOOR_BELOW_GROUND_QTY1 | 0.1293 | 1.14 |
| FLOOR_ABOVE_GROUND_QTY5 | 0.1232 | 1.13 |
| DEPTH_UOMIN | 0.0994 | 1.10 |
| BOOK_VALUE | 0.0954 | 1.10 |
| CONSTRUCT_MATERIAL_CDOTHR | 0.0948 | 1.10 |
| HISTORIC_STATUS_CDNREI | 0.0906 | 1.09 |
| CURRENT_PERIOD_DEP_AMT | 0.0849 | 1.09 |
| categoryGroup82 | 0.0495 | 1.05 |
| categoryGroup13 | 0.0449 | 1.05 |
| FLOOR_ABOVE_GROUND_QTY7 | 0.0444 | 1.05 |
| REPLACEMENT_ORG_CD54 | 0.0157 | 1.02 |
| utilizationpartial | 0.0108 | 1.01 |
| ANNUAL_OPERATING_COST_ZERO1 | 0.0067 | 1.01 |
| REPLACEMENT_ORG_CD1Y | 0.0047 | 1.00 |
| categoryGroup89 | 0.0039 | 1.00 |
| RESTORE_MOD_ORG_CD54 | 0.0016 | 1.00 |
| PREPONDERANT_USING_ORG_CD1Y | 0.0009 | 1.00 |
| FINANCIAL_REPORTING_ORG_CD54 | 0.0001 | 1.00 |
| PREPONDERANT_USING_ORG_CD54 | 0.0000 | 1.00 |
| RESTORE_MOD_FUND_CD3840 | 0.0000 | 1.00 |
| categoryGroup21 | 0.0000 | 1.00 |
| categoryGroup42 | -0.0008 | 1.00 |
| categoryGroup74 | -0.0039 | 1.00 |
| categoryGroup75 | -0.0042 | 1.00 |
| RESTORE_MOD_SUB_ACCT_CD5 | -0.0050 | 0.99 |
| TOT_ACCUM_DEP_AMT_ZERO1 | -0.0309 | 0.97 |
| CONSTRUCT_MATERIAL_CDBLCK | -0.0399 | 0.96 |
| REPLACEMENT_ORG_CD0J | -0.0628 | 0.94 |
| FLOOR_ABOVE_GROUND_QTY1 | -0.0686 | 0.93 |
| categoryGroup85 | -0.0865 | 0.92 |
| categoryGroup83 | -0.1011 | 0.90 |
| facilityClass2 | -0.1020 | 0.90 |

Using the results from the Fairchild AFB MDI beta test and assuming that all other covariates are equal, the following conclusions can be inferred about the predictors:

- Organization Code “1L” is 1.77 times more likely than other organization codes to be mission critical.
- When costs are shared among stakeholders, infrastructure is 1.51 times more likely to be mission critical than when costs are not shared.
- Infrastructure measured in LF is 1.42 times more likely to be mission critical.
- Infrastructure with funding code 3830 is 1.18 times more likely to be mission critical
- The likelihood of mission critical increases as the PRV value increases.
- Facilities with a single floor below ground are 1.14 times more likely to be mission critical than those that do not have a floor below ground.
- Facilities with five floors above ground are 1.13 times more likely to be mission critical than those that do not have five floors.
- The likelihood of mission critical increases as book value increases.
- Infrastructure with construction material code “other” is 1.10 times more likely to be mission critical than other material codes.
- Infrastructure with historic status code “NREI” is 1.09 times more likely to be mission critical than non “NREI” infrastructure.
- Category group 82 (heat and refrigeration), is 1.05 times more likely to be mission critical than those that are not category group 82.
- Category group 13 (Comm, Navigation Aids, Airfield Light) is 1.05 times more likely to be mission critical than infrastructure that is not in category group 13.
- Facilities with seven floors above ground are 1.05 times more likely to be mission critical than those that do not have seven floors above ground.
- Infrastructure with construction material code “BLCK” (concrete block) is 0.96 times as likely as non-concrete block infrastructure to be mission critical.
- Organization code “0J” is 0.94 times as likely to be mission critical compared to non-“0J” organization code infrastructure.
- Facilities with one floor above ground are 0.93 times as likely to be mission critical compared to infrastructure that is not a single floor above ground.
- Infrastructure in category group 85 (roads and other pavements) is 0.92 times as likely to be mission critical as non-roads and pavements.
- Infrastructure in category group 83 (sewage and waste) is 0.90 times as likely to be mission critical as non-sewage and waste infrastructure.
- Infrastructure in facility class 2 ( maintenance and production) is 0.90 times as likely to be mission critical as compared to non-facility class 2 infrastructure.

### Prediction

There are two noteworthy concerns with using the USAF MDI beta test data for a prediction model. First, the data set is very small at just over 300 observations. This limited sample size inhibits the viability of training a prediction model. Second, the target data set comes from just one USAF installation. As such, the data set cannot be considered representative of the entire USAF infrastructure population and so, the generalizability of a trained classifier is limited. An argument could be made for employing this data in a prediction model for similar installations with similar mission sets. Because of these limitations, the USAF data set is restricted to the inference objective. The Navy data set, however, contains ample observations for both the inference and prediction goals.

### Navy Data Set

Due to the limited number of features available for the Navy data set, no feature selection is employed and all features are employed in model training. Again, all potential classifiers are evaluated by comparing the respective ROC values. Eighty percent of the Navy data is utilized for training and 20 percent of the data is set aside for testing. Ten-fold cross validation resampling is employed with the training data. Figure 21 displays the 95 percent confidence intervals for each classifier's ROC values.

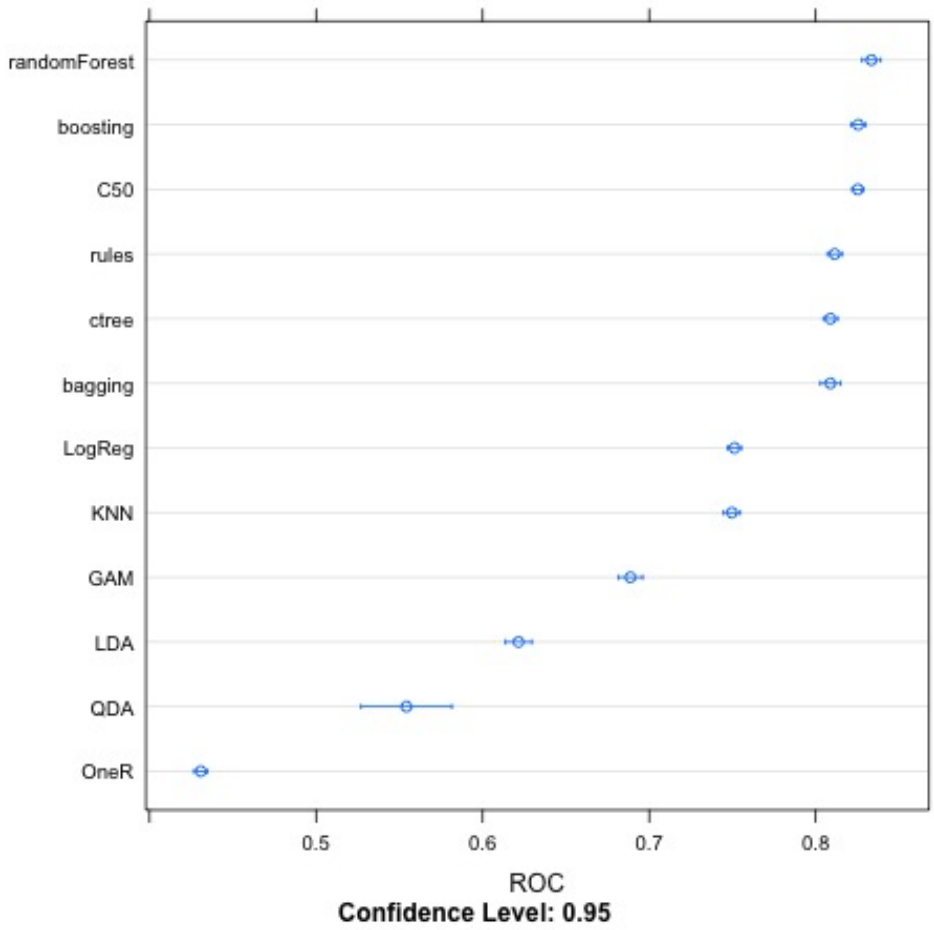



Figure 21. Navy Classifier Comparison: ROC AUC Values

Similar to the Fairchild data set, the decision tree models dominate the other classifiers. Specifically, the random forests classifier achieves the highest ROC AUC values, although the classification tree ensemble models all yield ROC AUC values above 0.8. Logistic regression leads the parametric models with a ROC AUC value between 0.7 and 0.8. Given these performance results, the random forests classifier is initially selected for the highest prediction accuracy and logistic regression is selected as the optimum model for inference.

