NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PREDICTING U.S. ARMY RESERVE UNIT MANNING USING MARKET DEMOGRAPHICS

by

Nathan L. Parker

June 2015

Thesis Advisor:	Samuel E. Buttrey
Co-Advisor:	Jonathan K. Alt
Second Reader:	Jeffrey B. House

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2015	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE
PREDICTING U.S. ARMY RESERVE UNIT MANNING USING MARKET DEMOGRAPHICS

5. FUNDING NUMBERS

6. AUTHOR(S) Nathan L. Parker

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943–5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

This thesis develops a data-driven, statistical model capable of predicting a U.S. Army Reserve (USAR) unit's manning level based on the demographics of the unit's location. This model will aid decision-makers involved in USAR stationing by assessing the ability of a proposed stationing location to support a unit's manning requirements. USAR units must recruit the majority of their personnel from the population within immediate proximity to the unit. Since the recruiting boundaries of multiple reserve centers often overlap, this thesis first develops an allocation method that ensures the population is not over-counted. This thesis then develops linear regression, classification tree, and logistic regression models to determine the ability of the location to support manning requirements. These models demonstrate that local demographic factors are a key driver in the ability of unit to meet its manning requirements. In particular, the logistic regression model delivers predictive results that allow decision-makers to identify locations with a high probability of meeting unit manning requirements. The recommendation of this thesis is that the USAR implement the logistic regression model.

14. SUBJECT TERMS
U.S. Army Reserve, USAR, manning, stationing, readiness, recruiting, data analysis, logistic regression, classification tree

15. NUMBER OF PAGES
85

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540–01–280–5500

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

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

PREDICTING U.S. ARMY RESERVE UNIT MANNING USING MARKET DEMOGRAPHICS

Nathan L. Parker
Captain, United States Army
B.S., United States Military Academy, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2015

Author: Nathan L. Parker

Approved by: Samuel E. Buttrey
Thesis Advisor

Jonathan K. Alt
Co-Advisor

Jeffrey B. House
Second Reader

Robert F. Dell
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This thesis develops a data-driven, statistical model capable of predicting a U.S. Army Reserve (USAR) unit's manning level based on the demographics of the unit's location. This model will aid decision-makers involved in USAR stationing by assessing the ability of a proposed stationing location to support a unit's manning requirements. USAR units must recruit the majority of their personnel from the population within immediate proximity to the unit. Since the recruiting boundaries of multiple reserve centers often overlap, this thesis first develops an allocation method that ensures the population is not over-counted. This thesis then develops linear regression, classification tree, and logistic regression models to determine the ability of the location to support manning requirements. These models demonstrate that local demographic factors are a key driver in the ability of unit to meet its manning requirements. In particular, the logistic regression model delivers predictive results that allow decision-makers to identify locations with a high probability of meeting unit manning requirements. The recommendation of this thesis is that the USAR implement the logistic regression model.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AFQT	Armed Forces Qualification Test
AR	Army Reserve
ARSS	Army Reserve Stationing Study
BLS	Bureau of Labor Statistics
CAA	Center for Army Analysis
CDC	Center for Disease Control
DA	Department of the Army
DOD	Department of Defense
FY	Fiscal Year
IET	Initial Entry Training
MSA	Metropolitan Statistical Area
MSS	market supportability study
MOS	military occupational specialty
OCAR	Office of the Chief of Army Reserves
QMA	Qualified Military Available
RA	regular Army
SL1	Skill Level 1
STAR	Stationing Tool Army Reserve
TPU	Troop Program Unit
USAR	United States Army Reserve
USAREC	United States Army Recruiting Command
ZIP	Zone Improvement Plan

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The process for selecting suitable locations for United States Army Reserve (USAR) units is both complex and important. Unlike regular Army units, the geographic location of a reserve unit has a direct impact on its ability to meet manning goals and readiness requirements. The USAR does not have the flexibility to move soldiers to meet manning shortfalls, so each USAR unit must be able to draw a sufficient number of qualified recruits from its local community.

This thesis focuses on the identification of potential stationing locations that have a high probability of supporting the unit's manning requirements in the Skill Level 1 (SL1) ranks, defined as E-1 through E-4. While many other factors are considered in selecting unit locations, the area's ability to fill required manning levels most directly affects unit readiness and is the dominant consideration. Once the USAR is able to identify the set of locations that are capable of supporting the unit's manning requirements, it can apply additional criteria to narrow the set to those that meet force structure and training facility requirements.

The USAR's primary decision-support tool to assess the potential stationing options is the Stationing Tool Army Reserve (STAR), which was developed by a Center for Army Analysis team led by Robert Bradford in 2007. This tool relies on subject matter expert elicited weightings to generate an overall utility score based on a location's ability to meet manning, force structure, and facilities requirements. Current USAR manning data shows that almost 20 percent of USAR locations, selected using the current methodology, are unable to support the manning requirements of their units. This under-manning may be a result of STAR recommending stationing locations outside of sufficient recruiting markets. This thesis uses a data-driven approach to develop a statistically based model that is capable of assessing a reserve location's ability to support manning requirements.

The first step in developing a model to assist the USAR in the stationing process involved gathering the required data. U.S. Army Recruiting Command (USAREC) and

USAR provided the bulk of the data for this analysis. We obtained the remaining data from publicly available sources: the Bureau of Labor Statistics, the Center for Disease Control and Prevention, and the U.S. Census Bureau. The population demographic data includes the number of assigned recruiters, regular Army and USAR accessions, Department of Defense accessions, Armed Forces Qualification Test scores, Qualified Military Available counts, obesity rates, unemployment rates, and post-secondary enrollments at the ZIP-code level. A separate data set includes unit level statistics such as current SL1 authorizations and fills status, along with attrition and location. Since ZIP codes often contain more than one unit, the unit level data is aggregated at the ZIP-code level. The remainder of this summary refers to these ZIP-code level aggregates as reserve locations.

The development of an allocation method is necessary since a population ZIP code may fall within the recruiting boundaries of multiple reserve locations. Without an allocation method, the population in urban areas will be over counted while the population in rural areas will be under counted. This allocation is accomplished by expanding the scope of a method initially developed by Stephen Mehay, in his 1989 report *An Enlistment Supply and Forecasting Model for the U.S. Army Reserve*. The resulting data set contains the population demographic and unit statistics for 599 reserve locations.

Using this data set, we build and compare three predictive models with fill rate as the dependent variable: a linear regression model, a classification tree model, and a logistic regression model. For the classification tree and logistic regression models the response variable is coded as a binary variable, with locations at or exceeding 100 percent fill coded as a one while locations not meeting this criteria are coded as a zero. The final linear regression model retains the number of SL1 authorizations, attritions, USAR accessions, obesity rate and location as the significant factors. This model produces an adjusted R-squared value of 0.292. The final classification tree and logistic regression models both retain the same factors as the linear model with the exception of obesity, which falls out. Both of these models produce a misclassification rate near 25 percent and an area under the curve, or AUC, near 0.75. The logistic regression model is

preferred due to its superior performance in correctly classifying those locations below the 100 percent fill level. All three models indicate that fill rate decreases as the number of SL1 authorizations increase and that fill rate increases as attrition and USAR accessions increase. The direction of influence for attrition is counterintuitive but remains consistent across all three modeling methods. Further research is necessary to determine the causal relationship between attrition and fill rate. All three models also indicate that locations in the southeast produces fill rates higher than those in the rest of the country.

The recommendation of this thesis is that the USAR implement the logistic regression model developed in the analysis as part of its existing decision support tool. This model provides a data-driven, statistically significant method to assess the ability of a reserve location to support a unit's manning requirements in an objective and repeatable manner. The implementation of the logistic regression model will allow the USAR to identify those locations with a high probability of supporting the unit's manning requirements.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to recognize and thank all of the people who supported me throughout this thesis process. Without your encouragement, wisdom and selfless assistance, I would not have been able to complete this work.

First, I would like to thank my family, especially my wife, Mya. Her support, encouragement, understanding, and love have been a key component to any achievement throughout my career, including my time at the Naval Postgraduate School. A big thanks is also due to my kids, Ali and Jake. Even though they did not know what they were doing, their passion for life and unconditional love were a constant, and needed, reminder that life is more than a job. One should always procrastinate on one's thesis to surprise the kids at the beach.

Next, I would like to thank my thesis team, Professor Sam Buttrey, LTC Jonathan Alt and LTC Jeffrey House. Gentlemen, thank you for always being available to answer questions and provide the guidance I needed. Also, thank you for encouraging me to take the more difficult path when you knew it would pay off for me.

I am also indebted to my colleagues from the U.S. Army Recruiting Command G2, the U.S. Army Reserve G1 and the Center for Army Analysis. Without the direct assistance and subject matter expertise from CPT Karey Speten, Joe Baird, MAJ Greg Whelan, Bobbie Anne Austin, LTC David Cloft, and Tucker Hughes, this thesis would not have been possible. Thank you for your patience as I learned about Recruiting and the Army Reserves. Thank you too for the quick response to many emails with what must have seemed like very random questions.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. PURPOSE

The process for selecting suitable locations, referred to as stationing, for United States Army Reserve (USAR) units, known as Troop Program Units (TPUs), is both complex and important. Unlike a regular Army (RA) unit, the geographic location of a TPU will have a direct impact on its ability to meet manning goals and related readiness requirements. The USAR does not have the flexibility to move soldiers to meet manning shortfalls, so it must be able to draw a sufficient number of qualified recruits from the local community (Department of the Army [DA] 2005a, 3). Additionally, the stationing process must take into account availability of training facilities and impacts on overall force structure when determining a location's suitability.

For a TPU to meet the readiness levels required to support its wartime mission, it must be able to meet its manning requirements across all ranks and occupy facilities that support the unit's individual and collective training requirements (DA 2010). As depicted in Figure 1, in recent years the USAR has been able to meet or approach its total authorized end strength. At the same time it has struggled to meet manning goals at the individual TPU level. Figure 2 shows that this has led to some reserve locations being significantly over-strength while others are significantly under-strength.

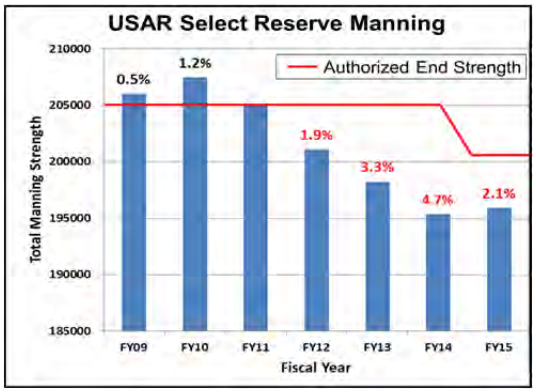


Figure 1. USAR Select Reserve Manning Level FY09-FY15 (after U.S. Army Reserve Command G1 ARIRB Strength Picture Brief dated March 4, 2015)

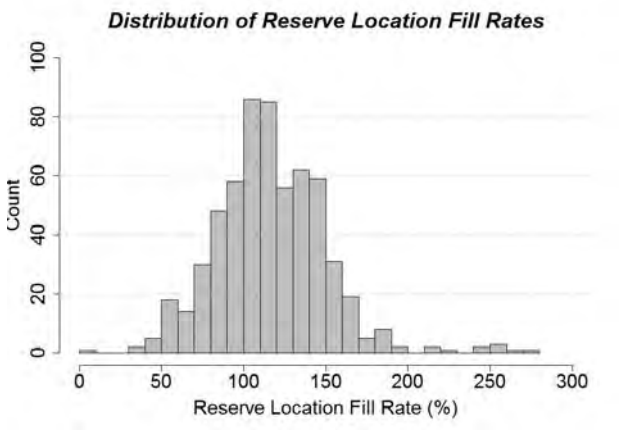


Figure 2. Distribution of Reserve Location Fill Rates

This thesis will focus on the identification of potential stationing locations that have a high probability of supporting the TPU's manning requirements in the Skill Level 1 (SL1) ranks, defined as E-1 through E-4. This represents just one area of concern in the larger stationing problem. Once the USAR identifies the set of locations that are capable of supporting the TPU's manning requirements, it can apply additional criteria to narrow the set to those that meet force structure and training facility requirements. This thesis will not address the criteria for evaluating the force structure and training facility requirements of potential stationing locations. By separating the evaluation of these three broad criteria, decision-makers will be able to more easily identify and quantify the risk associated with the selection of a specific stationing option.

B. PROBLEM STATEMENT

The identification and ranking of feasible stationing options for TPUs is a challenging, multi-attribute decision problem. Since 2008, Stationing Tool Army Reserve (STAR) has been the USAR's primary decision-support tool used in the stationing process. This tool relies on subject matter expert elicited weights to generate an overall utility score based on a location's ability to meet manning, force structure, and facilities requirements (Bradford and Hughes 2007). Current USAR manning data shows that almost 20 percent of USAR locations selected using the current methodology are unable to support the manning requirements of their TPUs (unpublished data). A data-driven

approach must be explored to understand if a stationing methodology informed by a statistical model could perform better.

This thesis will seek to address the following analysis questions:

- Can a model be developed to predict a location's ability to support a USAR TPU's Skill Level 1 manning requirements?
- What factors are the best predictors of a USAR TPU's ability to meet Skill Level 1 manning requirements?
- Is the data currently available within STAR sufficient to develop a useable model of a location's ability to support a TPU's Skill Level 1 manning requirements?

C. MOTIVATION

The Army Reserve is a critical component of the United States' National Defense Strategy. In 2010 the *Quadrennial Defense Review Report* stated that:

> Achieving the defense strategy's objectives requires vibrant National Guard and Reserves that are seamlessly integrated into the broader All-Volunteer Force. Prevailing in today's wars requires a Reserve Component that can serve in an operational capacity—available, trained and equipped for predictable routine deployment. Preventing and deterring conflict will likely necessitate the continued use of some elements of the Reserve Component—especially those that possess high-demand skill sets—in an operation capacity well into the future. (Department of Defense [DOD] 2010, 53)

The Reserve component allows the Army to maintain a ready and trained force that can be activated to meet strategic and operational needs without bearing the cost of maintaining that force in an active duty capacity (Klerman 2009, 13).