### Inference

The inference objective seeks to identify important predictors and relationships between the predictors and the response. Two specific inference questions for the Navy data set are as follows:

1. What real property features contribute to classifying mission critical infrastructure?
2. What are the relationships between category groups and mission critical infrastructure?

As with the Fairchild data, the lasso method is employed with logistic regression to increase interpretability. The lasso model is trained with all of the features in the Navy data set: unit of measurement, measurement value, PRV, facility class, and category group. The numeric features are standardized to have mean equal to zero and standard deviation of one. The categorical variables are decomposed into their respective factor levels. Ten-fold cross validation is utilized to select the optimum lambda with ROC as the performance metric. The highest ROC value attained is 0.75 at a lambda value of 0.002. A comparison of the training and test performance suggests that the model is not overfitting as the training and test errors are similar. The model output is the probability that a given observation is mission critical based on the features and coefficient values. The lasso model yields very low sensitivity levels and high specificity levels. This means that the classifier is primarily predicting non-mission critical (the majority class) except for a small fraction of the observations.

The lasso model retains 56 real property features as significant with respect to classifying mission critical infrastructure. The results indicate that three general categories of real property data that are significant in distinguishing between mission

critical and non-mission critical infrastructure. Each coefficient value equates to the respective change in the log odds for the respective feature, assuming all other feature values remain constant. Generally, coefficient estimates with a value of zero are indicative of no association, positive values indicate an increased likelihood of mission critical, and negative values indicate a decreased likelihood of mission critical. Furthermore, taking the exponent of the coefficient value yields the odds ratio value for the feature, which equates to the odds change for a one-unit change in the feature value. The 56 significant Navy real property predictors identified in the lasso model are presented Table 15.

Table 15. Lasso Model Results for Navy Data Set

| Feature Name | Coefficient | Odds Ratio |
|---|---|---|
| (Intercept) | -2.4099 | |
| categoryGroup13 | 0.2928 | 1.34 |
| UMKV | 0.2234 | 1.25 |
| PRV | 0.2109 | 1.23 |
| categoryGroup89 | 0.1989 | 1.22 |
| MEASUREMENT | 0.1773 | 1.19 |
| UMLF | 0.1572 | 1.17 |
| facilityClass3 | 0.1538 | 1.17 |
| categoryGroup42 | 0.1442 | 1.16 |
| UMSY | 0.1279 | 1.14 |
| categoryGroup15 | 0.1219 | 1.13 |
| categoryGroup73 | 0.0882 | 1.09 |
| categoryGroup84 | 0.0771 | 1.08 |
| categoryGroup81 | 0.0641 | 1.07 |
| categoryGroup14 | 0.0408 | 1.04 |
| UMKG | 0.0331 | 1.03 |
| UMOL | 0.0324 | 1.03 |
| categoryGroup21 | 0.0320 | 1.03 |
| categoryGroup39 | 0.0290 | 1.03 |
| categoryGroup16 | 0.0277 | 1.03 |
| UMGA | 0.0218 | 1.02 |
| categoryGroup86 | 0.0190 | 1.02 |
| categoryGroup32 | 0.0132 | 1.01 |
| UMTR | 0.0077 | 1.01 |
| categoryGroup43 | 0.0067 | 1.01 |
| UMMB | 0.0065 | 1.01 |
| categoryGroup12 | -0.0094 | 0.99 |
| categoryGroup55 | -0.0102 | 0.99 |
| UMBL | -0.0115 | 0.99 |
| categoryGroup82 | -0.0119 | 0.99 |
| UMKW | -0.0195 | 0.98 |
| categoryGroup83 | -0.0204 | 0.98 |
| categoryGroup51 | -0.0207 | 0.98 |
| categoryGroup54 | -0.0235 | 0.98 |
| UMFB | -0.0298 | 0.97 |
| facilityClass5 | -0.0335 | 0.97 |
| categoryGroup87 | -0.0398 | 0.96 |
| categoryGroup41 | -0.0423 | 0.96 |
| UMMI | -0.0465 | 0.95 |
| categoryGroup74 | -0.0517 | 0.95 |
| UMGM | -0.0568 | 0.94 |
| categoryGroup45 | -0.0621 | 0.94 |
| categoryGroup71 | -0.0624 | 0.94 |
| facilityClass6 | -0.0773 | 0.93 |
| categoryGroup76 | -0.0876 | 0.92 |
| UMMG | -0.0947 | 0.91 |
| categoryGroup22 | -0.1320 | 0.88 |
| UMEA | -0.1601 | 0.85 |
| categoryGroup44 | -0.1799 | 0.84 |
| categoryGroup85 | -0.2795 | 0.76 |
| categoryGroup69 | -0.4837 | 0.62 |
| categoryGroup75 | -0.5074 | 0.60 |
| facilityClass7 | -0.5642 | 0.57 |

To further investigate the infrastructure functions designated by the category group codes, the lasso model is trained again using only the category group feature. This allows for an apples-to-apples comparison of the likelihood that a given category group is classified as mission critical. Given the estimated coefficient values, the odds ratio is calculated by taking the exponent of the coefficient. An odds ratio greater than one indicates a greater likelihood of mission critical and an odds ratio less than one indicates a decreased likelihood of mission critical. The results and interpretation of the lasso model with the category group feature are presented in Table 16.

Table 16. Lasso Model Results for Navy Data Set Category Group Feature

| Category Group | Description | Coefficient | Odds Ratio | Interpretation |
|---|---|---|---|---|
| (Intercept) | | -1.8398 | | |
| 51 | Medical Centers & Support Facilities | 1.0763 | 2.93 | As Odds Ratio Increases, Likelihood of Mission Critical Increases |
| 15 | Waterfront Operational Facilities | 1.0444 | 2.84 | |
| 13 | Comm, Navigation Aids, Airfield Light | 0.9924 | 2.70 | |
| 81 | Electrical Power | 0.6859 | 1.99 | |
| 31 | RDT&E Buildings | 0.5361 | 1.71 | |
| 89 | Miscellaneous Utilities | 0.4421 | 1.56 | |
| 84 | Water | 0.4044 | 1.50 | |
| 39 | RDT&E Facilities Other Than Buildings | 0.2955 | 1.34 | |
| 43 | Cold Storage | 0.2434 | 1.28 | |
| 42 | Ammunition Storage | 0.2202 | 1.25 | |
| 16 | Harbor & Coastal Operational Facilities | 0.1411 | 1.15 | |
| 82 | Heat and Refrigeration | 0.0626 | 1.06 | |
| 21 | Maintenance Facilities | -0.0173 | 0.98 | As Odds Ratio Decreases, Likelihood of Mission Critical Decreases |
| 87 | Ground Improvement Structures | -0.0648 | 0.94 | |
| 12 | Liquid Fueling and Dispensing Facilities | -0.0875 | 0.92 | |
| 37 | RDT&E Range Facilities | -0.0988 | 0.91 | |
| 14 | Land Operational Facilities | -0.1744 | 0.84 | |
| 17 | Training Facilities | -0.2076 | 0.81 | |
| 83 | Sewage and Waste | -0.2313 | 0.79 | |
| 61 | Administrative Buildings | -0.4435 | 0.64 | |
| 41 | Liquid Storage; Fuel & Non-propellants | -0.4993 | 0.61 | |
| 55 | Dispensaries and Clinics | -0.5827 | 0.56 | |
| 62 | Underground Administrative Structures | -0.6511 | 0.52 | |
| 53 | Medical and Medical Support Facilities | -0.6743 | 0.51 | |
| 85 | Roads and Other Pavements | -0.8610 | 0.42 | |
| 45 | Open Storage | -0.9212 | 0.40 | |
| 44 | Covered Storage | -1.2032 | 0.30 | |
| 73 | Personnel Support & Services Facilities | -1.2588 | 0.28 | |
| 72 | Unaccompanied Personnel Housing | -1.6092 | 0.20 | |
| 22 | Production Facilities | -1.8958 | 0.15 | |
| 54 | Dental Clinics | -1.9512 | 0.14 | |
| 74 | Indoor MWR Facilities | -1.9705 | 0.14 | |
| 71 | Family Housing | -2.2637 | 0.10 | |
| 76 | Museums & Memorials | -3.1229 | 0.04 | |
| 69 | Admin Structures Other Than Buildings | -4.2931 | 0.01 | |
| 75 | Outdoor MWR Facilities | -4.3067 | 0.01 | |

The category group rankings appear to be consistent with intuitions about mission critical infrastructure. Two of the common themes associated with the highest ranked infrastructure functions include direct ties operational missions (e.g. navigational aids) and uniqueness. For example, ammunition storage facilities have very specific requirements and it would be unadvisable to store ammunition in a facility not intended for ammunition. This supports the intuition that redundancy, or a lack thereof, plays a role in mission critical infrastructure.

Alternatively, for category groups with a negative coefficient, there are typically alternatives or workarounds available and delays or inaccessibility will most likely have little negative impact on mission execution. For example, administration structures tend to be fairly generic and redundant capability is likely available, if required. Additionally, morale, welfare, and recreational (MWR) facilities are not required for mission execution. The category group lasso model results provide a notional hierarchy for generic function codes. This general interpretation of the Navy real property inventory functions and their respective likelihood of being mission critical aids in shaping the intuitions surrounding mission criticality and built infrastructure.

### Prediction

For the prediction goal, the random forests classifier outperforms the other classification tree models by a small margin. The random forests algorithm is similar to bagging with the exception that a random subset of features is employed in the learning process. As such, the unique tuning parameter of interest for random forests is the total number of features used in the random feature selection known as “mtry”. In order to select the best tuning parameter, cross validation is employed with classifiers trained at

all possible tuning parameter values. Figure 22 presents the results of the random forests "mtry" tuning iterations for the Navy training data.

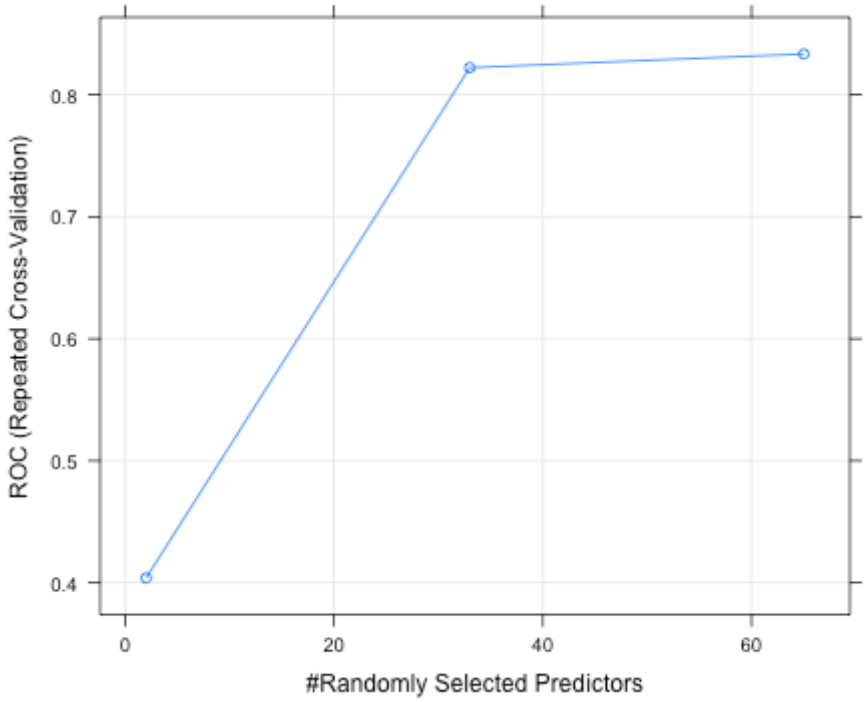



Figure 22. Navy Random Forests Classifier Tuning Parameter

The tuning parameter plot reveals that there is a sharp increase in the ROC value up to 32 randomly selected predictors, after which the performance improves only slightly as the number of features increases to the maximum of 65. Using the total number of predictors is essentially the bagging classification method because there is no random feature subset. Given that the classification tree models performed similarly well, it is prudent to further analyze their performance by comparing their respective sensitivity and specificity results. Figure 23 displays the ROC values, sensitivity, and specificity for the respective tree models.

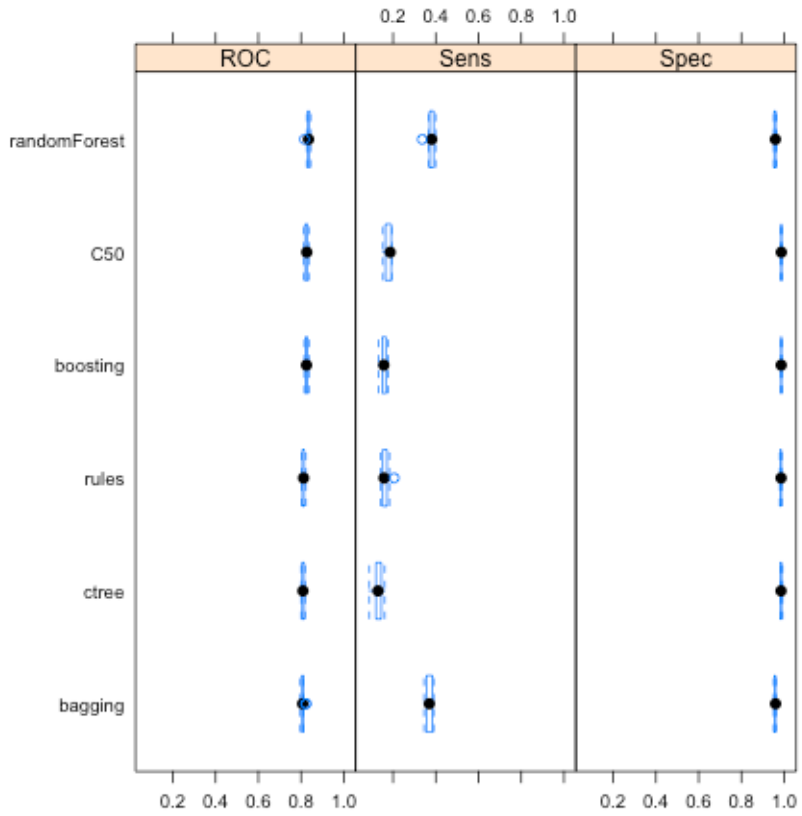



Figure 23. Navy Decision Tree Comparison

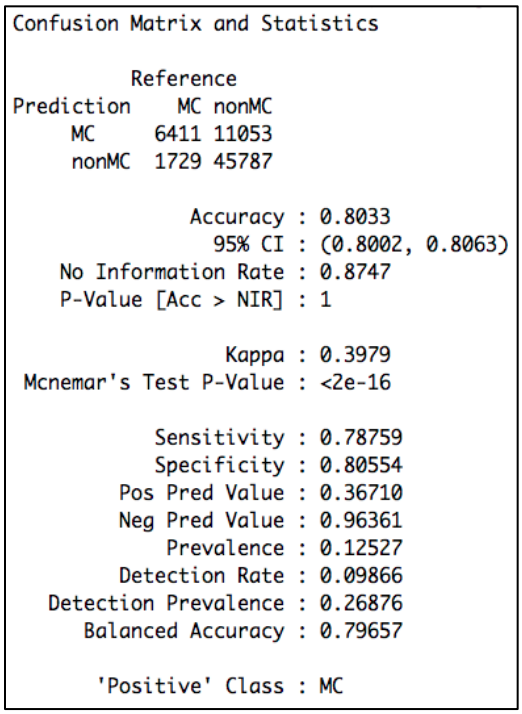
The decision tree comparison reveals that each model yields low sensitivity and high specificity. This means that the model is very good at identifying non-mission critical infrastructure and not good at identifying mission-critical infrastructure. The class imbalance may be causing the bias towards the majority class, non-mission critical. Despite having decent ROC values, the classifier performance for the Navy training data is unacceptable for use as a decision support tool. The next step is to pursue a means of increasing the model sensitivity.

The C5.0 classification tree algorithm enables the use of a "cost matrix", which penalizes classification mistakes. The cost matrix specifies the penalization costs for both false positive and false negative prediction errors. Initially, the C5.0 model is implemented using 10-fold cross validation with cost values between one and ten as tuning parameter levels. The five features employed in model training are PRV, measurement, unit of measure, facility class, and category group with each factor variable decomposed into dummy variables. The tuning parameter comparison indicates that a cost value of five yields the best compromise between model accuracy, sensitivity, and specificity. The cost matrix employed with the C5.0 algorithm is displayed in Table 17.