In recent years, the USAR has been forced to temporarily augment TPUs that are entering a deployment cycle with reservists from other units to meet the deploying unit's manning requirements. Of the 22 TPUs included in a 2009 Government Accountability Office study, 21 required augmentation from non-deploying units to meet manning requirements for deployment (Pickup 2009, 14). This significant cross-leveling of personnel induces considerable stress in the individual reservists and both the gaining and losing units (Laurent 2005, 28). While less than ideal, the cross-leveling of personnel has at least been sustainable due to the predictable nature of force requirements in sustained

campaigns that allows units to transition through a defined train-up cycle. However, many of the current campaign plans require large numbers of USAR units to be deployed within the first 30 to 45 days of operations, a period that would not allow time for a major cross-leveling of personnel (DOD 2011). If USAR is to continue meeting the readiness requirements of the United States' national defense strategy, TPUs must be located in areas where the recruitable market is able to meet and sustain the unit's manning requirements.

The stationing process, and its impact on TPU manning, is of such significance that the Chief of Staff of the Army issued a tasking to the USAR in February 2014 in which he suggested "perhaps it is not the mission itself, but the location of the Army Reserve units that is the problem [for recruiting]" (Cloft 2014, 1).

D. SCOPE AND STRUCTURE OF THIS THESIS

This research first gathers the data necessary to capture the demographic profile of an area as it relates to a TPU's ability to draw recruits from the local population. This data set will then be used to build multiple regression and classification models in an attempt to develop a model capable of predicting a recruitable market's ability to support the SL1 manning requirements of a proposed TPU.

Chapter II covers the mission, structure and manning challenges of the USAR along with a literature review of past work relevant to this thesis. Chapter III details the data collection process and pre-processing methodologies necessary to develop the model data set. Chapter IV captures the model development process while Chapter V reports the findings and conclusions of this analysis.

II. BACKGROUND

A. MISSION

The United States Army Reserve (USAR) serves as a critical force provider that is available to augment the regular Army (RA). The U.S. Code formally defines the purpose of the USAR:

> To provide trained units and qualified persons available for active duty in the armed forces, in time of war or national emergency, and at such other times as the national security may require, to fill the needs of the armed forces whenever more units and persons are needed than are in the regular components. (2006, Title 10, § 10102)

In the post-Vietnam era, General Abrams directed a robust restructuring of both the active and reserve components of the Army as the United States transitioned from the draft to an all-volunteer military. Under this restructuring, referred to as the Laird-Abrams Doctrine, the USAR assumed ownership of a significant portion of the Army's combat support and combat service support capabilities (Jones 2004). From the early 1970s to early 1990s, the USAR served as a strategic reserve that would only be activated to support a major armed conflict. Following the large activation of USAR elements for the Gulf War, military decision-makers increasing relied on USAR assets to fill operational requirements. This reliance on USAR elements would continue to increase as the United States entered the protracted conflicts in Iraq and Afghanistan.

Since the initial deployments to Afghanistan in 2001, the USAR has deployed over 170,000 soldiers in support of the Global War on Terrorism (USAR, unpublished data). In 2008, Secretary of Defense Robert Gates directed changes to formally redefine the role of the nation's reserve forces, including the USAR, from a strategic reserve to an operational reserve. Today, the USAR supplies 75 percent of key support units and capabilities such as logistics, medical, engineering, military information support, and civil affairs that comprise half of the Army's combat support and combat service support forces. These forces total nearly 20 percent of the Army's total force while using less than 6 percent of the total Army budget (Office of the Chief, Army Reserve [OCAR] 2015).

B. STRUCTURE

The USAR is composed exclusively of individuals who are not assigned to the RA or the Army National Guard. The three major sub-groups within the USAR are the Select Reserve, the Individual Ready Reserve, and the Retired Reserve. The Select Reserve contains those soldiers who are most readily available to respond to activations and mobilizations. This force is further broken down into Troop Program Units (TPUs), Active Guard and Reserve, and individual mobilization augmentees, as depicted in Figure 3.

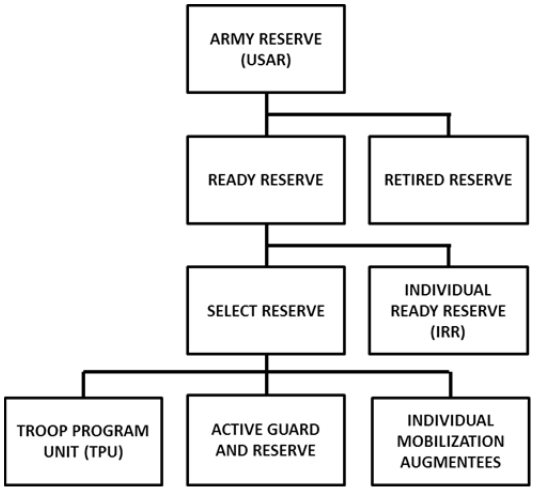


Figure 3. Structure of the Army Reserve

Soldiers assigned to TPUs traditionally train with their assigned unit one weekend per month along with an additional two weeks of annual training during the year. As TPUs form the core of the USAR force structure, they will be the focus of this research. The majority of these TPUs have organizational structures that parallel those found in the RA: platoons, companies, and battalions, along with brigade and higher headquarter elements. The USAR currently has an authorized end-strength for the Select Reserve of 202,000 soldiers who serve in over 3,500 units dispersed across the United States, Puerto Rico, Guam and other overseas locations (USAR, unpublished data). Figure 4 displays the geographic dispersion of those TPUs located in the continental United States.

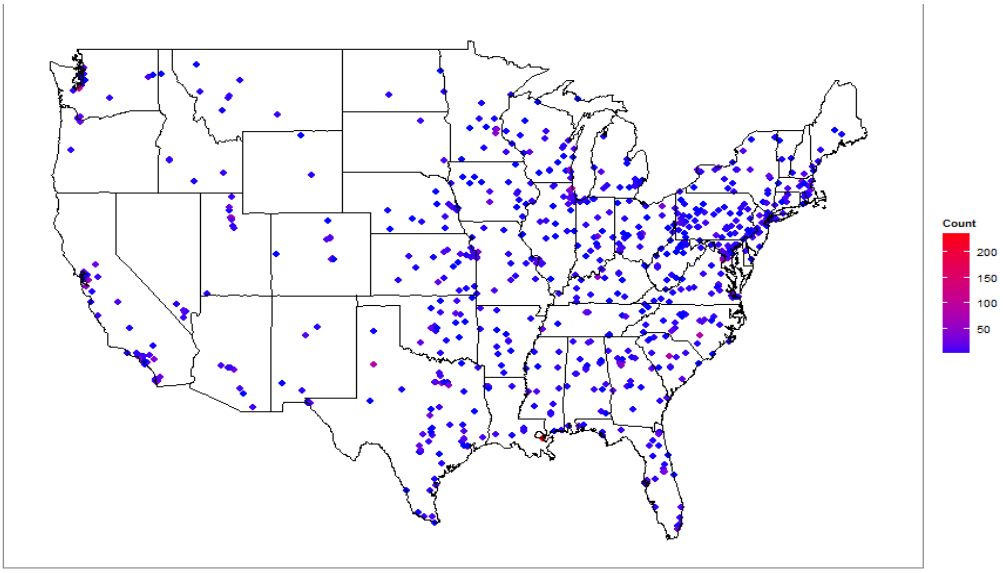


Figure 4. Number of USAR TPUs by ZIP code

C. MANNING

By regulation, members of a TPU must reside within a 50-mile radius or 90-minute drive of the reserve center though individual commanders have the discretion to approve waivers for this requirement (DA 2005a, 3). This geographic restriction on the TPU's market for recruiting directly ties the unit's manning to the population that lives within its immediate vicinity. While the RA draws soldiers from the entire national population and moves them wherever required, within the USAR, each TPU draws the bulk of its soldiers from the local population. This makes the demographics of the local population a critical factor when evaluating a location's ability to support a TPU.

The USAR also differs from the RA in the type of employment that it provides. As a part-time employer, the USAR competes in the secondary labor market while as a full-time employer, the RA competes in the primary labor market. This allows the USAR to attract potential recruits uninterested in an RA enlistment, such as those individuals enrolled in college or other post-secondary education and those establishing a civilian career. Since the USAR is unable to provide full-time employment opportunities, stationing solutions must place TPUs in areas where civilian employers are able to provide sufficient full-time or part-time employment.

D. RECRUITING

The Army is unique as the only Department of Defense (DOD) component that combines its active and reserve recruiting efforts. The U.S. Army Recruiting Command (USAREC) is responsible for all non-prior service recruiting for both the active and reserve components. In Fiscal Year (FY) 2014, USAREC utilized a force of 7,096 RA and 1,356 USAR recruiters to accomplish the recruiting mission. An additional 458 recruiters supported this mission in staff positions throughout the USAREC organization (U.S. Army Recruiting Command [USAREC] 2014). Each year, USAREC receives both an RA and USAR recruiting mission from the Department of the Army. USAREC breaks this overall mission down into assigned missions for each of its five subordinate recruiting brigades, each of which covers a specific geographic region.

The vast majority of non-prior service recruits who enlist in the USAR enter on a 6+2 contract. This contract obligates the future soldier to six years of service in the USAR followed by two years of service in the Individual Ready Reserve. As soldiers enter the end of their initial contract, they have the opportunity to enter into a contract extension (re-enlistment) contingent on their prior performance and Army's continued requirement of their service.

The process by which recruits move from a signed enlistment contract to their first assigned unit differs significantly between the USAR and RA. When future RA soldiers sign enlistment contracts, they enter the Future Soldier Program which acts as a holding pool until the time that they depart for Initial Entry Training (IET). Soldiers do not count against the RA's authorized end-strength until they begin IET (DA 2015, 6). Upon completion of IET, these soldiers are available to fill any vacancy for their military occupational skill (MOS) across the entirety of the RA and do not count against a particular unit's authorizations until they arrive at the unit (DA 2015, 10). In the case of the USAR, soldiers immediately count against both the USAR and the individual TPU's authorized end-strength even though it may be several months before they begin IET and several more months until these soldiers return to a TPU with the training necessary to fill their assigned billet (DA 2005b). The USAR has authorized all TPUs to exceed their

authorized Skill Level 1 (SL1) manning, without limitation, to alleviate the effects of these unqualified soldiers being counted on the rolls of individual units (Talley 2015).

E. READINESS

The Army defines a unit's readiness as "the ability to provide capabilities required by the combatant commanders to execute [its] assigned missions. This is derived from the ability of each unit to deliver the outputs for which it was designated" (DA 2010, 100). To assess each unit's readiness level, the Army looks at four sub-levels: personnel (P-Level), equipment/supplies on-hand (S-Level), equipment readiness/serviceability (R-Level), and unit training (T-Level), each measured on a one to four scale using sub-level specific scoring rules (DA 2010). The assessment of a unit's overall readiness in core missions, its C-Level, uses a combination of all four sub-level scores. A graphical representation of the Army's methodology for overall unit readiness assessments is shown in Figure 5.

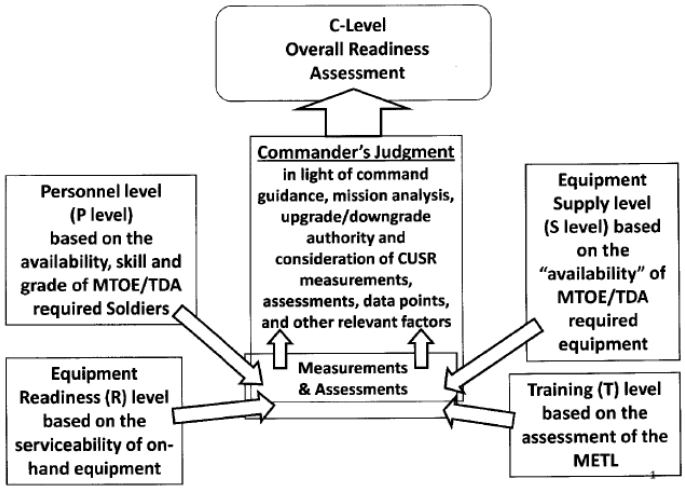


Figure 5. Army Methodology for Readiness Assessment (from DA 2010, 15)

In assessing personnel- (or manning-) related readiness of a unit, three different metrics are assessed:

- Available Strength: The number of soldiers assigned divided by the number authorized.

- Available Duty MOS Qualified (DMOSQ): The number of soldiers holding the correct training for their assigned position divided by the number authorized.
- Available Senior Grade: A measure of the number of senior grade (E-5 and above) authorized positions that are filled (DA 2010, 15).

The lowest of the three metrics determines the unit's P-Level score. Table 1 depicts the parameters for each of the P-Level scoring rules.

Table 1. Metrics for Determining Personnel Levels (from DA 2010, 44)

			Available senior grade	
Level	Available strength	Available DMOSQ	By category	Composite
1	100–90 percent	100–85 percent	100–85 percent	1.54 or less
2	89–80	84–75 percent	84–75 percent	1.55 – 2.44
3	79–70 percent	74–65 percent	74–65 percent	2.45 – 3.34
4	69 percent or less	64 percent or less	64 percent or less	3.35 or more

The manning level of a unit also has an indirect effect on its training (T-Level) score. An undermanned TPU will not be able to complete its mission-essential tasks, resulting in a lower T-Level score. Though not directly assessed in this research, it is worth noting that an undermanned unit will not be able to fulfill its wartime requirement within a reasonable timeframe.