Table 17. Cost Matrix for C5.0 Algorithm

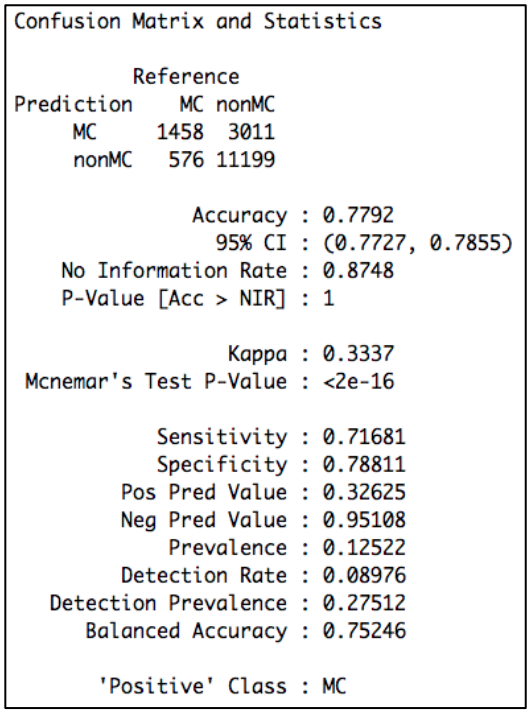
| | Predicted Class | |
|---|---|---|
| True Class | Mission Critical | Non-Mission Critical |
| Mission Critical | 0 | 1 |
| Non-Mission Critical | 5 | 0 |

The C5.0 algorithm with the cost matrix yields an overall training set accuracy of 0.80, sensitivity of 0.79, and specificity of 0.81 on the Navy training data. Next, the training and test set results are compared to determine if the model is overfitting on unseen examples. The training set and test set results are displayed in Figure 24 and Figure 25, respectively.



```
Confusion Matrix and Statistics

          Reference
Prediction    MC nonMC
     MC     6411 11053
     nonMC  1729 45787

               Accuracy : 0.8033
                 95% CI : (0.8002, 0.8063)
    No Information Rate : 0.8747
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3979
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.78759
            Specificity : 0.80554
         Pos Pred Value : 0.36710
         Neg Pred Value : 0.96361
             Prevalence : 0.12527
         Detection Rate : 0.09866
   Detection Prevalence : 0.26876
      Balanced Accuracy : 0.79657

       'Positive' Class : MC
```

Figure 24. Training Set Results for Navy C5.0 Classifier with Cost Matrix



```
Confusion Matrix and Statistics

          Reference
Prediction    MC nonMC
     MC     1458  3011
     nonMC   576 11199

               Accuracy : 0.7792
                 95% CI : (0.7727, 0.7855)
    No Information Rate : 0.8748
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3337
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.71681
            Specificity : 0.78811
         Pos Pred Value : 0.32625
         Neg Pred Value : 0.95108
             Prevalence : 0.12522
         Detection Rate : 0.08976
   Detection Prevalence : 0.27512
      Balanced Accuracy : 0.75246

       'Positive' Class : MC
```

Figure 25. Test Set Results for Navy C5.0 Classifier with Cost Matrix

As expected, the training set accuracy is higher than the test set accuracy but only by a margin of about two percent. The sensitivity experiences a larger decrease in performance dropping from a 0.79 true positive rate on the training data to 0.71 on the test data. The specificity decreases 0.81 on the training data to 0.79 on the test set. The decrease in accuracy between the training and test data is not indicative of significant overfitting.

Finally, limited inference capability is provided with the C5.0 decision tree classifier. The caret package in R enables variable importance estimates with the "varImp()" function. For C5.0 decision trees, the variable importance function calculates the percentage of the training samples in the terminal nodes after the split (inside-R.org, 2016). For example, the predictor used for the first split affects the rest of the splits and, therefore, has a feature importance of 100 percent. The C5.0 classifier employs the facility class, category group, PRV, measurement, and unit of measure features. The feature importance function is limited, however, in that it does not provide insight into the underlying decisions made. For this reason, less flexible models like logistic regression are favored for inference. Figure 26 presents the top 20 features employed in C5.0 classifier.

```
C5.0Cost variable importance

  only 20 most important variables shown (out of 65)

                Overall
facilityClass7   100.00
PRV               91.27
categoryGroup22   81.43
categoryGroup44   80.85
MEASUREMENT       80.71
categoryGroup85   77.78
facilityClass6    69.24
UMKV              59.95
categoryGroup13   56.17
UMGM              52.67
UMBL              49.54
UMEA              48.13
facilityClass2    41.39
categoryGroup17   40.44
UMMB              35.50
categoryGroup45   35.43
categoryGroup89   33.03
categoryGroup12   30.30
facilityClass5    29.78
categoryGroup14   29.62
```

Figure 26. Variable Importance: C5.0 Classifier with Cost Matrix

The primary features employed by the C5.0 decision tree algorithm reveal that physical characteristics, monetary value, and general facility use categories are instrumental in distinguishing between non-mission critical and mission critical infrastructure. The Navy real property features available do appear to have limitations with predictive accuracy, however. With additional real property features, the prediction accuracy may improve.

**Step 9: Using Discovered Knowledge**

The final step of the KDD process is to employ the discovered knowledge. The KDD goals in this research are to develop a prediction model for classifying mission critical infrastructure and identify relationships between real property features and the mission dependency index metric. The prediction model is intended for use as a decision support tool in the USAF MDI adjudication process. By training the prediction model on real property assets with MDI labels determined via stakeholder input, the model provides a more user-oriented prediction as compared to the interim USAF MDI assignment via CATCODE.

AFCENT provided real property data for the three primary air bases in Southwest Asia: Ali Al Salem Air Base, Kuwait; Al Dhafra Air Base, United Arab Emirates; and Al Udeid Air Base, Qatar. The original real property data contained 2,037 observations with 15 features. The AFCENT RP data provided contained the features listed in Table 18.

Table 18. AFCENT Real Property Data Features

| Feature Name | Data Type |
|---|---|
| FACILITY NUMBER | Integer |
| INTEREST CODE | Integer, Text |
| FACILITY TYPE | Text |
| TYPE CONSTRUCTION | Text |
| CATEGORY CODE | Integer |
| LOCAL DESIGNATION | Text |
| RPA DESCRIPTION | Text |
| UNIT OF MEASURE QUANTITY | Integer |
| RPA UNIT OF MEASURE CODE | Text |
| MDI | Integer |
| COST BASIS | Integer |
| PLANT REPLACEMENT VALUE | Integer |
| CREATE DT YEAR | Year |
| MAJCOM CLAIMANT | Integer, Text |

The C5.0 cost matrix prediction model trained with the Navy real property data yielded the best classification results and is employed in predicting AFCENT mission critical infrastructure. Prior to using the model for predictions on the AFCENT infrastructure, the AFCENT real property data is preprocessed to align with the real property data set utilized to train the classifier. Next, the model is employed in classifying AFCENT infrastructure as either mission critical or non-mission critical. The classifier predictions are then compared with the original USAF mission critical labels to identify specific facilities that do not align. This comparison provides AFCENT personnel with a subset of facilities to investigate for possible MDI adjudication. The classifier results for each installation are presented in Table 19.

Table 19. Classifier Results for AFCENT Installations

| | Number of Facilities Identified | Increased to MC | Decreased to non-MC |
|---|---|---|---|
| Ali Al Salem Air Base | 66 | 57 | 9 |
| Al Dhafra Air Base | 45 | 31 | 14 |
| Al Udeid Air Base | 236 | 228 | 8 |
| **Totals:** | **347** | **316** | **31** |

The facility classes and category groups are of particular interest with respect to the classifier predictions. Subsetting the mission critical prediction discrepancies by function codes allows for a better understanding the primary disconnects between the Navy MDI assignment process and the USAF MDI assignment process. The Navy and USAF discrepancy frequencies by facility class (one-digit function codes) and category group (two-digit function codes) are presented in Figure 27 and Figure 28, respectively.