F. UNIT STATIONING PROCESS

The stationing of a new TPU, or the re-stationing of an existing TPU, is a complex process requiring coordination between numerous stakeholders at multiple levels of the USAR command structure. The USAR gives the following as the stated purpose of this stationing process:

> [to] integrate force structure with facilities providing [Operational, Function, Training, and Support] OFTS Commands the best possible overall unit readiness, enhance career progression, increase recruiting, maximize facility utilization, address demographic changes, and provide improved Mission Command. (Colon 2012, 2)

The USAR's "Stationing Memorandum of Instruction" provides the following reasons for the initiation of a stationing action:

- Activations: Initial stationing for a new organization created and approved as a result of Total Army Analysis, Concept Plan or to satisfy Army requirements.
- Split Stationing: Stationing actions originated by an existing TPU's owning command which desires to split the existing TPU between two or more reserve centers.
- Relocation: Initiated by a TPU's owning command to relocate the TPU to a different reserve center. These result from a requirement to improve a TPU's readiness or when known future force structure changes will exceed the current location's capacity.
- Conversions/Reorganizations: Action initiated by the TPU's owning command in response to force structure changes directed by a higher command. (Colon 2012, 13)

The life-cycle of an individual stationing action typically spans 24 to 30 months. In addition to the time required to complete the stationing action, a newly stationed TPU has 36 months until it must meet the unit readiness reporting requirements specified in Army Regulation 220–10 (DA 2010, 20). This five-year lag from the initiation of a stationing action until the time that the TPU must be able to fill wartime requirements makes the accuracy of the stationing process critical to the sustained readiness of the USAR as a whole. Due to current fiscal constraints the USAR expects that most stationing actions will involve placing TPUs into existing reserve locations. By developing a model that predicts a reserve location's ability to meet a TPU's manning requirements this research will support USAR's ability to maintain a manned, trained, and ready force.

G. LITERATURE REVIEW

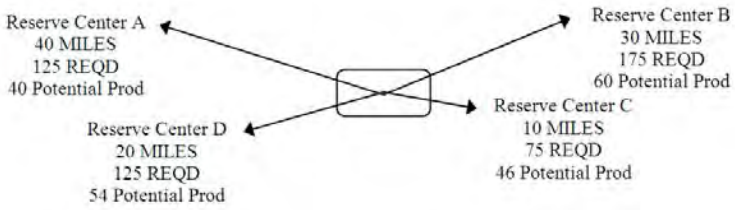
Since General Abram's restructuring of the USAR in the 1970s, the stationing process for TPUs has continued to be an area of active research. How the demographic characteristics of a unit's recruiting market will affect its manning and readiness levels is the unifying theme across these academic and policy studies. A high-level view of the timeline of this research shows that the topic becomes ripe for investigation every five to seven years as both technology and the granularity of demographic data improves. From this large body of research, three primary sources capture the latest methods and

techniques for informing USAR stationing decisions. The following sections will discuss the significant contributions and identified shortcomings of each work.

1. USAREC Market Supportability Study

For more than 25 years, all reserve stationing actions have required a formal market supportability study. The requirement for these studies comes from DOD Directive 1225.7, *Reserve Component Facilities Programs and Unit Stationing*, that directs services to review the manpower potential of an area to determine its adequacy for meeting and maintaining authorized officer and enlisted strengths (Deputy Secretary of Defense 1996). In the early 1990s, USAREC developed the *Market Supportability Study* (MSS) to meet these requirements. At that time, the USAREC G2 was responsible for completing the MSS along with producing a recommendation on whether the proposed USAR stationing action was supportable. In 2007, a portion of the Stationing Tool Army Reserve (STAR) replaced both the MSS methodology and the USAREC review process.

The portion of the MSS that relates to this research is the algorithm by which it allocates portions of a ZIP code's population when it falls within a 90 minute drive of multiple reserve centers. In this algorithm, the distances between the centroid of a population ZIP code and each reserve center within 50 miles, along with the relative sizes of each reserve center, determine the allocation of the population. The MSS algorithm uses a distance factor weighting of .333 and a relative unit size (defined by the number of authorized personnel) weighting of .667. The criteria used in determining these weightings are unclear since the full documentation of the MSS could not be located. Figure 6 depicts the allocation of a single population ZIP code's potential production of 200 soldiers between four reserve centers. In the tabular portion of Figure 6, columns (b), (d), (e), and (g) show the method for calculating the distance ratio while columns (c), (f), and (h) show the method for calculating the size ratio. Column (i) shows the combination of the distance and size ratios to arrive at the adjusted total ratio used to determine the distribution of the population's potential production to each reserve center.



(a) Reserve Center	(b) Distance	(c) Reqd	(d) Distance -50Miles	(e) (d/100) Distance Ratio	(f) (c/500) Reqd Ratio	(g) (e*.333) Weighted Distance Ratio	(h) (f*.667) Weighted Reqd Ratio	(i) (g+h) Adjusted Total Ratio	(j* 800) Allocated Potential Production
A	40	125	10	10%	25%	3.3%	16.7%	20%	40
B	30	175	20	20%	35%	6.7%	23.3%	30%	60
C	10	75	40	40%	15%	13.3%	10%	23.3%	46
D	20	125	30	30%	25%	10%	16.7%	26.7%	54
		500	100						200

Figure 6. MSS Allocation Methodology (from USAREC, unpublished data)

Since the full documentation for the MSS is not available, it is difficult to ascertain the origins and research behind this allocation algorithm. It appears that this algorithm is a refinement of one proposed by Stephen Mehay in a 1989 USAREC Study Report (36–37).

The original implementation of the MSS could only process pre-selected lists of potential stationing sites to determine whether they were supportable or non-supportable. This was likely due to the limited automated data access and computational power available at the time of the MSS's development. The data pre-processing portion of this research will use a variation of the MSS allocation scheme. This variation expands the underlying fundamentals of the MSS methodology by applying it to all population ZIP codes and reserve centers to determine the appropriate allocation.

2. Unit Positioning and Quality Assessment Model

As part of his Naval Postgraduate School thesis, Fair (2004) developed the Unit Positioning and Quality Assessment Model to improve the USAR stationing process. In this work, Fair first constructed a single database capturing demographic statistics at the ZIP-code level. Whereas the MSS used a limited scope of information related to the size and volume of the recruitable market, Fair extended the information available for analysis to include factors related to the population quality and vocation. In the development of

the ZIP-code level demographic database Fair included the following: Bureau of Labor Statistics vocational inclination data groups, the military available population, Microvision 50 lifestyle segmentation categorized by groups, quality of accessions via Armed Forces Qualification Test (AFQT), and the unemployment rate (Fair 2004).

Fair (2004) then developed a linear regression model in which the vocational groups, lifestyle segments, military available population, quality of accessions, and unemployment rate are the independent variables and total USAR production is the dependent variable. This regression model predicts the maximum expected number of USAR recruits a particular population ZIP code can produce annually. Fair also proposed the extension of this model to predict the maximum number of recruits in each population ZIP code who would qualify for specific MOSes. This extension included development of regression models for the top five MOSes in the USAR force structure (Fair 2004). Fair's work does not address the distribution of a population between multiple reserve centers. While the USAR did not incorporate the results of this research directly into its stationing process, the Center for Army Analysis (CAA) team used many of his data source in their study (Bradford and Hughes 2007, C-2).

3. Army Reserve Stationing Study

The Office of the Chief of the Army Reserve (OCAR) identified an urgent need for help with stationing in 2006. At that time, the USAR was in the process of realigning its command structure. This included shifting the bulk of the stationing workload from the regional commands to a centralized function within USAR Force Management staff. At the same time, the USAR expected to expand by 340 TPUs between FY08 and FY13 under the "Grow the Army" and "Army Reserve Rebalancing" initiatives (Bradford and Hughes 2007). In response to this request for assistance, a team of six analysts led by Robert Bradford from the Center for Army Analysis (CAA) completed the year-long *Army Reserve Stationing Study* in 2007. The stated purpose of the Army Reserve Stationing Study (ARSS) project is as follows:

> To develop a unit stationing methodology and tool that considers important factors including: capacity of a local area to recruit and maintain unit personnel, the ability to provide career progression opportunities for

> USAR soldiers, and the location and capacity of existing Reserve facilities. To use this methodology to support stationing decisions for the 340 units associated with Army Growth and Army Reserve Rebalancing. (Bradford and Hughes 2007, iii)

Recognizing that the project centered on complex decisions that included competing objectives, the ARSS team focused on multiple-objective decision analysis as the core of their analysis. The team identified 18 separate measures and developed a value function for each measure. These value functions took the raw measurements and converted them to a scale from 0 to 10. Based on their relative importance, each measure received a weighting that allowed for the generation of an overall value score between 0 and 10 for each metropolitan statistical area (MSA) and existing reserve center. The development of the value functions and comparative weights drew primarily from the input of subject matter experts from the stationing teams within the regional commands. Figures 7 and 8 depict the model hierarchy and measures, and their associated weights, respectively.

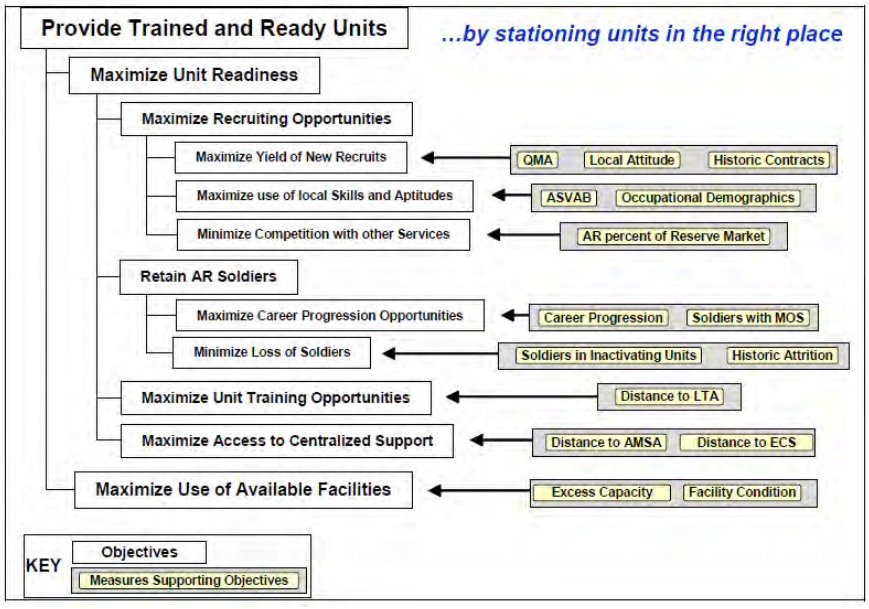


Figure 7. ARSS Objective Hierarchy and Measures (from Bradford and Hughes 2007)

Measures	Global weight	Matrix weight
QMA	0.127	100
Local Attitude	0.064	50
Historic Contracts	0.006	5
ASVAB	0.115	90
Occupational Demographics	0.064	50
Sister Service	0.096	75
Career Progression1	0.038	30
Career Progression2	0.038	30
Career Progression3	0.038	30
Manning	0.115	90
Backfill	0.096	75
Attrition	0.025	20
Distance to LTA	0.025	20
Distance to Major Training Area	0.000	0
Distance to AMSA	0.025	20
Distance to ECS	0.025	20
Facility Capacity	0.064	50
Facility Condition	0.038	30

Figure 8. ARSS Measure Weights (from Bradford and Hughes 2007)

Following the completion of the ARSS, OCAR initiated two follow-on studies through CAA: the *Army Reserve Stationing Study—Phase II* (Hughes 2008) and the *Army Reserve Stationing Portfolio Study* (Hughes 2010). These studies made minor adjustments to the base model and developed extensions to accommodate the use of ARSS products for specialized units such as medial and training units.

The primary input to the ARSS model is the type of unit, by standard requirement code, under consideration for stationing. From this input, the model returns two primary reports. One report includes the value score for all MSAs and the other includes the value scores for each existing reserve center. As an initial recommendation, the CAA team considered any MSA or reserve center in the top third to be supportable, the middle third to be marginally supportable, and the bottom third to be unsupportable. The CAA team also noted that this analysis served only as a starting point for determining the appropriate stationing location for a given TPU and that further detailed analysis would be necessary in the decision-making process.

In 2008, the OCAR also utilized the CAA expertise and methodologies developed during the ARSS series to assist in the developing STAR. This web-based tool automates

the process developed by CAA, allowing USAR analysts to quickly conduct the initial analysis required in a stationing action. An extension of the CAA methodology produces the market supportability studies as required by DOD Directive 1225.7. This change entirely removed USAREC from the USAR stationing process. STAR is now the primary analytic and decision-support tool used by USAR to determine the feasibility and supportability of stationing actions.

The models and products developed and supported by the CAA team in the ARSS series represent a significant improvement to the analysis used in the USAR stationing process. The most significant improvement over previously used analyses is the ability to evaluate the feasibility of all possible stationing locations simultaneously. The model is easy for non-technical decision-makers to understand and represents the priorities of the USAR decision-makers in place at the time of the study's completion. While easy to understand, the use of a multiple-objective value model has the potential to discount weaknesses in an MSA or reserve center that still achieves a supportable score. In some cases, the high-value contributions from facility and career advancement measures may mask weaknesses in a location's ability to generate the necessary number of recruits. By separating the ability of a location's recruiting market to support a TPU's manning requirements from the facility and force structure portions of the stationing problem, this research aims to provide decision-makers a better understanding of the benefits and drawbacks of a stationing decision. Additionally, the use of a data-driven approach in the development of statistically based models enables the assessment of stationing options to be both object and repeatable to a degree not provided by subject matter expert based models.

H. THE WAY AHEAD

Drawing on methods and data sources used in the research detailed above this work develops a model capable of predicting a potential stationing location's ability to meet the proposed TPU's Skill Level 1 manning requirements. The first portion of this work covers the collection of the demographic data necessary to predict a recruiting market's ability to support TPU's manning requirement. The second portion covers

allocating the population data to the appropriate reserve center using an extension of the MSS allocation algorithm. Finally, the predictive model development utilizes classification and regression models in which a reserve center's current manning level is the response variable and the recruiting market demographics are the dependent variables.

III. DATA AND METHODOLOGY

A. DATA COLLECTION

The first step in developing a model to assist the U.S. Army Reserve (USAR) in the stationing process involved gathering the data required. This process included compiling data from many disparate sources, reviewing for obvious errors, formatting into compatible file types, and eventually combining the data into a format usable in a statistical software package. U.S. Army Recruiting Command (USAREC) and USAR provided the bulk of the data for this analysis. The remainder is publicly available from the Bureau of Labor Statistics (BLS), the Center for Disease Control and Prevention (CDC), and the U.S. Census Bureau. The following sections discuss the individual data sets utilized in this research.

1. USAR Unit and Personnel Data Set

This unpublished data set provided by USAR G1 includes information for each Troop Program Unit (TPU) along with each individual allocation, or *line number*, within the Select Reserves. This information was used to determine the number of Skill Level 1 (SL1) authorizations for each unit and whether each SL1 authorization was vacant or filled. Individual manning information for each TPU was then grouped by the units' ZIP code to determine the SL1 manning statistics associated with each recruitable market. These ZIP-code level aggregates will be referred to as reserve locations or reserve ZIP codes.

2. USAR Cohort Data Set

This unpublished data set provided by USAR G1 includes information on all USAR enlisted accessions along with their current characterization of service and assigned unit between Fiscal Year (FY) 2008 and FY2014. This data set was used to determine the number of attritions from each TPU. Attritions were classified to determine the number of soldiers leaving at the end of their service obligations and those separating

due to adverse action. The unit level data was grouped by ZIP code to produce the number of Adverse and Non-Adverse attritions.

3. USAREC Production Data Set

This unpublished data set provided by USAREC G2 includes all enlisted accessions processed by USAREC between FY2011 and FY2014, including each recruit's home of record at time of enlistment, age, Armed Forces Qualification Test (AFQT) score, and component of service (USAR or regular Army [RA]). This data set was used to calculate the annual production rate average AFQT score, for both RA and USAR, by ZIP code.

4. Department of Defense Production

This unpublished data set provided by USAREC G2 includes all enlisted accessions by ZIP code processed by Department of Defense (DOD) entities between FY2011 and FY2014. This data was used to provide insight into the level of competition that USAR faces from other DOD entities when seeking recruits within each ZIP code.

5. USAREC Recruiter Laydown

This unpublished data set provided by USAREC G2 includes information on each recruiter including Army component and the ZIP code of the recruiting center where the recruiter is assigned. This information was used to determine the number of recruiters per component assigned within each ZIP code.

6. Unemployment Rate

This unpublished data set provided by USAREC G2 contains county-level unemployment rates, which are publically released by the BLS. Specifically, this data set uses the U-3 unemployment rate, more commonly known as the official unemployment rate, which measures total unemployed as a percentage of the civilian labor force. Other unemployment measures, such as youth unemployment, which might better represent the unemployment within the USAR primary population for recruiting, were not publically available for geographic areas below the state level. Figure 9 depicts how the U-3

unemployment rate, indicated by the dashed black line, generally tracks these other metrics. As such, it was determined that the U-3 rate was a suitable proxy for the youth unemployment rates.

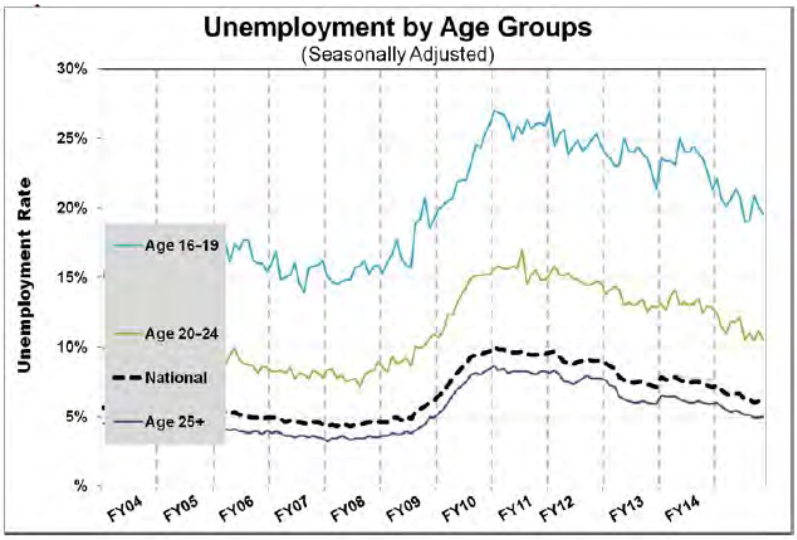


Figure 9. Unemployment by Age Groups for FY04 to FY14 (from USAREC 2014)

7. Obesity Rate

Data for the obesity rate was extracted from "Community Health Status Indicators" survey data set published by the CDC in 2010. The scope of this survey, both in measured statistics and sampled population, varies from year to year depending on the requirements of the CDC. The data from 2010 was the only data set that provided obesity data for the entire United States at the county level.

8. Qualified Military Available Population

USAREC G2 provided the unpublished Qualified Military Available (QMA) population data set. As part of a 2013 study, the Lewin Group developed this data in support of the Joint Advertising, Market Research, and Studies requirement. The Lewin Group used multiple demographic factors, including health, crime, and education, to estimate the number of individuals 17 to 24 years old within each ZIP code who met the medical and moral requirements to enlist in the military. This ZIP-code total was then

broken down into an estimate of how many of those qualified would fall into each of the six Armed Forces Qualification Test (AFQT) categories. The QMA data set excludes those individuals enrolled in post-secondary institutions since it was primarily developed to support active-duty recruiting efforts.

9. Post-Secondary Enrolled Population

The post-secondary enrolled population was derived from the Census Bureau's American Community Survey published in 2014. This survey provides estimates for the number of individuals enrolled in public and private post-secondary institutions by ZIP code. Historically, those individuals pursuing post-secondary education have been viable recruiting markets for the USAR. This data set was included to offset the exclusion of those enrolled in post-secondary institutions from the QMA data set.

10. Regional Location

The regional location of each reserve location was determined by the Army Recruiting Brigade that supports the units within that location. The selection of this classification was influenced by the initial results of research conducted by Marmion (2015) in his research involving recruiter production. Figure 10 depicts the five regional location classifications.

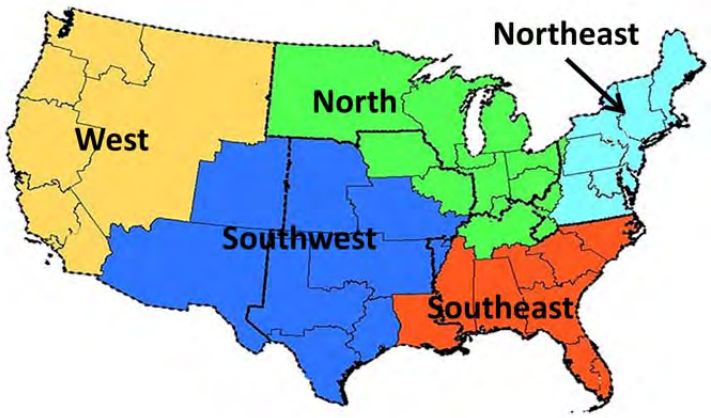


Figure 10. Map of Regional Locations of Reserve Locations (after USAREC, http://www.usarec.army.mil/hq/recruiter/brigade.aspx)

11. Population ZIP Code to Reserve Center Distance/Time Data Set

Prior to this study, the USAREC G2 prepared an unpublished table of distances and drive times from the centroid of each ZIP code containing a USAR reserve center to the centroid of each population ZIP code within either a 50-mile radius or a 90-minute drive.

B. DATA PROCESSING METHODOLOGY

Following the collection and initial cleaning of the individual data sets, several steps were required to construct the final data set the used in the classification and regression models. The following is a list of the individual steps used in preparing the final data set. Additional details for the methods used in steps 2 and 4 appear later in this chapter.

1. For all data sets containing county-level data, the data was translated to ZIP-code level data using a Federal Information Processing Standards (FIPS) to ZIP code crosswalk.
2. For all data sets with observations spanning multiple years, the data was combined to generate a single value for each ZIP code. This was accomplished using the weighted average method.
3. A master population data set was constructed from the individual data sets. This was done by joining the production, recruiter, attrition, unemployment, obesity, QMA, and post-secondary enrolled data sets by ZIP code. The resulting data set contained 17 demographic statistics for each of the 22,680 population ZIP codes within the continental United States.
4. The descriptive statistics for each reserve location were calculated using the population demographics to reserve location allocation method. These descriptive statistics were then joined with each reserve location's manning data to form the data set that was used in the classification and regression models. This master reserve location data set contains 17 demographic statistics for each of the 667 reserve locations.
5. The data set was then examined to identify reserve locations with missing values or other data anomalies. This resulted in the removal of 68 reserve locations from the data set. The majority of these corresponded to locations outside the continental United States such as those located in Hawaii, Alaska, Guam, Puerto Rico and Europe. The final reserve location data set contained the 599 observations used in model development and analysis.

1. Weighted Average Method

The application of a weighted average technique to the multi-year data sets allowed for the representation of data from 2012, 2013, and 2014 in a single data point for each ZIP code. This technique used a 20-percent weighting for the 2012 value, a 30-percent weighting for the 2013 value, and a 50-percent weighting for the 2014 value. As a simplified implementation of exponential smoothing, this combination represents a tradeoff between reducing the impact of cyclical changes in the data while capturing the most relevant portion of any trend in the data (Taha 2007). Equation (1) shows an example of the weighted average formulation.

$$Value_{Average} = .2*Value_{2012} + .3*Value_{2013} + .5*Value_{2014} \quad (1)$$

2. Population Demographics to Reserve Location Allocation Method

Since an individual population ZIP code can be within the recruitable market range (50-mile radius or 90-minute drive) of multiple reserve centers, it was necessary to develop a method to determine the allocation of each population ZIP code to a reserve location. Such a method is imperative to accurately capture the recruitable market available to each reserve location. The allocation method uses the 90-mintue drive metric, instead of the 50-mile radius metric, as it better represents an individual reservist's burden in commuting to a specific reserve center. As depicted in Figure 11, in areas of high TPU concentrations, a single population ZIP code can be within a 90-minute drive of up to 25 different reserve centers.

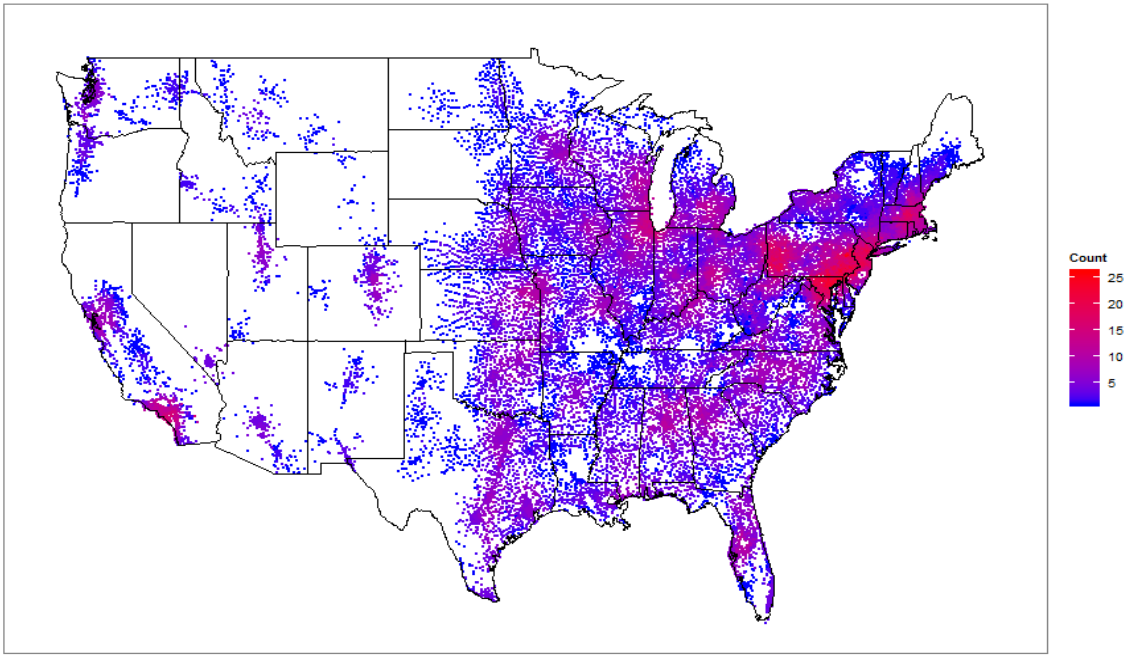


Figure 11. Number of Reserve Centers within a 90-minute drive of each Population ZIP code

Without the application of an allocation method that takes into account the multiple reserve centers drawing from a single population ZIP code, the combined data set will over count the population available to reserve centers in high concentration areas. The fundamentals of the MSS allocation method provided the basis for the allocation method used in this research. By expanding the scope of the MSS method to include all population ZIP codes and reserve centers, it was possible to avoid any over-representation of the recruitable market.

The first step of the allocation method was determining the portion of each population ZIP code allocated to each of the reserve locations that fall within the 90-minute drive. The allocation was determined by two factors: the relative size of the reserve centers, measured in number of SL1 authorizations, and the drive-time from the population ZIP code to the respective reserve locations. The determination of the weightings for the size and distance factors used results from a sensitivity analysis. Table 2 depicts the results of this analysis in which the Adjusted R-squared values from a saturated, first-order linear regression model serves as the measure of performance.

Table 2. Sensitivity Analysis Results for Allocation Method Weightings

Distance Weighting	Size Weighting	Adjusted R-squared
1	0	0.282
0	1	0.221
0.333	0.667	0.274
0.667	0.333	0.290
0.5	0.5	0.290

The results of the sensitivity analysis indicate that the model performance increases as weighting of the distance and size factors approach equality. The 0.5/0.5 weighting scheme was selected based on its performance and simplicity. Figure 12 shows a graphic example of this allocation method. Table 3 shows the supporting calculations for the allocation of a single population ZIP code to four competing reserve locations. The calculation steps involving drive time are highlighted blue, those involving size are highlighted yellow, and the final combination highlighted green.