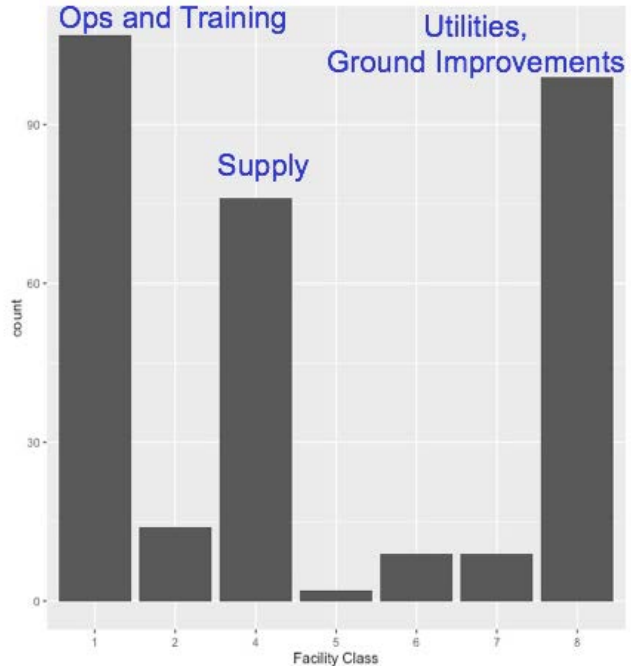



Figure 27. Facility Class Frequencies for Mission Critical Predictions

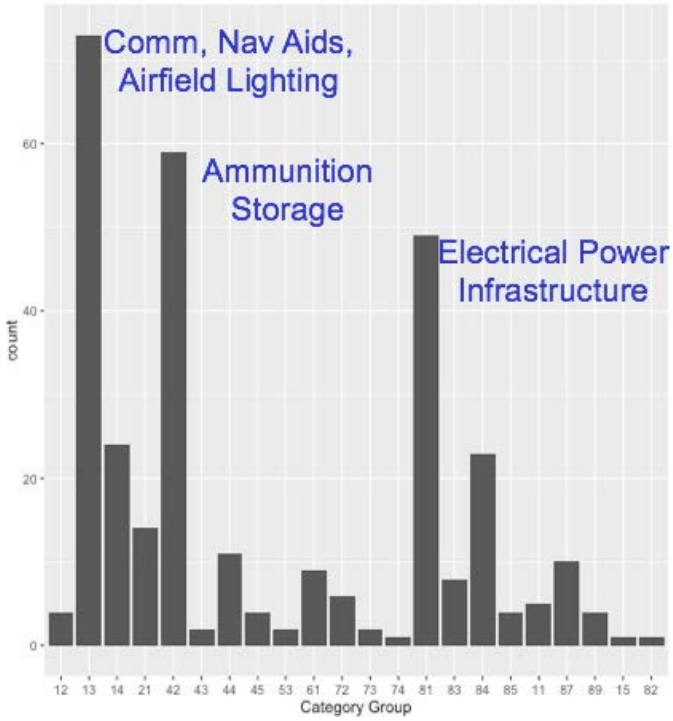



Figure 28. Category Group Frequencies for Mission Critical Predictions

The facility classes with the most discrepancies are one, four, and eight. Facility class one is operations and training, facility class four is supply, and facility class eight is utility and ground improvements. Furthermore, the frequencies indicate that the category groups with the most discrepancies are 13, 42, and 81. Category group 13 encompasses communication, navigation aids, and airfield lighting; category group 42 is ammunition storage; and category group 81 is electrical power infrastructure. The infrastructure in these specific facility classes and category groups enable mission execution and indicate general infrastructure functions that should be investigated further for MDI adjudication.

## Chapter Summary

In conclusion, the decision tree models yielded the best prediction results for classifying mission critical infrastructure. The final classifier yields an accuracy of 77.9 percent on the test data sensitivity and specificity values of 71.6 and 78.8, respectively. This model is utilized to predict mission critical facilities for three AFCENT installations. In this capacity, the classifier provides a decision support tool for identifying potential MDI discrepancies for further investigation. The classifier identified 347 AFCENT infrastructure assets as possible MDI discrepancies. These results enable AFCENT engineers to narrow in on specific facilities for possible MDI adjudication. Furthermore, supervised learning methods provide insight into real property features through inferential analysis. Logistic regression is employed with the lasso method to identify important real property features and supplement intuitions associated with mission critical infrastructure.

# V. Conclusions and Recommendations

## Chapter Overview

The KDD process provides a unique framework for better understanding the mission dependency index problem. Real property data in the federal government is abundant due to recent changes in infrastructure asset management. In many ways, the plethora of real property data represents a dormant resource. This research scratches the surface of the possibilities that machine learning techniques provide with respect to mining real property databases for useful knowledge about government real property portfolios. This chapter answers the investigative questions and summarizes the conclusions and recommendations from the knowledge discovery process.

## Investigative Questions Answered

### 1. How can machine learning techniques, specifically supervised learning, be applied to predict mission critical USAF facilities?

The research indicates that non-parametric learning algorithms yielded the best classification performance. Specifically, classification trees are best suited for real property data, which consists of both numeric and categorical features. Logistic regression yielded the best platform for model interpretability and inference, however, the prediction accuracy is sub-optimal with very low sensitivity levels. Similarly, despite achieving a good ROC value, the random forests algorithm does not provide adequate sensitivity to employ as a decision support tool for the MDI adjudication process. The C5.0 algorithm employed with a corresponding cost matrix provided the best platform for reasonable sensitivity and specificity levels. Overall, supervised learning techniques

offer a myriad of tools to gain knowledge from real property features. Specifically, the real property data features in the research provided limited predictive capability for identifying mission critical infrastructure (MDI greater than or equal to 85).

**2. What features should be collected for such an algorithm?**

Existing real property data, while extensive, provides limited prediction capability for discriminating between mission critical and non-mission critical infrastructure. Algorithm selection relies heavily on the available data. For example, LDA and QDA are employed with numeric data, while logistic regression and classification trees can utilize both numeric and categorical data. The data features employed in the final classification model included generic function codes (facility class and category groups), plant replacement value, measurement type, and measurement value. These features provided enough information about the infrastructure to yield a 75 percent balanced accuracy between the true positives and true negatives on the test data set. More specific four-digit function codes (FACs) were not employed in the final model due to the significantly increased computational costs and observed overfitting to the training data.

In pursuit of the inference objective, logistic regression with the lasso method identified six feature categories within the USAF real property data that proved significant for the given classification task. The six categories include (1) infrastructure characteristics, (2) financial characteristics, (3) organization codes, (4) regulatory codes, (5) historic status, and (6) infrastructure function. The data features within these six categories ultimately suggest that the themes surrounding mission critical infrastructure include uniqueness of the physical infrastructure, organizational ties, age, and infrastructure function.

Given that the best classification model resulted in sensitivity and specificity levels between 0.70 and 0.80 on the test data, real property data appears to be limited in discriminating between mission critical and non-mission critical infrastructure. This suggests that data collected via the MDI survey process captures information about infrastructure-mission relationships that do not exist in real property data. This research identifies that additional data is necessary to classify mission critical infrastructure with high accuracy.

Proposed data features that would contribute to classifying mission critical infrastructure include (1) the occupying organization's relationship to specific mission(s), (2) maximum infrastructure down time without mission degradation, and (3) the number of co-located redundant infrastructure assets. The occupying organization plays a major role in determining mission criticality. There is a natural hierarchy with respect to mission execution. This is sometimes captured via CATCODEs for infrastructure like airfield pavements, however, many CATCODEs do not specifically identify the occupant and their relationship to the mission. A common functional or organizational hierarchy framework for facility occupants could be useful as an additional predictor. Second, infrastructure is in place to serve a specific function. Levels of service for infrastructure differs by the specific support provided and the specific mission(s) supported. Also, infrastructure downtime measured in units of time provides an easily understood metric for data collection. This metric alone could serve as a strong discriminator for mission critical infrastructure. Third, redundancy is a key element of mission criticality. High value assets with no redundant capability present a higher risk than an asset with multiple back-up options. A data feature that captures the number of legitimate redundant

infrastructure assets could serve as a strong discriminator when coupled with infrastructure function. For example, an air traffic control tower with no co-located redundant capability is indicative of high mission criticality for a flying mission.

**3. What is the appropriate architecture for such an algorithm?**

Out of the classification algorithms evaluated, multiple decision tree classifiers produced the highest ROC values. Specifically, the C5.0 decision tree algorithm with a cost matrix yields the most suitable compromise between prediction accuracy, sensitivity, and specificity for an MDI adjudication decision support tool. Alternatively, logistic regression provides the best classifier for interpretably and inference but suffers from low sensitivity levels.

**4. What are the costs and benefits associated with employing machine learning in Air Force asset management facility prioritization?**

There are two primary potential benefits associated with using machine learning for Air Force asset management prioritization. First, machine learning provides a means of automating tedious tasks. Given adequate labeled observations and relevant data features, supervised learning techniques provide the opportunity to automate record reviews such as the MDI adjudication process. The MDI adjudication process currently requires manual record reviews and inter-agency coordination. Second, machine learning could be beneficial in minimizing data collection required for specific tasks. By collecting data from a subset of a given population, a machine learning model could be employed for prediction with the rest of the population. For example, data collection for MDI data could be minimized to a subset of installations within each MAJCOM in order

to build predictive models for facility prioritization. This application would minimize the resources and costs associated with enterprise-wide data collection.

While machine learning offers powerful tools for data analysis, there are five primary costs associated with employing machine learning techniques for facility prioritization. These costs include the requirement for labeled data, models versus subject matter expert judgment, analytical expertise, concerns with "black box" models, and computational resources. First, labeled data is a prerequisite for supervised learning and requires deliberate investment. USAF real property data represents facts about infrastructure at a very basic level. Real property data is typically collected immediately upon commissioning of a given facility and is reviewed and updated annually. This research reveals that real property data is insufficient for training strong classifiers for mission critical infrastructure.

Second, machine learning techniques should not be employed in place of subject matter expert judgment for complex decision-making tasks such as facility prioritization. Mission dependency index is essentially an attempt to capture tacit knowledge from experts. Enterprise-wide application of the MDI should rely on a solid foundation of data collection with deliberate metrics aimed at capturing tacit knowledge from USAF personnel who are intimately familiar with the facilities that support mission execution.

The third cost associated with machine learning techniques is analytical expertise and time for analysis. Machine learning combines technical aspects from numerous fields including statistics and computer science. As such, machine learning techniques require a certain amount of expertise and experience to employ effectively. Generally speaking, the skills required to execute machine learning techniques are not common in

the USAF civil engineer career field. As USAF asset management evolves and data quality improves, opportunities for machine learning applications will increase. Emphasis should be placed on employing the right people with the knowledge, skills, and abilities to discover knowledge from databases. Furthermore, time is a commodity in all USAF functional areas and attempting to employ machine learning through side projects or as an additional duty is suboptimal. The iterative nature of the KDD process necessitates dedicated personnel with time to focus on analysis. The KDD process establishes domain knowledge as a prerequisite for data mining. USAF civil engineering personnel have the domain knowledge required for asset management problems. It is time to start investing in the right people and skills to supplement civil engineer domain knowledge and discover knowledge from data.