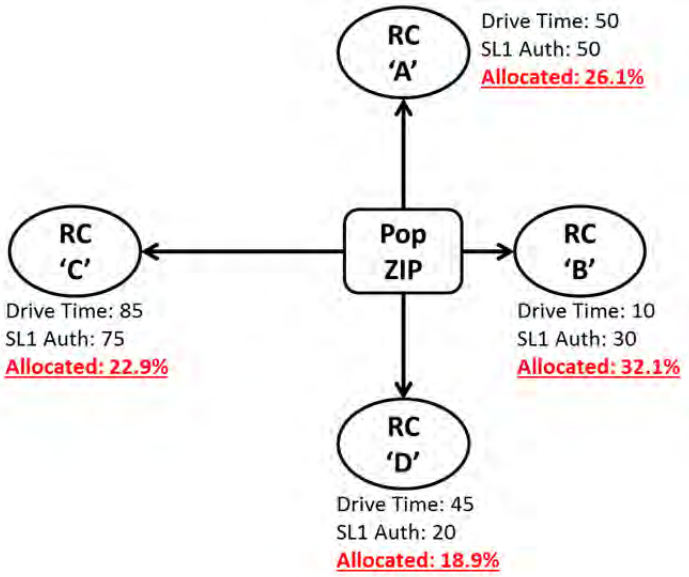


Figure 12. Distribution of a single Population ZIP code between Four Reserve Locations

Table 3. Calculations for the Distribution of a single Population ZIP code between Four Reserve Locations

Reserve Center	Drive Time (DT)	90 - DT	Drive Time Ratio (DTR)	(a) Weighted Drive Time Ratio (DTR * .5)	SL1 Authorized	SL1 Ratio	(b) Weighted SL1 Ratio (SL1R * .5)	Adjusted Total Ratio (a + b)
A	50	40	23.5%	11.8%	50	28.6%	14.3%	**26.1%**
B	10	80	47.1%	23.5%	30	17.1%	8.6%	**32.1%**
C	85	5	2.9%	1.5%	75	42.9%	21.4%	**22.9%**
D	45	45	26.5%	13.2%	20	11.4%	5.7%	**18.9%**
		170			175			

The output from the first step in the allocation method was a 22,680 by 667 data table containing the allocation weighting for all possible population ZIP code to reserve location pairs.

The second step of the allocation method was to calculate the values of the 17 demographic statistics for each of the 667 reserve locations, referred to as the reserve location demographic matrix (RLDM), using the allocation weighting matrix and the master population demographic matrix. This was accomplished using the steps listed below and depicted graphically in Figure 13:

1. Prepare the allocation weighting matrix (AW-M) using the transpose operation to form a 667 by 22,680 matrix.
2. Calculate the initial RLDM (I-RLDM) by multiplying the transposed allocation weighting matrix (AW-M-T) by the master population demographic matrix (M-PD-M).
3. Preform corrective calculations on all normalized demographic factors (unemployment, obesity, AFQT scores and attrition rates) to produce the master RLDM (M-RLDM). Due to the additive nature of matrix multiplication these factors must be divided by the sum of the allocation weighting factors. This divisor is specific to each normalized factor for each reserve location.

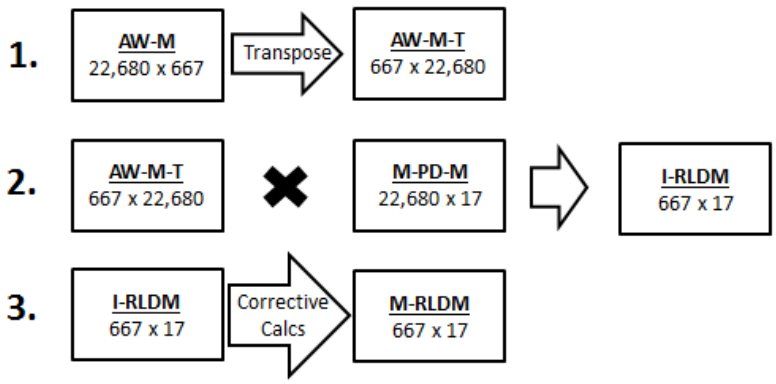


Figure 13. Steps to calculate the Master Reserve Location Demographic Matrix

IV. MODEL DEVELOPMENT AND ANALYSIS

This chapter contains the classification and regression models developed to predict the ability of a reserve location to support the location's manning requirements. The first section contains an analysis of the descriptive statistics of the data set. The subsequent sections discuss the linear regression, classification tree, and logistic regression models, along with the analysis of these models. Descriptive statistic calculations and model developments discussed in this chapter were completed using the R statistical software program (R Core Team 2013).

A. DESCRIPTIVE STATISTICS

This section provides a summary of the descriptive statistics of the 599 observations used during model generation. The descriptive statistics presented below provide the information necessary to understand the range, variance and basic distribution of the data.

1. Dependent Variables

Fill rate serves as the dependent variable in all of the models developed in this research. Figure 14 shows the distribution of the reserve location fill rates. A binary fill rate variable was developed for the classification tree and logistic regression models. Reserve locations with fill rates less than 100 percent were coded as zeros and reserve locations with fill rates greater than or equal to 100 percent were coded as ones. Table 4 shows the number of reserve locations for each classification.

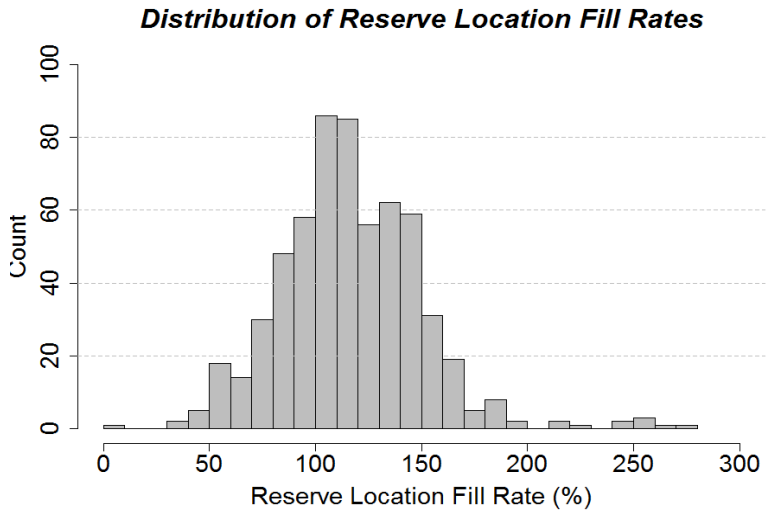


Figure 14. Distribution of Reserve Location Fill Rates

Table 4. Binary Split on Reserve Location Fill Rate

Binary Value	Number of Reserve Locations	Classification Criteria
0	176	Fill Rate < 100%
1	423	Fill Rate ≥ 100%

2. Independent Variables

Tables 5–8 show the descriptive statistics of the independent variables considered by the classification tree and regression models. These statistics provide the information necessary to place the binary splits of the classification tree and the coefficient values of the regression models into context. These statistics show that the population count base factors such as attritions, accessions, and qualified military available (QMA) follow an exponential type distribution while the rate based factors such as obesity, unemployment and Armed Forces Qualification Test (AFQT) scores follow a normal type distribution.

Table 5. Descriptive Statistics of Reserve Location Attrition Data

	Minimum	1st Quartile	Mean	3rd Quartile	Max	Stan. Dev.
Adverse	0.00	2.50	8.12	11.10	53.70	8.42
Non-Adverse	0.00	1.80	5.69	7.50	55.80	6.53

Table 6. Descriptive Statistics of Reserve Location Recruiting Data

	Minimum	1st Quartile	Mean	3rd Quartile	Max	Stan. Dev.
Recruiters	0	3.70	9.85	12.48	81.16	9.74
AR Accessions	1	9.09	23.56	30.34	172.66	21.64
RA Accessions	3.72	32.52	97.70	124.32	764.53	101.13
DOD Accessions	15.93	112.98	251.95	330.75	1633.74	218.0
AFQT	44.89	57.92	59.90	62.16	75.50	3.65

Table 7. Descriptive Statistics of Reserve Location Unemployment and Obesity Data

	Minimum	1st Quartile	Mean	3rd Quartile	Max	Stan. Dev.
Unemployment	3.20	6.04	6.88	7.68	12.70	1.45
Obesity	15.67	21.68	23.53	25.28	31.50	2.81

Table 8. Descriptive Statistics of Reserve Location QMA and Post-secondary Enrollment Data

	Minimum	1st Quartile	Mean	3rd Quartile	Max	Stan. Dev.
QMA I	79.3	730.50	2592.8	3241.0	29138.4	2973.9
QMA II	217.8	1634.9	4630.9	5772.1	43194.0	4507.6
QMA IIIA	110.1	729.6	2021.3	2509.1	18849.9	2066.8
QMA IIIB	131.6	776.8	2180.4	2648.8	20034.2	2260.6
QMA IV	96.1	824.1	2527.8	2909.2	29624.4	3052.8
Post-secondary Enrolled	1690	12599	36839	45979	371950	39702.5

B. MODEL DEVELOPMENT

The primary goal guiding the process of selecting which predictive modeling techniques to employ in this research was to keep the models as simple as possible without sacrificing accuracy. The first technique explored was least-squares linear regression using location fill rate as the dependent variable. The second technique explored a classification tree using the binary split of location fill rate as the dependent variable. The final technique explored was logistic regression that again used the binary

split as the dependent variable. The development of each model and analysis of the results for each technique are discussed in the following sections.

1. Linear Regression Model Development

The simple and well-understood structure of the linear regression model made it a natural choice as an initial modeling technique. This model used location fill rate as the dependent variable. The distribution of the reserve location fill rates is depicted in Figure 14.

In early exploratory linear regression models, it was observed that locations with very small or very large fill rates had a significant effect on the model. Removing them from the model not only had significant effects on model performance but also on those regressors the model deemed statistically significant. To develop a model that best captured the performance of the majority of the reserve locations, the initial data set was reduced to include only those locations whose fill rate was greater than 50 percent and less than 150 percent. Table 9 and Figure 15 depict the removal of those reserve locations that fell outside the specific range and the distribution of the fill rates for the retained reserve locations.

Table 9. Removal of Reserve Locations with Fill Rates <50% and >150%.

	Total	**% of Total**
Original Observations	599	100%
Fill Rate < 50%	8	1.3%
Fill Rate > 150%	73	12.2%
Observations Used	518	86.4%

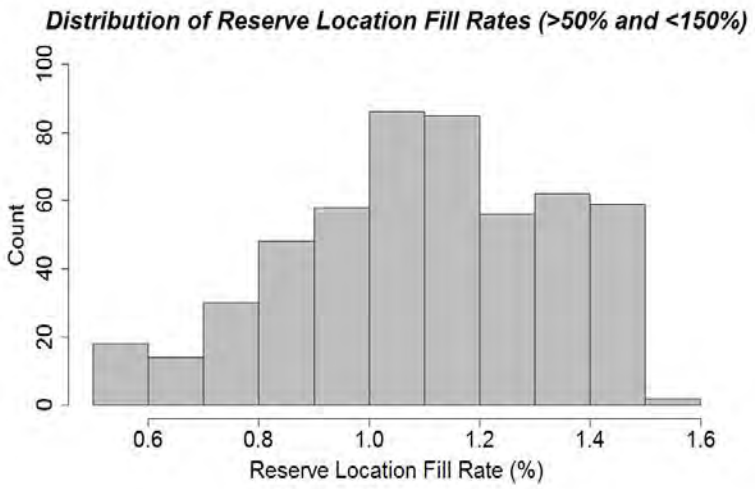


Figure 15. Distribution of Reserve Location Fill Rates (>50% and <150%)

The exploratory models also highlighted the collinearity of several of the independent variables including recruiters, qualified military available (QMA) populations, regular Army (RA) accessions, Army Reserve (AR) accessions, Department of Defense (DOD) accessions, and post-secondary enrollment. Even though these variables are highly collinear, the exploratory models indicated that many of them are statistically significant with p-values below 0.05. However, retaining all of the statistically significant variables would cause problems in accurately estimating the coefficient values, as well as accurately interpreting the model (Faraway 2005, 83). The high collinearity of these variables is understandable since two subsets of them (QMA I-IV and Post-secondary Enrollment) and (Recruiters) are inputs to the third subset (RA, AR and DOD accessions). The decision was made to remove the first two subsets and allow subsequent models to only consider RA, AR, and DOD accessions.

The development of the final linear regression model started with a saturated main effects model and used manual variable deletion to remove variables that had a p-value greater than 0.05. Following the variable reduction process the two attrition variables (Adverse and Non-Adverse) were combined. This produced a slight increase in model performance and produces a simpler model. Table 10 contains the final model, including coefficients and associated p-values. Additionally, Table 11 depicts the goodness-of-fit performance measures of the model.

Table 10. Linear Regression Model Coefficients

	Coefficient	p-value
(Intercept)	1.1230	< 0.001
RQD	-0.0021	< 0.001
Attrition	0.0105	< 0.001
AR Prod	0.0082	< 0.001
Obesity	-0.0077	0.0215
Region NE	0.0304	0.2270
Region SE	0.1298	< 0.001
Region SW	0.0463	0.0712
Region W	0.0205	0.5424

Table 11. Linear Regression Model Goodness-of-Fit Performance Metrics

GOF Metric	Value
Residual Standard Error	.2004
R-squared	.3031
Adjusted R-squared	.2922
Degrees of Freedom	508

Regression diagnostic tests were completed on the final model to determine whether it met the underlying assumptions of a linear regression model. These three tests included constant variance of the errors, normal distribution of the errors, and detection of unusual or overly influential observations (Faraway 2005, 53). The paragraphs below discuss the details of the individual diagnostic tests and their corresponding results.

The contestant variance, or homoscedasticity, assumption test is a visual inspection of the residual versus the fitted values plot of the linear regression model (Faraway 2005, 53–54). Figure 16 depicts the residual versus fitted value plot for the linear regression model. The dashed lines highlight the clipping effect caused by removing the observations with fill rates less than 0.05 and greater than 1.5. Figure 18 shows some evidence of heteroscedasticity, that is, the range of the residuals appears to vary with the fitted values. This minor appearance of heteroscedasticity is not significant enough to discount the model.

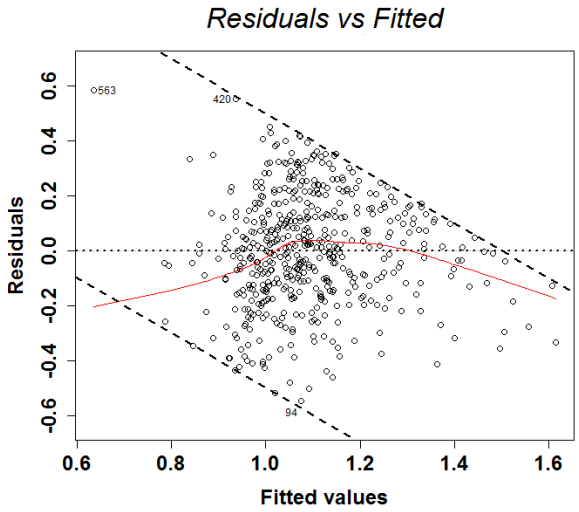


Figure 16. Residuals versus Fitted values plot of the linear regression model

The normal distribution of errors assumption test is a combination of a visual inspection of the model's Q-Q plot and the formal Shapiro-Wilk test (Faraway 2005, 60). Visual inspection of the Q-Q plot, depicted in Figure 17, shows no significant evidence of non-normality in the residuals. The Shapiro-Wilk test returns a p-value of .3686. This p-value indicates that the null hypothesis of the Shapiro-Wilks test, that the residuals are normally distributed, cannot be rejected (Faraway 2005, 60). The results of these two tests indicate that the residuals from the linear regression model can be assumed to be normally distributed.

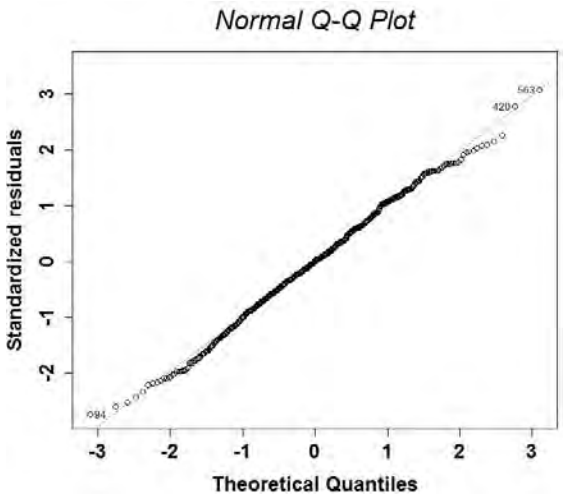


Figure 17. Q-Q Plot of the Linear Regression Model Residual Values

The final diagnostic test is the Cook's distance criterion to identify any overly influential observations (Faraway 2005, 70). Figure 18 shows a plot of Cook's distance for each observation considered by the linear regression model. This plot shows no points that approach the 0.5 value normally considered significant. No changes to the model are necessary based on these results.

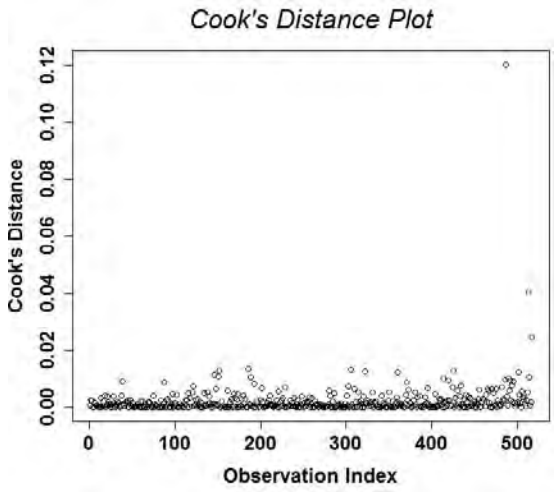


Figure 18. Cook's Distance Plot of the Linear Regression Model

2. Linear Regression Model Analysis

The analysis of the linear regression model focuses on two aspects: the model's goodness of fit, and what the model's structure indicates about a location's ability to meet manning requirements.

As depicted in Table 11, the adjusted R-squared value of the final linear regression model is slightly larger than 0.29. This value can be interpreted to mean that the linear regression model is able to explain 29 percent of the variance in the reserve location fill rates (Faraway 2005, 17). Though adjusted R-squared values in this range are not abnormal for data sets dealing with socio-economic data, the resulting prediction interval widths are too large for this model to be useful in the stationing process.

The structural composition of the linear regression model nonetheless provides some insight into how the demographics factors influence the reserve location's ability to

meet manning requirements. The first aspect of analyzing the structural composition is to look at which variables the model retained along with the magnitude and sign of the corresponding coefficients (see Table 10). The following paragraphs discuss the analysis of each of the variables retained by the linear regression model.

For the *RQD* variable, the model indicates that an increase in the SL1 authorizations is associated with an expected decrease in the fill rate. The magnitude of the *RQD* coefficient also appears plausible, as an increase of 10 SL1 authorizations will reduce the overall manning level by two percent. Here both the direction of effect and the magnitude align with expectations.

For the *Attrition* variable, the model indicates that an increase in the number of attritions is associated with an expected increase in the fill rate. This is not the direction, or sign, of the coefficient expected with variables that reduce a reserve center's manning. Two possible explanations exist for this effect, though there is not enough evidence to arrive at a conclusion. The first explanation comes from the strong positive correlation between the *RQD* and *Attrition* variables. It is likely that, at least to some extent, this correlation is masking the true direction and magnitude of influence for both the *Attrition* and *RQD* variables. Since locations with higher *RQD* values are likely to have higher *Attrition* values, it is possible that the direction of influence between these two variables is reversed. An alternative explanation is that higher levels of *NonAdverse* attrition indicate an area of increased economic activity. This allows current soldiers to find employment that either does not support continued service in the USAR or provides enough income so that service in the USAR is no longer attractive. Additionally, higher levels of *Adverse* attrition could indicate units that are more likely to hold to and enforce standards. While this causes higher *Adverse* attrition levels, the remaining soldiers are more committed to the unit as their *espirit de corps* rises.

For the *Obesity* variable, the model indicates that an increase in the obesity rate is associated with an expected decrease in the fill rate. The overall influence of this variable is relatively small since 50 percent of the observations have *Obesity* values between 21.68 and 25.26, as displayed in Table 7. The direction, or sign, of the coefficient makes

sense because as the population near each reserve location becomes less physically fit, on average, it is less able to support the reserve center's manning requirement.

For the *AR Prod* variable, the model indicates that an increase in the number of Army Reserve accessions is associated with an increase in the fill rate. The direction of influence is as expected since *AR Prod,* or the number of Army Reserve accessions, is how locations receive the recruits necessary to fill vacancies. The magnitude of the variable is also in line with expectations since the average number of SL1 authorizations, for the locations used in this model, is 113. One unit of *AR Prod*, that is one new Army Reserve soldier, increases the average location's fill rate by 0.8 percent, matching the magnitude of the model coefficient.

The coefficients for the four categorical *Region* variables also match expectation. Only the *Region SE* categorical variable is retained at a statistically significant level and is also the region with the largest coefficient value. This is interpreted as meaning that, all other things being equal, a location in the Southeast will, on average, have an almost 13 percent higher fill rate than a location in the North, which is the categorical variable included in the intercept value.

It is also worth noting that *Unemploy* is not retained as statistically significant by the linear regression model. The model suggests that reserve location fill rates are relatively insensitive to the unemployment rate within their local communities. Discussion of the unemployment data in Chapter III showed that the U-2 unemployment rate roughly tracked with the youth unemployment rate, at least at the national level. Future analysis should consider youth unemployment at the local, or ZIP-code level, to determine whether such data would produce a statistically significant effect on reserve location fill rates.

Overall, the linear regression model provides multiple insights into how an area's demographics factors influence the fill rate of the reserve location it supports. While the model performance is likely not at a level necessary for implementation within the USAR stationing process, it does provide a solid basis for further analysis. The full output from the linear regression model is included in Appendix A for reference.

3. Classification Tree Model Development

The classification tree method was selected for its ability to handle binary variables, its ability to perform automatic stepwise variable selection, and the intuitive nature of the completed model (Breiman, et al. 1984, 56-58). The *rpart* package, a software extension for R, used to construct and evaluate the classification tree model primarily implements the methodology developed by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone in the 1984 edition of *Classification and Regression Trees* (Therneau, Atkinson, and Ripley 2013, 1). This method "grows" the tree by attempting to reduce the diversity, or impurity, at each node by selecting the best binary splitting criteria from the set of independent variables. This continues until the tree reaches a set of specified stopping criteria (Therneau and Atkinson 2015, 5). The following classification tree models use the default stopping criteria found in *rpart*. The *rpart* package also conducts a 10-fold cross-validation of the model to provide criteria for "pruning" the tree back from its original size (Therneau, Atkinson, and Ripley 2013, 22).

In the classification tree model developed for this research, the binary split on reserve location fill rate is used as the dependent variable (see Table 4). Prior to constructing the model, the data set was divided into a training set containing 400 observations, and a test set containing 199 observations. Initial exploratory models consider all independent variables listed in Tables 5–8 as candidate variables in node splitting. Further refinement showed that classification tree models using only those variables retained by the linear regression model, minus *Obesity*, produced equal or better levels of accuracy and resulted in less complex models. Figure 19 depicts the complexity parameter (cp) verses cross-validated (X-val) relative error for the original classification tree. For this research, the classification tree was pruned back to a complexity parameter value of .019, resulting in a tree with 8 terminal nodes, or leaves. Figure 20 depicts the resulting classification tree (Milborrow 2015).

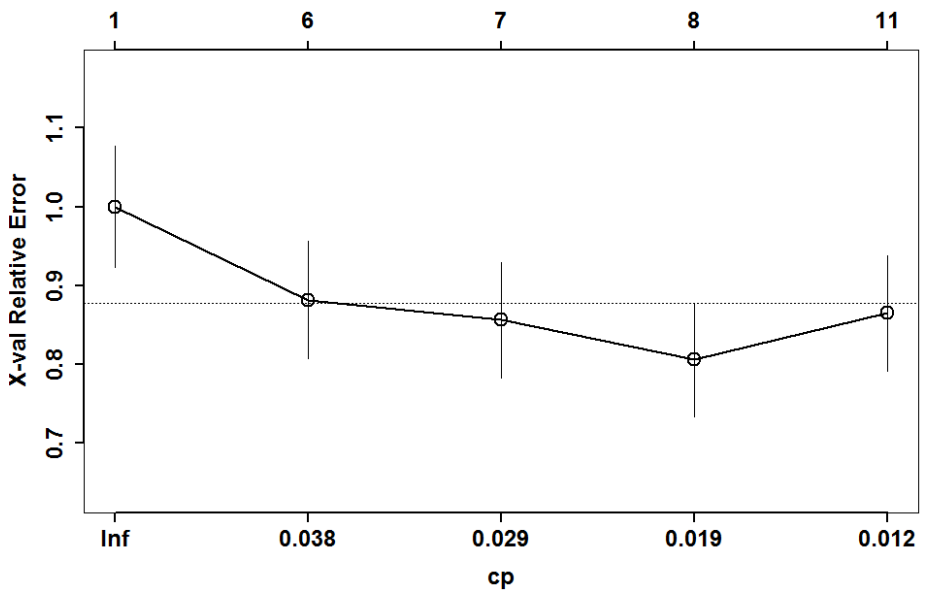


Figure 19. Complexity Parameter versus X-val Relative Error for Classification Tree Model

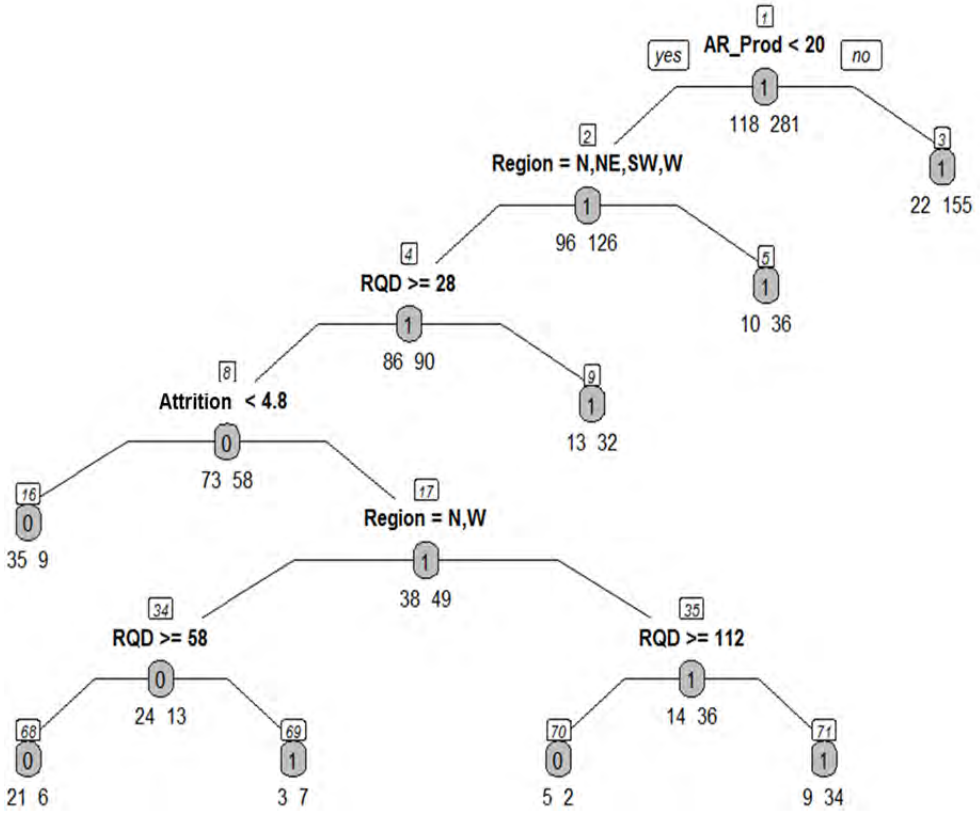


Figure 20. Pruned Classification Tree Model

The annotation of the classification tree depicted in Figure 20 used the following conventions:

- Node Labels—small numbers in boxes above the nodes.
- Node Value—numbers inside the grey boxes. These correspond to the Binary Value (0/1) with the highest number of observations at that node or leaf.
- Splitting Criteria—bold Boolean expressions above the nodes. If the expression evaluates as true, the observation moves down the tree to the left; conversely if the expression evaluates as false, the observation moves down the tree to the right.
- Node Results—numbers below each node representing the count of the Binary Value (0/1) observations at each node.