The fourth cost associated with machine learning techniques is the potential danger of applying "black-box" techniques to complex problems. A "black-box" is a model that offers limited explanation of the model inter-workings and the general transition from input to output. Using "black-box" methods can lead to deceptive models and can severely limit credibility with decision makers. Furthermore, "black-box" models do not necessarily contribute to understanding the underlying phenomenon, which limits the usefulness of the knowledge discovery process.

Finally, employing machine learning techniques requires dedicated time and computational resources, including hardware and software. While computing power is ever-increasing, complex models require significant processing. Adequate investment in capable hardware and software is a necessary precursor to dedicated data analysis. This

could pose a limitation in applying machine learning techniques within the civil engineer career field.

5. **How can the Knowledge Discovery in Databases (KDD) process be applied to facilitate MDI reviews for AFCENT facilities?**

This research employed the KDD process as a means of automating the MDI review and adjudication process. The KDD process emphasizes a solid understanding of the problem and domain. Data collection and preparation are conducted within the context of the MDI and military infrastructure prioritization problem. Algorithm selection and implementation are based on the identification of mission critical infrastructure with an MDI value of 85 or higher. The final classification model is utilized to identify likely AFCENT mission critical infrastructure in order to minimize the requirement for manual MDI reviews. Ultimately the results of this KDD application provide a means of decreasing the personnel and time requirements for AFCENT MDI adjudication.

## Conclusions of Research

In conclusion, the KDD process provided a solid framework for the USAF civil engineering-specific MDI problem using real property data from the United States Navy. This is just one example of how machine learning techniques can be applied to automate tedious tasks and provide relevant and objective tools for domain-specific problems. The results provided enhanced intuitions about mission critical infrastructure in the context of real property. Furthermore, the classification model provides AFCENT civil engineers with a tool to minimize personnel-hours associated with manual MDI reviews by identifying a subset of facilities for further investigation.

## Significance of Research

Infrastructure asset management is built on a foundation of data including physical characteristics, condition, and function. This data is contained in databases that represent dormant resources in the absence of the analytical expertise required for knowledge discovery. The future of asset management lies in high quality data and analytical techniques to better forecast mission requirements, life cycle costs, and resource allocation decisions. The USAF civil engineer career field should embrace machine learning and commit to training, organizing, and equipping personnel to employ these techniques and enhance asset management practices.

## Recommendations for Action

One recommendation for consideration by USAF civil engineer leadership is a "Civil Engineering Data Analysis Center of Excellence". Data analysis requires a certain level of analytical expertise and training. By investing in personnel with data mining and data science expertise, thorough and objective analysis will improve enterprise-wide decision-making. Additionally, appropriate investment in computational resources is necessary to execute complex data analysis in a timely fashion. Predictive analytics is commonplace in industry and the USAF should not neglect the powerful tools machine learning has to offer.

## Recommendations for Future Research

This research emphasizes existing limitations associated with the USAF mission dependency index. Mission dependency is a function of infrastructure purpose and the relationship with specific mission objectives. Currently, MDI is determined solely by CATCODE, a generic infrastructure function identifier. The missing piece of the USAF mission dependency index is the linkage with specific mission objectives. The USAF missions are clearly identified in the 2023 implementation plan, however, the connection between these missions and the infrastructure required for mission execution remains uncertain. Future MDI research should focus on two primary lines of efforts, mapping specific infrastructure assets to specific missions using measureable data features and developing a reliable and repeatable process for stakeholder data collection.

The first line of effort should focus on developing metrics to capture facts about the consequence of failure associated with specific USAF infrastructure. This research reveals that real property data has limited capability for predicting mission critical infrastructure. Existing real property data does not capture the linkage between infrastructure assets and the missions they support or to what degree they contribute to mission execution. Essentially, interdependencies and intradependencies are not taken into account with USAF MDI values. Generic infrastructure functions identified by CATCODES do not provide the level of granularity necessary to compare infrastructure assets. Data should be collected from the field in order to truly identify mission critical infrastructure with any degree of certainty. In the context of mission dependency, infrastructure can be viewed as a "network" of nodes with varying degrees of connection to and impact on executing the specific USAF mission priorities listed in the 2023

implementation plan. Future research should include developing specific data features that (1) link specific infrastructure to specific missions, (2) capture infrastructure redundancies at the base, MAJCOM, and USAF levels, and (3) capture time-based metrics for allowable infrastructure downtime.

The second line of effort should focus on developing a reliable and repeatable data collection process for mission dependency data elements. Data collection options abound with web-based survey and data collection tools. The USAF MDI methodology suffers from a lack of data collection. The mission dependency problem transcends real property data and requires a more holistic solution than assignment by CATCODE. We cannot use data features of the past to solve existing and future problems. Infrastructure asset management requires deliberate data collection for effective decision-making. Data collection for MDI should not have to cost millions of dollars per year. Further research into web-based streamlined data collection is necessary to improve upon the existing MDI metric and resource allocation framework.

**Summary**

This research shows that machine learning is a strong contender for solving asset management specific problems. As infrastructure data relevance and quality improves, knowledge of machine learning techniques can provide domain experts with options for gleaning knowledge from databases. Furthermore, data science and machine learning should be incorporated into the USAF asset management framework in order to guide the future of asset management and establish best practices in managing a vast real property portfolio.

## Appendix A. Data Mining Algorithms

### Logistic Regression

Logistic regression is very popular model for binary classification problems. Logistic regression is a linear model that predicts the probability that a given observation belongs to a specific class. Given the probabilities associated with the observations, a threshold for the probabilities is selected to classify each observation. As the logistic regression outputs a probability, the range of values will always fall between 0 and 1. The logistic function is presented in Equation 2 (James et al., 2013).

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (2)$$

The logistic function requires a different interpretation from the standard linear regression equation. With some manipulation, the logistic function takes on the form known as the odds. The odds is the probability of belonging to the specified class divided by the probability of not belonging to the specified class. The odds ratio is presented in the left hand side of Equation 3 (James et al., 2013).

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \quad (3)$$

Another important concept in logistic regression is the log-odds or logit, which amounts to a further manipulation of the odds presented previously. The log-odds is presented in the left hand side of Equation 4 (James et al., 2013).

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \quad (4)$$

Ultimately, the beta values (β) are determined based on the training data and the maximum likelihood method. The maximum likelihood method seeks to estimate the beta values that will produce a value close to zero for training observations that do not belong to the class and a value close to one for training observations that do fall in the class. The beta values are chosen to maximize the likelihood equation presented in Equation 5 (James et al., 2013).

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} (p(x_i)) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \quad (5)$$

Once the beta weights are determined via the maximum likelihood, the model can be used for prediction on unseen observations by simply inserting the attribute values into the logistic function.

**Linear Discriminant Analysis**

Linear Discriminant Analysis (LDA) is another a widely used classification model and, like logistic regression, provides a probability for each observation. LDA models the distribution of each predictor separately against the response classes (James et al., 2013). LDA employs Bayes' theorem to produce estimates for the class probability given the value of X, or Pr(Y=k|X=x). LDA makes specific assumptions about the data. "The LDA classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance $\sigma^2$, and plugging estimates for these parameters into the Bayes classifier" (James et al., 2013). Understanding the model assumptions is necessary for comparing LDA with other classifiers. For example, James et al. (2013) describe that LDA can perform better than logistic regression when the Gaussian distribution assumption holds.

### Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) shares the same assumptions made in LDA except that QDA assumes that each class has a distinct covariance matrix (James et al., 2013). Furthermore, QDA assumes a quadratic decision boundary whereas LDA assumes a linear decision boundary. This quadratic decision boundary means that QDA is more flexible than LDA and "can accurately model a wider range of problems than can the linear methods" (James et al., 2013).

### K-nearest Neighbors

The K-nearest Neighbors (KNN) classification method operates in a completely different manner than logistic regression, LDA, and QDA. Specifically, KNN uses measures of distance between observations for classification. In KNN, the "K" term represents the number of neighbors selected for classification. For example, if K is set equal to three, the algorithm will identify the three observations closest to a given observation using a specified distance measurement. Two common distance measurements are Euclidian distance and Manhattan distance. Euclidian distance is the most direct distance between two points and Manhattan distance is measured at right angles along specified axes. Once K is selected, the algorithm assigns the observations to the majority class among their respective neighbors. KNN is a completely non-parametric approach meaning that no assumptions are made about the decision boundary (James et al., 2013). As such, KNN is much more flexible than logistic regression, LDA, and QDA. There are tradeoffs with the increased flexibility, however, as KNN does not

provide any information about which predictor variables are important. Also, the KNN classifier requires selection of an appropriate K value.