4. Classification Tree Model Analysis

The observations from the test set were evaluated using the classification tree model from Figure 20. Table 12 displays these results in a standard confusion matrix style.

Table 12. Actual versus Predicted Values for Classification Tree Model

		Predicted Value	
		0	1
Actual Value	0	27	31
	1	16	125

The results in Table 12 show that the classification tree model produced an overall misclassification rate of 23.6 percent. The classification tree model had higher misclassification rate of 53.4 percent on those locations below the 100 percent fill level and a lower misclassification rate of 11.3 percent for those locations above the 100 percent fill level.

In addition to misclassification rate, the receiver operating characteristics (ROC) and Accuracy versus Cutoff plots provides additional information on the classification tree performance. The ROC plot depicted in Figure 21 shows the trade space between the true positive rate and the false positive rate. By varying the cutoff point at which an observation is classified as either a zero or a one a decision-maker can chose to accept

different combinations of true positive and false positive rates. As a reference the confusion matrix for the classification tree displayed in Table 12 used a cutoff value of 0.5. The area under the curve of the ROC plot, which is a standard measure of performance for classification models, produced by the classification tree model is 0.753.

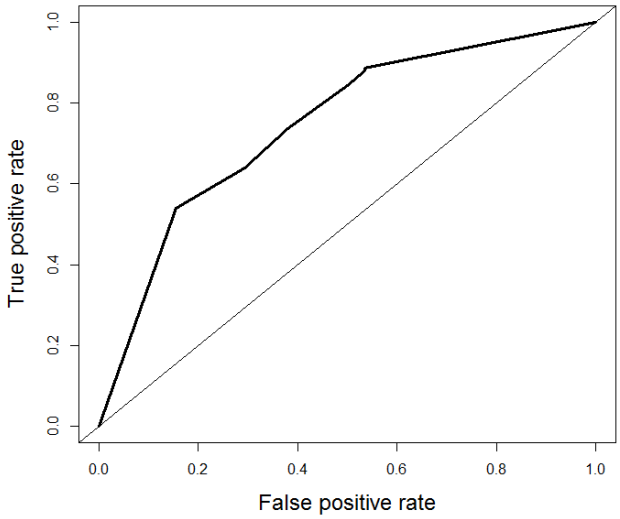


Figure 21. Receiver Operating Characteristics (ROC) Plot for Classification Tree Model

The Accuracy versus Cutoff plot depicted in Figure 22 provides information into how varying the cutoff point will impact the accuracy of the predictions. Figure 22 shows an almost cosistant level of accuracy between a cutoff value of 0.3 and 0.7. This indicates that a decision-maker could alter the cutoff point between these ranges and expect similar levels of accuracy. This indicates that the classification tree model may provide decision-makers with some flexibility in determining their desired true positive versus false positve rates without sacarificing accuracy.

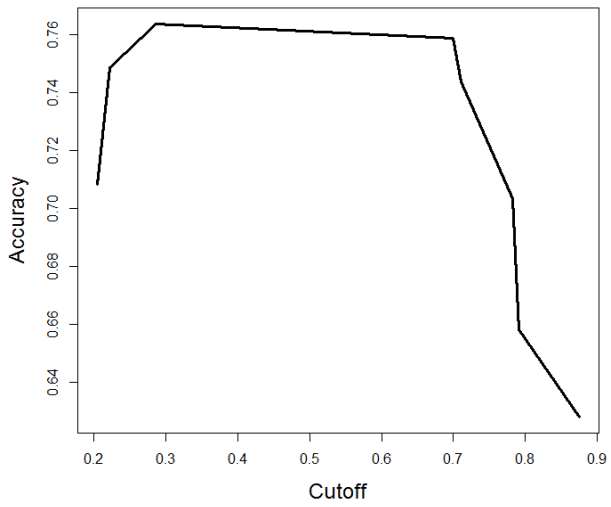


Figure 22. Accuracy vs. Cutoff Plot for the Classification Tree Model

Examination of the classification tree structure, shown in Figure 20, yields further insights. The classification tree model produces results consistent with those from the linear regression model. At Node 1 the model indicates that locations with higher *AR Prod* are more likely to have fill rates above 100 percent. At Node 2 the model indicates that locations in *Region Southeast* are more likely to have fill rates above 100 percent. Additional similarities are observed at the nodes using *Attrition* and *RQD* as the splitting criteria. The full output from the classification tree model is included in Appendix C for reference.

The importance that the classification tree model places on each of the independent variables is another valuable insight. Therneau and Atkinson (2015, 11) developed the variable importance calculation in the *rpart* package as "the sum of the goodness of split measurements for each split for which it was the primary variable, plus goodness times adjusted agreement for all splits in which it was a surrogate." They then scale the variable importance values so that their sum total is equal to 100 (Therneau and Atkinson 2015, 11). This variable importance metric provides insight into the impact each independent variable has on the model regardless of whether it appears as a primary splitting criteria. Table 13 shows the variable importance levels for the classification tree model.

Table 13. Variable Importance for Classification Tree Model

Variable	Importance
RQD	31
AR Prod	27
Attrition	26
Region	15

5. Logistic Regression Model

The binomial logistic regression model also used the binary split on reserve location fill rate as the dependent variable (see Table 4). Additionally, the same training and test sets used in the classification model were used in the construction and testing of the logistic regression model. The logistic regression model used the *glm* function from the base *stats* package included in R.

The logistic regression model development started with a saturated, main effects model. Variables below a p-value of 0.05 were systematically removed from the model starting with those identified in the linear regression model as being highly collinear. The logistic regression model retained both the *NonAdverse* and *Adverse* variables independently, but further analysis indicated that the model produced better performance when these two variables were combined to create a single *Attrition* variable. The final logistic regression model structure was verified using the *step* function from the base *stats* package included in R. Table 14 displays the final logistic regression model, including coefficients and associated p-values.

Table 14. Logistic Regression Model Coefficients

	Coefficient	p-value
(Intercept)	-0.8037	0.0116
RQD	-0.0325	< 0.001
Attrition	0.1506	< 0.001
AR Prod	0.1388	<0.001
Region NE	0.2487	0.4887
Region SE	1.5560	<0.001
Region SW	0.2969	0.4188
Region West	0.4310	0.3965

A overdispersion test was completed to validate the underlying assumption of a binomial distribution by dividing the residual deviance of the model by the degrees of freedom. If the model assumptions are correct then value should be less than or equal to one (Faraway 2006, 45). The overdisperson test for the final logistic regression model produces a value of 0.913 from a residual deviance value of 357.1 on 391 degrees of freedom. The logistic regression model is determined to meet the underlying model assumptions.

6. Logistic Regression Model Analysis

The observations from the test set were evaluated using the logistic regression model from Table 14. Table 15 displays these results in a standard confusion matrix style.

Table 15. Actual versus Predicted Values for the Logistic Regression Model

		Predicted Value	
		0	1
Actual Value	0	39	19
	1	33	108

The results in Table 15 show that the logistic regression model produced an overall misclassification rate of 26.1 percent. The logistic regression confusion matrix was generated using the same cutoff value used for classification tree confusion matrix. The logistic regression model had a higher misclassification rate of 32.7 percent on those locations below the 100 percent fill level and a lower misclassification rate of 23.4 percent for those locations above the 100 percent fill level.

The ROC plot depicted in Figure 23 and the Accuracy versus Cutoff plot depicted in Figure 24 show how the logistic regression model compares to the classification tree model (see Figure 21 and Figure 22). The logistic regression model produces a slightly higher area under the curve of 0.765 and the Accuracy versus Cutoff plot produces a similar range of stability between the cutoff values of 0.3 and 0.7.

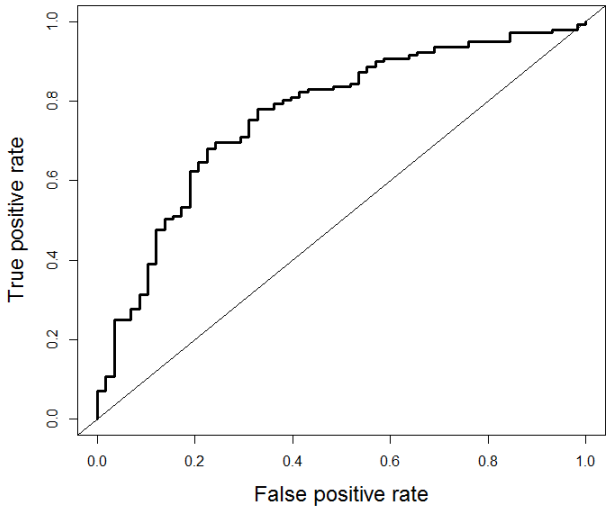


Figure 23. Receiver Operating Characteristics (ROC) Plot for Logistic Regression Model

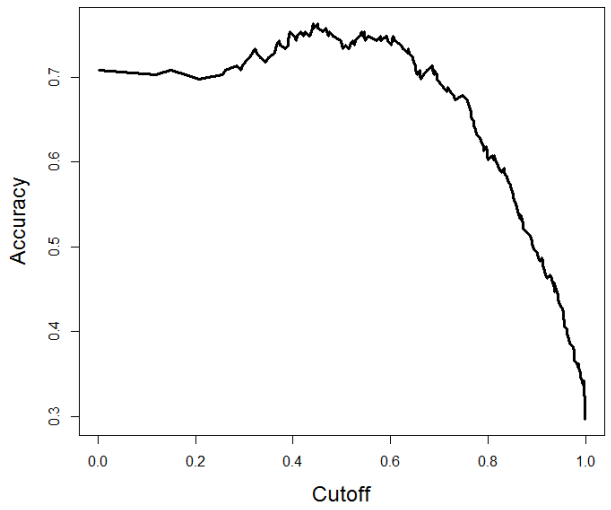


Figure 24. Accuracy vs. Cutoff Plot for the Logistic Regression Model

Analysis of the coefficient values of the logistic regression model (see Table 14) provides insights similar to those gained from the coefficient values of the linear regression model (see Table 10). Like the linear regression model, the logistic regression model indicates that location's with higher *RQD* values are expected to have lower fill

rates. The same similarities hold true for *Attrition*, *AR Prod* and *Region*. The full output from the logistic regression model is included in Appendix C for reference.

C. SUMMARY

All three of the models discussed above provide insight into how the population demographics are likely to influence the ability of the reserve location to meet its manning requirements. The fact that all three models place similar levels of importance on the same independent variables is significant. This similarity indicates robustness in the reserve location demographic factors the impact its ability to support manning requirements. Future analysis can use these factors as a starting point when conducting research into a population's ability to support USAR manning requirements.

Both the classification tree and logistic regression model produce levels of accuracy that will provide valuable recommendations to decision-makers involved in the USAR stationing process. The logistic regression model is judged to be the superior of the two models since it is more likely to correctly classify those locations with fill rates below the 100 percent level.

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY AND RECOMMENDATIONS

This chapter provides a summary of the analytic approach and results discussed in the previous sections along with recommendations and the identification of areas for future research.

A. SUMMARY

The overarching goal of this analysis was the development of data-driven, statistical models that would aid decision-makers in the U.S. Army Reserve (USAR) stationing process. These models were created to assess the ability of a potential stationing location to meet the manning requirements of a Troop Program Unit (TPU) in a repeatable and objective fashion. To accomplish this goal, three questions for analysis were addressed within this thesis:

- Can a model be developed to predict a location's ability to support a USAR TPU's Skill Level 1 manning requirements?
- What factors are the best predictors of a USAR TPU's ability to meet Skill Level 1 manning requirements?
- Is the data currently available within Stationing Tool Army Reserve (STAR) sufficient to develop a useable model of a locations ability to support a TPU's Skill Level 1 manning requirements?

To support the development of statistical models this research aggregated demographic data from eight separate data sets. The final data sets contained 17 demographic factors for each ZIP code within a 90 minute drive of any reserve center. The development of an allocation method allowed these demographic factors to be accurately attributed to the reserve locations thus providing the data set necessary for model development. Finally, three separate models including a linear regression model, a classification tree model, and a logistic regression model were developed to provide USAR stationing decision-makers with the information necessary to make informed stationing decisions.

This thesis has demonstrated that both a classification tree and logistic regression model can predict a location's ability to support a TPU's manning requirements. These models only require four factors: number of Skill Level 1 authorizations, number of

Attritions per year, number of Army Reserve accessions per year, and the regional location. The current USAR stationing decision support tool already contains the data necessary to implement these models. The allocation method, detailed in Figure 11 and Table 3, is necessary to support the implementation of either of these models.

Future analysis in this area should focus on the further refinement of the population to reserve location allocation method and the identification of additional unit and demographic factors that could be affecting manning levels. By addressing these two areas it is possible that higher levels of accuracy can be achieved from the existing model structures.

B. RECOMMENDATIONS

We recommend that USAR implement the logistic regression model as a decision support tool for use in it basing decisions. This model should be used independently to identify the locations most likely to support SL1 manning when units are repositioned. This model provides a data-driven, statistically significant method to assess the ability of a reserve location to support a unit's manning requirements in an objective and repeatable manner. The implementation of the logistic regression model will allow the USAR to identify those locations with a high probability of supporting the unit's manning requirements.

APPENDIX A. LINEAR REGRESSION MODEL

This appendix contains the full linear regression model output produced by the *stats* package. This output provides the ability to observe additional information provided by the model output that was not included in body of the thesis.

```
Call:
lm(formula = Fill ~ RQD + Obesity + AR_Prod + Region + I(Adverse +
    NonAdverse), data = Fill2.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54594 -0.12984  0.00519  0.13426  0.58348

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.1299110  0.0845765  13.360  < 2e-16 ***
RQD                     -0.0020929  0.0002361  -8.863  < 2e-16 ***
Obesity                 -0.0077105  0.0033446  -2.305   0.0215 *
AR_Prod                  0.0082120  0.0009628   8.529  < 2e-16 ***
RegionNE                 0.0304044  0.0251377   1.210   0.2270
RegionSE                 0.1296775  0.0294264   4.407 1.28e-05 ***
RegionSW                 0.0463066  0.0256134   1.808   0.0712 .
RegionW                  0.0205343  0.0336820   0.610   0.5424
I(Adverse + NonAdverse)  0.0104826  0.0015448   6.786 3.23e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2004 on 508 degrees of freedom
Multiple R-squared:  0.3031,  Adjusted R-squared:  0.2922
F-statistic: 27.62 on 8 and 508 DF,  p-value: < 2.2e-16
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. CLASSIFICATION TREE MODEL

This appendix contains the full classification model output produced by the *rpart* package. This output provides the ability to observe potential splits considered, but not utilized, by the model along with surrogate splitting criteria developed by the model.

```
Call:
rpart(formula = Fill ~ RQD + AR_Prod + Region + I(Adverse +
NonAdverse), data = Fill5.train)
  n= 399
          CP nsplit rel error    xerror       xstd
1 0.04237288      0 1.0000000 1.0000000 0.07725484
2 0.03389831      5 0.6864407 0.8813559 0.07431205
3 0.02542373      6 0.6525424 0.8559322 0.07360382
4 0.01800000      7 0.6271186 0.8050847 0.07209915

Variable importance
RQD    AR_Prod      I(Adverse + NonAdverse)   Region
 31        27                 26                15

Node number 1: 399 observations,    complexity param=0.04237288
  predicted class=1  expected loss=0.2957393  P(node) =1
    class counts:   118   281
   probabilities: 0.296 0.704
  left son=2 (222 obs) right son=3 (177 obs)
  Primary splits:
      AR_Prod < 19.96355 to the left,improve=18.701470, (0 missing)
      I(Adverse + NonAdverse)< 5.45 to the left, improve=15.083750, (0
missing)
      RQD    < 127.5  to the left, improve= 5.317794, (0 missing)
      Region splits as LLRLR, improve= 5.003542, (0 missing)
  Surrogate splits:
      RQD    < 98.5     to the left,  agree=0.837, adj=0.633, (0 split)
      I(Adverse + NonAdverse) < 10.45 to the left,agree=0.802,
adj=0.554, (0 split)
      Region splits as LLLLR, agree=0.589, adj=0.073, (0 split)

Node number 2: 222 observations,    complexity param=0.04237288
  predicted class=1  expected loss=0.4324324  P(node) =0.556391
    class counts:    96   126
   probabilities: 0.432 0.568
  left son=4 (176 obs) right son=5 (46 obs)
  Primary splits:
      Region splits as  LLRLL, improve=5.366254, (0 missing)
      RQD < 14.5 to the right, improve=5.139401, (0 missing)
      AR_Prod  < 7.94394  to the left,  improve=4.257245, (0 missing)
      I(Adverse + NonAdverse) < 4.8 to the left,  improve=4.122084, (0
missing)



Node number 3: 177 observations
```

```
  predicted class=1  expected loss=0.1242938  P(node) =0.443609
    class counts:    22   155
   probabilities: 0.124 0.876
Node number 4: 176 observations,    complexity param=0.04237288
  predicted class=1  expected loss=0.4886364  P(node) =0.4411028
    class counts:    86    90
   probabilities: 0.489 0.511
  left son=8 (131 obs) right son=9 (45 obs)
  Primary splits:
      RQD  < 27.5to the right, improve=4.824435, (0 missing)
      I(Adverse + NonAdverse) < 4.8 to the left,  improve=2.873072, (0
missing)
      AR_Prod < 8.216849 to the left, improve=2.685124, (0 missing)
      Region splits as LR-RL, improve=1.091298, (0 missing)
  Surrogate splits:
      I(Adverse + NonAdverse) < 2.75 to the right, agree=0.835,
adj=0.356, (0 split)
      AR_Prod < 4.405236 to the right, agree=0.778,adj=0.133,(0 split)

Node number 5: 46 observations
  predicted class=1  expected loss=0.2173913  P(node) =0.1152882
    class counts:    10    36
   probabilities: 0.217 0.783
Node number 8: 131 observations,    complexity param=0.04237288
  predicted class=0  expected loss=0.4427481  P(node) =0.3283208
    class counts:    73    58
   probabilities: 0.557 0.443
  left son=16 (44 obs) right son=17 (87 obs)
  Primary splits:
      I(Adverse + NonAdverse) < 4.8 to the left, improve=7.518442, (0
missing)
      AR_Prod < 8.216849 to the left, improve=6.511446, (0 missing)
      Region splits as LR-RL, improve=2.199663, (0 missing)
      RQD < 112.5 to the right, improve=1.636337, (0 missing)
  Surrogate splits:
      RQD < 41.5 to the left, agree=0.763, adj=0.295, (0 split)
      AR_Prod < 7.577033 to the left, agree=0.756, adj=0.273, (0 split)

Node number 9: 45 observations
  predicted class=1  expected loss=0.2888889  P(node) =0.112782
    class counts:    13    32
   probabilities: 0.289 0.711
Node number 16: 44 observations
  predicted class=0  expected loss=0.2045455  P(node) =0.1102757
    class counts:    35     9
   probabilities: 0.795 0.205
Node number 17: 87 observations,    complexity param=0.04237288
  predicted class=1  expected loss=0.4367816  P(node) =0.2180451
    class counts:    38    49
   probabilities: 0.437 0.563
  left son=34 (37 obs) right son=35 (50 obs)
  Primary splits:
      Region splits as LR-RL, improve=5.7797330, (0 missing)
      RQD < 59 to the right, improve=4.5977010, (0 missing)
      AR_Prod  < 15.44043 to the left, improve=2.3558800, (0 missing)
```

```
      I(Adverse + NonAdverse) < 10.4 to the right, improve=0.9577609,
(0 missing)
  Surrogate splits:
      RQD  < 77.5 to the right, agree=0.609, adj=0.081, (0 split)
      AR_Prod < 9.925941 to the left, agree=0.609, adj=0.081, (0 split)

Node number 34: 37 observations,    complexity param=0.03389831
  predicted class=0  expected loss=0.3513514  P(node) =0.09273183
    class counts:    24    13
   probabilities: 0.649 0.351
  left son=68 (27 obs) right son=69 (10 obs)
  Primary splits:
      RQD < 57.5     to the right, improve=3.3315320, (0 missing)
      AR_Prod  < 15.58513 to the left, improve=2.5136740, (0 missing)
      I(Adverse + NonAdverse) < 6.85 to the right, improve=0.7848649,
(0 missing)
  Surrogate splits:
      I(Adverse + NonAdverse) < 7.05 to the right, agree=0.865,
adj=0.5, (0 split)
      AR_Prod < 10.71297 to the right, agree=0.784, adj=0.2, (0 split)

Node number 35: 50 observations,    complexity param=0.02542373
  predicted class=1  expected loss=0.28  P(node) =0.1253133
    class counts:    14    36
   probabilities: 0.280 0.720
  left son=70 (7 obs) right son=71 (43 obs)
  Primary splits:
      RQD < 112.5 to the right, improve=3.0702990, (0 missing)
      I(Adverse + NonAdverse) < 10.4 to the right, improve=0.6669444,
(0 missing)
      AR_Prod < 10.81081 to the right, improve=0.4056140, (0 missing)
      Region splits as  -R-L-, improve=0.1600000, (0 missing)

Node number 68: 27 observations
  predicted class=0  expected loss=0.2222222  P(node) =0.06766917
    class counts:    21     6
   probabilities: 0.778 0.222
Node number 69: 10 observations
  predicted class=1  expected loss=0.3  P(node) =0.02506266
    class counts:     3     7
   probabilities: 0.300 0.700
Node number 70: 7 observations
  predicted class=0  expected loss=0.2857143  P(node) =0.01754386
    class counts:     5     2
   probabilities: 0.714 0.286
Node number 71: 43 observations
  predicted class=1  expected loss=0.2093023  P(node) =0.1077694
    class counts:     9    34
   probabilities: 0.209 0.791
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. LOGISTIC REGRESSION MODEL

This appendix contains the full logistic regression model output produced by the *stats* package. This output provides the ability to observe additional information provided by the model output that was not included in body of the thesis.

```
Call:
glm(formula = Fill ~ RQD + AR_Prod + Region + I(Adverse + NonAdverse),
    family = "binomial", data = Fill4.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3847  -0.7358   0.3573   0.7329   2.3224

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.803730   0.318574  -2.523 0.011639 *
RQD                     -0.032465   0.004867  -6.670 2.56e-11 ***
AR_Prod                  0.138769   0.021757   6.378 1.79e-10 ***
RegionNE                 0.248691   0.359242   0.692 0.488771
RegionSE                 1.559562   0.436359   3.574 0.000352 ***
RegionSW                 0.296914   0.367230   0.809 0.418789
RegionW                  0.430989   0.508305   0.848 0.396497
I(Adverse + NonAdverse)  0.150623   0.033502   4.496 6.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 484.55  on 398  degrees of freedom
Residual deviance: 357.17  on 391  degrees of freedom
AIC: 373.17

Number of Fisher Scoring iterations: 6
```

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Bradford, Robert, and Tucker Hughes. 2007. *Army Reserve Stationing Study*. CAA-2006249. Fort Belvoir, VA: Center for Army Analysis.

Breiman, Leo, Jerome H. Freidman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Belmont, CA: Wadsworth, Inc.

Center for Disease Control and Prevention. 2010. Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer Database. Object name RISKFACTORSANDACCESSTOCARE. Accessed March 9, 2015. https://www.healthdata.gov/data/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer.

Cloft, David. 2014. *USAR Recruiting Strategy Review*. Tasker: 140277680, CSA: 38. Washington, DC: Office of the Chief, Army Reserve.

Colon, Pedro J. Memorandum for All Army Reserve Units sent July 18, 2012. "Army Reserve Stationing Memorandum of Instruction." Fort Bragg, NC: U.S. Army Reserve Command.

Department of the Army. 2005a. *Assignments, Attachments, Details, and Transfers* (Army Regulation 140–10). Washington, DC: Government Printing Office.

———. 2005b. *Service Obligations, Methods of Fulfillment, Participation Requirements, and Enforcement Procedures* (Army Regulation 135-91). Washington, DC: Government Printing Office.

———. 2010. *Army Unit Status Reporting and Force Registrations* (Army Regulation 220–1). Washington, DC: Government Printing Office.

———. 2015. *Personnel Accounting and Strength Reporting* (Army Regulation 600-8-6). Washington, DC: Government Printing Office.

Department of Defense. 2010. *Quadrennial Defense Review*. Washington, DC: Government Printing Office.

———. 2011. *Comprehensive Review of the Future Role of the Reserve Component: Volume I Executive Summary & Main Report*. Washington, DC: Office of Vice Chairman of the Joint Chiefs of Staff and Office of Assistant Secretary of Defense for Reserve Affairs.

Deputy Secretary of Defense. 1996. *Reserve Component Facilities Programs and Unit Stationing*. DOD Directive 1225.7. Washington, DC: Deputy Secretary of Defense, March 18.

Fair, Martin L. 2004. "Geo-demographic Analysis In Support of the United States Army Reserve (USAR) Unit Positioning and Quality Assessment Model (UPQUAM)." Master's thesis, Naval Postgraduate School.

Faraway, Julian J. 2005. *Linear Models with R.* Boca Raton, FL: Chapman & Hall/CRC.

———. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Boca Raton, FL: Chapman & Hall/CRC.

Hughes, Tucker. 2008. *Army Reserve Stationing Study—Phase II*. CAA-2007177. Fort Belvoir, VA: Center for Army Analysis.

———. 2010. *Army Reserve Portfolio Stationing Study*. CAA-2008162. Fort Belvoir, VA: Center for Army Analysis.

Jones, Brian D. 2004. "The Abrams Doctrine: Total Force Foundation or Enduring Fallacy?" Strategic Research Project. Carlisle Barracks, PA: U.S. Army War College.

Klerman, Jacob A. 2009. *Rethinking the Reserves*. Santa Monica, CA: RAND Corporation.

Laurent, Janet S. 2005. *An Integrated Plan Is Needed to Address Army Reserve Personnel and Equipment Shortages*. GAO-05–660. Washington, DC: U.S. Government Accountability Office.

Marmion, William N. 2015. "Evaluating and Improving the SAMA (Segmentation Analysis and Market Assessment) Recruiting Model." Master's thesis, Naval Postgraduate School.

Mehay, Stephen L. 1989. *An Enlistment Supply and Forecasting Model for the U.S. Army Reserve.* USAREC Study Report 89–2. Monterey, CA: Naval Postgraduate School.

Milborrow, Stephen. 2015. rpart.plot: Plot rpart Models. An Enhanced Version of plot.rpart. R package version 1.5.2. http://CRAN.R-project.org/package=rpart.plot.

Office of the Chief of Army Reserves. 2015. *U.S. Army Reserve At A Glance: Twice The Citizen.* Washington, DC: Army Reserve Communications. Accessed February 10. http://www.usar.army.mil/resources/Media/ARAG%20V3%20Final%20PDF.pdf.

Pickup, Sharon L. 2009. *Army Needs to Reevaluate Its Approach to Training and Mobilizing Reserve Component Forces.* GAO-09–720. Washington, DC: U.S. Government Accountability Office.

R Core Team. 2013. R: A language and environment for statistical computation. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org

Taha, Hamdy A. 2007. *Operations Research: An Introduction.* Upper Saddle River, NJ: Pearson Education, Inc.

Talley, Jefferey W. Memorandum for All Army Reserve Units sent January 26, 2015. "Army Reserve (AR) Troup Program Unit (TPU) Manning Guidance." Washington, DC: Office of the Chief, Army Reserve.

Therneau, Terry M., and Elizabeth Atkinson. 2015. "An Introduction to Recursive Partitioning Using the RPART Routines." R-project.org. February 24. http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Therneau, Terry M., Elizabeth Atkinson, and Brian Ripley. 2013. rpart: Recursive Partitioning. R package version 4.1-3. http://CRAN.R-project.org/package=rpart.

U.S. Army Recruiting Command. 2014. *U.S. Army Recruiting Command's Fiscal Year 14 End of Year Talking Points.* Fort Knox, KY.

U.S. Census Bureau 2009–2013 5-Year American Community Survey. 2014. Object name B14004. Accessed March 9, 2015. http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_13_5YR_B14004&prodType=table.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California