### Generalized Additive Models

Generalized additive models (GAMs) are a general framework used to extend linear models with non-linear functions for the predictor variables (James et al., 2013). GAMs can be used for both regression and classification as extensions of linear regression and logistic regression, respectively. The GAM extension for logistic regression is presented in Equation 6 (James et al., 2013).

$$\boldsymbol{log}\left(\frac{\boldsymbol{p(X)}}{\boldsymbol{1-p(X)}}\right)=\boldsymbol{\beta_0}+\boldsymbol{f_1(X_1)}+\boldsymbol{f_2(X_2)}+\cdots+\boldsymbol{f_p(X_p)} \quad (6)$$

GAMs offer a more flexible alternative to linear or logistic regression. Of course, the GAM's applicability depends on the data.

### Decision Trees

Classification and regression trees are a very popular supervised learning method. This section focuses on classification trees to align with the MDI classification problem. Classification trees use recursive partitioning to create a flow-chart-like decision tree for data sets with a qualitative response (Lantz, 2013). Recursive partitioning identifies the best predictor of all features and splits the data into a smaller subset. This process is repeated until some threshold is met or the terminal node reaches an acceptable level of homogeneity. There are numerous methods for determining the best split for a given subset. James et al. (2013) prefer Gini index or cross-entropy over the classification error rate. The Gini index and cross-entropy are both measures of node purity where a small

value indicates that the region consists primarily of the same class. Finally, pruning decision trees may increase prediction accuracy and interpretability (James et al., 2013).

Classification trees differ from regression trees in that the observations are classified by the majority class in a given subset, whereas regression uses the mean response value for the subset (James et al., 2013). Decision trees are lauded for their simplicity and interpretability despite lower accuracy compared to more flexible supervised learning methods (James et al., 2013).

While single decision trees tend to yield limited predictive capability with high variance, more advanced methods exist that employ multiple trees for prediction. These advanced methods include bagging, random forests, and boosting, which tend to provide significant improvements in predictive performance (James et al., 2013).

The term "bagging" is short for bootstrap aggregation. While the bagging concept is generalizable to other learning methods, bagging is often used with decision trees to minimize variance (James et al., 2013). James et al. (2013) describe the value associated with reducing variance:

> A natural way to reduce variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions.

Because training data is finite, the bootstrap method employs random sampling with replacement to generate numerous training data sets. In bagging, individual trees are not pruned so they typically yield high variance and low bias (James et al., 2013). With the trees generated, predictions are then averaged over the bootstrapped training data sets (B)

in order to reduce the variance. Equation 7 presents the bootstrap aggregation equation (James et al., 2013).

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \quad (7)$$

For classification trees, bagging culminates in a majority vote from the individual trees to determine the class label.

Random forests are very similar to bagging except that the random forest method seeks to mitigate correlation among the generated trees by randomly selecting a subset of features available for each split (James et al., 2013). A general rule of thumb for the number of features to provide is the square root of the total number of features or even a single predictor (Friedman, Hastie, & Tibshirani, 2001). Ultimately, the random forests algorithm yields many unique individual trees that, when averaged together, yield lower variance thereby mitigating overfitting on the training data (Friedman et al., 2001).

Finally, the boosting method can be applied to other learning methods but is commonly used with decision trees to improve predictions (James et al., 2013). In employing boosting with decision trees, numerous trees are grown sequentially. Each iteration of tree building uses the the previous tree to improve prediction performance. The three tuning parameters associated with boosting include (1) the number of trees, (2) the shrinkage parameter, and (3) the number of splits in each tree (James et al., 2013).

## Bibliography


AETC. (2015, September 14). Air Education and Training Command - “About Us.” Retrieved September 14, 2015, from http://www.aetc.af.mil/Home.aspx

AF/A7C. (2008, August 14). Real Property Accountability and Reporting. USAF.

AFCEC. (2014a). FY18-22 AFAMP Business Rules. AFCEC.

AFCEC. (2014b, June 26). Real Property Accountability and Inventory Playbook. AFCEC.

AFCEC. (2015, January 15). MDI Refinement Playbook. AFCEC/CPAD.

AFGSC. (2015). Air Force Global Strike Command - “Welcome.” Retrieved September 14, 2015, from http://www.afgsc.af.mil/main/welcome.asp

Albrice, D., Branch, M., & Lee, T. (2014). Municipal portfolio stewardship with limited budgets: the application of matrix correlations as a tool to support resource allocation decisions in the public good. In *Asset Management Conference 2014* (pp. 1–12). IET. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7129238

Amadi-Echendu, J. E., Willett, R., Brown, K., Hope, T., Lee, J., Mathew, J., … Yang, B.-S. (2010). What Is Engineering Asset Management? In J. E. Amadi-Echendu, K. Brown, R. Willett, & J. Mathew (Eds.), *Definitions, Concepts and Scope of Engineering Asset Management* (pp. 3–16). Springer London. Retrieved from http://link.springer.com/chapter/10.1007/978-1-84996-178-3_1

Antelman, A. (2008). *United States Air Force Mission Dependency Index (MDI) Proof of Concept Report*. Port Hueneme, California: Naval Facilities Engineering Service Center.

Beretta, L., & Santaniello, A. (2011). Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets. *Journal of Biomedical Informatics*, *44*(2), 361–369.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information & Management*, *39*(3), 211–225.

Cox, A. (2008). What’s wrong with risk matrices? *Risk Analysis*, *28*(2), 497–512.

Dempsey, J. (2006, October 31). Facility Asset Management Doctrine: A Strategy for Making Better Decisions at Lower Risk and Costs.

Dempsey, J. (n.d.). Mission Dependency Index. Retrieved from http://www.gsa.gov/graphics/ogp/03-PRA-011_R2M-y8V_0Z5RDZ-i34K-pR.doc

DOD. (2005, April 6). Real Property Management. DOD.

DOD. (2013, September 30). Department of Defense Base Structure Report.

DOD. (2014, January 17). Real Property Inentory (RPI) and Forecasting. DOD.

DOD. (2015, February 4). DOD Real Property Categorization.

Eulberg, D. (2008). Managing Air Force Assets. *Air Force Civil Engineer Magazine*, *16*(1). Retrieved from http://www.afcec.af.mil/shared/media/document/AFD-120926-125.pdf

Executive Order No. 13327. (2004, February 6). Federal Real Property Asset Management. Federal Register. Retrieved from http://www.gpo.gov/fdsys/pkg/FR-2004-02-06/pdf/04-2773.pdf

Fayyad, U. M., Piatetsky-Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 panel: data mining: the next 10 years. *ACM SIGKDD Explorations Newsletter*, *5*(2), 191–196.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, *13*(3), 57.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin. Retrieved from http://statweb.stanford.edu/~tibs/book/preface.ps

FRPC. (2011, September 20). 2011 Guidance for Real Property Inventory Reporting. Federal Real Property Council.

FRPC. (2015, June 9). 2015 Guidance for Real Property Inventory Reporting. GSA.

GAO. (1998, December). Executive Guide: Leading Practices in Capital Decision-Making. Retrieved from http://www.gao.gov/products/GAO/AIMD-99-32

GAO. (2003, February 19). Defense Infrastructure: Changes in Funding Priorities and Strategic Planning Needed to Improve the Condition of Military Facilities. Retrieved January 4, 2016, from http://www.gao.gov/products/GAO-03-274

GSA. (2015, April 30). Federal Real Property Counci 2015 Guidance for Real Property Inventory Reporting. General Services Administration Office of Government-wide Policy.

Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R Journal*, *2*(1), 30–38.

Hastie, T., & Qian, J. (2014, June 26). Glmnet Vignette. Retrieved January 25, 2016, from https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet_beta.html#log

Hastie, T., & Tibshirani, R. (2013, December). *Introduction to Statistical Learning*. Video. Retrieved from https://www.youtube.com/watch?v=LvaTokhYnDw

Hodkiewicz, M. R. (2015). The Development of ISO 55000 Series Standards. In *Engineering Asset Management-Systems, Professional Practices and Certification* (pp. 427–438). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-09507-3_37

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(3), 299–310.

Hubbard, D. W. (2014). *How to measure anything: Finding the value of intangibles in business*. John Wiley & Sons. Retrieved from https://books.google.com/books?hl=en&lr=&id=EAPXAgAAQBAJ&oi=fnd&pg=PA175&dq=hubbard+how+to+measure+anything&ots=CqccAZGpG0&sig=_3oIIHMUtD8VCS04MC_yQztU3Dg

IBM. (2015, February 26). What is big data? [CT000]. Retrieved May 25, 2015, from http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

inside-R.org. (2016). varImp {caret}. Retrieved January 24, 2016, from http://www.inside-r.org/packages/cran/caret/docs/varimp

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf

John, G. H., Kohavi, R., Pfleger, K., & others. (1994). Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference* (pp. 121–129).

Kira, K., & Rendell, L. (1992). A Practical Approach to Feature Selection. *Proceedings of the Ninth International Workshop on Machine Learning*.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *Machine Learning: ECML-94* (pp. 171–182). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-57868-4_57

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Kuhn, M. (2012). Variable selection using the caret package. *URL< Http://cran. Cermin. Lipi. Go. id/web/packages/caret/vignettes/caretSelection. Pdf*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.168.1655&rep=rep1&type=pdf

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-6849-3

Kujawski, E., & Miller, G. (2009). The mission dependency index: Fallacies and misuses. In *INCOSE International Symposium* (Vol. 19, pp. 1565–1580). Wiley Online Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2334-5837.2009.tb01035.x/abstract

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, *38*(11), 54–64.

Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd. Retrieved from https://books.google.com/books?hl=en&lr=&id=ZQu8AQAAQBAJ&oi=fnd&pg=PT12&dq=machine+learning+with+R&ots=_7ue-l2lda&sig=aD7q1WDpJOdq77SM7_CyA7nOGls

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*(6), 861–867.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311.

Madaus, M. (2009, June). *Asset Management Optimization Tools*.

Maddox, L. (2014). Making The List. *Air Force Civil Engineer Magazine*, *22*(2).

Maimon, O., & Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook* (1 edition). Boston: Springer.

Maletic, J. I., & Marcus, A. (2010). Data cleansing: A prelude to knowledge discovery. In *Data Mining and Knowledge Discovery Handbook* (pp. 19–32). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-09823-4_2

McElroy, R. S. (1999). Update on national asset management initiatives: facilitating investment decision-making. In *Innovations in Urban Infrastructure Seminar of the APWA International Public Works Congress* (pp. 1–10). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.316&rep=rep1&type=pdf

Michael Grussing, Gunderson, S., Canfield, M., Falconer, E., Antelman, A., & Hunter, S. (2010, September). Development of the Army Facility Mission Dependency Index for Infrastructure Asset Management. USACE.

National Research Council. (2008). *Core Competencies for Federal Facilities Asset Management through 2020: Transformational Strategies*. The National Academies Press Washington, DC.

NAVFAC. (2008a, July). Real Property Inventory (RPI) Procedures Manual. NAVFAC. Retrieved from http://www.navfac.navy.mil/content/dam/navfac/Asset%20Management/PDFs/final_P78_%20july_08_%20for_%20posting.pdf

NAVFAC. (2008b, October). Public Works Department Management Guide. NAVFAC.

Nichols, M. (2015). *A Delphi Study Using Value-Focused Thinking For United States Air Force Mission Dependency Index Values*. Air Force Institute of Technology, Wright-Patterson AFB.

Oreski, D., & Novosel, T. (2014). Comparison of Feature Selection Techniques in Knowledge Discovery Process. Retrieved from http://www.temjournal.com/documents/vol3no4/journals/1/articles/vol3no4/Comparisonoffeatureselectiontechniquesinknowledgediscoveryprocess.pdf

Peng, R. (2015). *R Programming for Data Science*. Leanpub.

Revolution Analytics. (n.d.). Companies Using R. Retrieved January 29, 2015, from http://www.revolutionanalytics.com/companies-using-r

Romanski, P., & Kotthoff, L. (2014, October 25). Package “FSelector.” Retrieved from https://cran.r-project.org/web/packages/FSelector/FSelector.pdf

r-project.org. (n.d.). The R Project for Statistical Computing. Retrieved January 29, 2015, from http://www.r-project.org/

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

SAF/FMB. (2014, February 27). United States Air Force Fiscal Year 2015 Budget Overview. SAF. Retrieved from http://www.saffm.hq.af.mil/shared/media/document/AFD-140304-039.pdf

Sánchez-Maroño, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007). Filter methods for feature selection–a comparative study. In *Intelligent Data Engineering and Automated Learning-IDEAL 2007* (pp. 178–187). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-77226-2_19

Sharp, C. (2002). *An Evaluation of Facility Maintenance and Repair Strategies of Select Companies*. Air Force Institute of Technology, Wright-Patterson AFB.

Smith, D. (2015, January 27). Why now is the time to learn R. Retrieved January 29, 2015, from http://opensource.com/business/14/12/r-open-source-language-data-science

Teicholz, E., Nofrei, C., & Thomas, G. (2005). Executive order# 13327 for real property asset management. *IFMA Journal, November/December*. Retrieved from http://www.graphicsystems.biz/gsi/articles/ifma_executive_order13327.pdf

Tufte, E. R. (2006). Beautiful evidence. *New York*. Retrieved from http://www.maa.org/publications/maa-reviews/beautiful-evidence

USAF. (2007, August 6). Air Force Policy Directive 32-90, Real Property Asset Management.

USAF. (2008). Air Force Instruction 32-9005, Real Property Accountability and Reporting.

USAF. (2013, December 9). Air Force 2023 Implementation Plan. Headquarters Air Force.

USN. (2010, July 2). Operational Risk Management. Retrieved from https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#safe=off&q=opnavinst+3500.39c

Vance, A. (2009, January 6). R, the Software, Finds Fans in Data Analysts. *The New York Times*. Retrieved from http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html

Vanier, D. (2001). Why Industry Needs Asset Management Tools. *Journal of Computing in Civil Engineering*, *15*(1), 35–43. http://doi.org/10.1061/(ASCE)0887-3801(2001)15:1(35)

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10).

Woodhouse, J. (2001). Asset management: Asset management processes and tools. *The Woodhouse Partnership Ltd, UK*.

Wood, S. (2015). *Core Statistics* (1 edition). New York, NY: Cambridge University Press.

Zumel, N., Mount, J., & Porzak, J. (2014). *Practical data science with R*. Manning. Retrieved from http://toc.dreamtechpress.com/toc_978-93-5119-437-8.pdf

| REPORT DOCUMENTATION PAGE | | *Form Approved OMB No. 0704–0188* |
|---|---|---|
| The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, 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 Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704–0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS. | | |
| 1. REPORT DATE (DD–MM–YYYY)<br>24-03-2016 | 2. REPORT TYPE<br>Master's Thesis | 3. DATES COVERED *(From — To)*<br>Sept 2014 – Mar 2016 |
| 4. TITLE AND SUBTITLE<br>Mission Dependency Index of Air Force Built Infrastructure: Knowledge Discovery with Machine Learning | | 5a. CONTRACT NUMBER<br>5b. GRANT NUMBER<br>5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S)<br>Smith, Clark W, Captain, USAF | | 5d. PROJECT NUMBER<br>5e. TASK NUMBER<br>5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>Air Force Institute of Technology<br>Graduate School of Engineering and Management (AFIT/EN)<br>2950 Hobson Way<br>Wright-Patterson AFB OH 45433-7765 | | 8. PERFORMING ORGANIZATION REPORT NUMBER<br>AFIT-ENV-MS-16-M-184 |
| 9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)<br>United States Air Forces Central Command<br>Maj Andrea Griffin<br>Shaw AFB, SC<br>Andrea.Griffin@afcent.af.mil | | 10. SPONSOR/MONITOR'S ACRONYM(S)<br>AFCENT<br>11. SPONSOR/MONITOR'S REPORT NUMBER(S) |
| 12. DISTRIBUTION / AVAILABILITY STATEMENT<br>**DISTRUBTION STATEMENT A.** APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. | | |
| 13. SUPPLEMENTARY NOTES<br>This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States. | | |


14. ABSTRACT

Mission Dependency Index (MDI) is a metric developed to capture the relative criticality of infrastructure assets with respect to organizational missions. The USAF adapted the MDI metric from the United States Navy's MDI methodology. Unlike the Navy's MDI data collection process, the USAF adaptation of the MDI metric employs generic facility category codes (CATCODEs) to assign MDI values. This practice introduces uncertainty into the MDI assignment process with respect to specific missions and specific infrastructure assets. The uncertainty associated with USAF MDI values necessitated the MDI adjudication process. The MDI adjudication process provides a mechanism for installation civil engineer personnel to lobby for accurate MDI values for specific infrastructure assets. The MDI adjudication process requires manual identification of MDI discrepancies, documentation, and extensive coordination between organizations.

Given the existing uncertainty with USAF MDI values and the effort required for the MDI adjudication process, this research pursues machine learning and the knowledge discovery in databases (KDD) process to identify and understand relationships between real property data and mission critical infrastructure. Furthermore, a decision support tool is developed for the MDI adjudication process. Specifically, supervised learning techniques are employed to develop a classifier that can identify potential MDI discrepancies. This automation effort serves to minimize the manual MDI review process by identifying a subset of facilities for potential adjudication.



15. SUBJECT TERMS

Mission Dependency Index, Infrastructure Asset Management, Machine Learning, Knowledge Discovery in Databases


| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Maj Vhance V. Valencia, AFIT/ENV |
| U | U | U | UU | 133 | 19b. TELEPHONE NUMBER (Include Area Code)<br>(937) 255-3636 x4826 Vhance.Valencia@afit.edu |

**Standard Form 298 (Rev. 8–98)**
***Prescribed by ANSI Std. Z39.18***