NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DETERMINING MARKET CATEGORIZATION OF UNITED STATES ZIP CODES FOR PURPOSES OF ARMY RECRUITING

by

Brandon M. Fulton

June 2016

Thesis Advisor:	Lyn R. Whitaker
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 2016	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE DETERMINING MARKET CATEGORIZATION OF UNITED STATES ZIP CODES FOR PURPOSES OF ARMY RECRUITING	**5. FUNDING NUMBERS**
6. AUTHOR(S) Brandon M. Fulton	
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)

The Army relies on Zone Improvement Plan (ZIP) codes to assign recruiters and to track recruit production. ZIP codes have different densities of potential recruits; the Army uses commercial market segmentation data to analyze markets and past accessions to assign recruiters and quotas to maximize production. We use 347 variables from publicly available United States government agencies for each of 34,007 ZIP codes to cluster ZIP codes into similar groups.

We use between 2 and 18 clusters for each of five categories of data, using three dissimilarity calculation methods, and three clustering algorithms. Using national recruiting leads as a proxy for market potential, we find the best cluster assignment by fitting Poisson regressions predicting leads from ZIP code cluster membership.

Economic cluster assignments predict leads with a pseudo R-squared value of 0.69, reducing the need for United States Army Recruiting Command to rely on proprietary data with 66 market segments per ZIP code for market analysis and predicting recruiting potential. These 18 clusters provide an easier tool for recruiting commanders. Additionally, these clusters offer a new method of identifying potentially high-production ZIP codes without using previous accessions and the highly correlated number of recruiters assigned as predictor variables.

14. SUBJECT TERMS recruiting, tree clusters, unsupervised	**15. NUMBER OF PAGES** 95
	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

DETERMINING MARKET CATEGORIZATION OF UNITED STATES ZIP CODES FOR PURPOSES OF ARMY RECRUITING

Brandon M. Fulton
Major, 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 2016

Approved by: Lyn R. Whitaker
Thesis Advisor

Jeffrey B. House
Second Reader

Patricia A. Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Army relies on Zone Improvement Plan (ZIP) codes to assign recruiters and to track recruit production. ZIP codes have different densities of potential recruits; the Army uses commercial market segmentation data to analyze markets and past accessions to assign recruiters and quotas to maximize production. We use 347 variables from publicly available United States government agencies for each of 34,007 ZIP codes to cluster ZIP codes into similar groups.

We use between 2 and 18 clusters for each of five categories of data, using three dissimilarity calculation methods, and three clustering algorithms. Using national recruiting leads as a proxy for market potential, we find the best cluster assignment by fitting Poisson regressions predicting leads from ZIP code cluster membership.

Economic cluster assignments predict leads with a pseudo R-squared value of 0.69, reducing the need for United States Army Recruiting Command to rely on proprietary data with 66 market segments per ZIP code for market analysis and predicting recruiting potential. These 18 clusters provide an easier tool for recruiting commanders. Additionally, these clusters offer a new method of identifying potentially high-production ZIP codes without using previous accessions and the highly correlated number of recruiters assigned as predictor variables.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION....1
A. PROBLEM DESCRIPTION....1
B. PURPOSE....1
C. RESEARCH QUESTIONS....2
D. SCOPE OF STUDY....3
1. Constraints....3
2. Limitations....3
3. Assumptions....4
E. STUDY OUTLINE....5

II. BACKGROUND....7
A. RECRUITING CONSIDERATIONS....7
1. Recruiting Boundaries....7
2. Recruiting Personnel....8
3. Recruiting Mission....8
B. EXISTING GEOGRAPHIC CLASSIFICATIONS....9
1. ZIP Codes....9
2. Recruiting Regions and Boundaries....10
3. Counties....11
4. U.S. Census Bureau....11
C. PAST ANALYSIS OF FACTORS INFLUENCING RECRUITING....12
1. Studies of Market Characteristics....14
2. Studies of Recruiting Incentives and Efforts....17

III. DATA SOURCES....19
A. COMMUNITY HEALTH STATUS INDICATORS....19
B. INDIVIDUAL INCOME TAX RETURNS....20
C. ECONOMIC CENSUS....21
D. COUNTY BUSINESS PATTERNS....21
E. MILITARY BASES....22
F. VETERAN POPULATIONS....23
G. SECONDARY EDUCATION INSTITUTIONS....23
H. NATIONAL LEADS....24

IV. CLUSTERING AND ANALYSIS....25
A. CLUSTER BUILDING....25

1. Variables 26
2. Dissimilarity Calculation 27
3. Clustering Algorithm 27
4. Number of Clusters 28
B. ASSESSING CLUSTERS 29
C. COMPARING TREECLUST PARAMETERS 30
1. Comparing Clustering Algorithms 31
2. Comparing Dissimilarity Calculation Methods 33
3. Comparing K-Values 35
D. COMPARING CLUSTERS TO PRIZM NE 38
E. COMBINING MODELS 40

V. CONCLUSION AND RECOMMENDATIONS 41
A. SUMMARY OF RESULTS 41
B. RECOMMENDATIONS FOR FUTURE WORK 42
C. CONCLUSION 44

APPENDIX A. VARIABLES USED 45

APPENDIX B. DATASET TRANSFORMATIONS 59
A. COMMUNITY HEALTH STATUS INDICATORS 59
B. INDIVIDUAL INCOME TAX RETURNS 61
C. ECONOMIC CENSUS AND COUNTY BUSINESS PATTERNS 61
D. MILITARY BASES 63
E. VETERAN POPULATIONS 64
F. SECONDARY EDUCATION INSTITUTIONS 64

LIST OF REFERENCES 67

INITIAL DISTRIBUTION LIST 73

LIST OF FIGURES

Figure 1. United States Army Recruiting Command Boundaries. Source: United States Army Recruiting Command (n.d.). 8

Figure 2. Metropolitan and Micropolitan Statistical Areas of the United States and Puerto Rico. Source: United States Census Bureau (2013b). 12

Figure 3. Recruiting Year 2013 Penetration Rates. Source: United States Army Recruiting Command (2013). 13

Figure 4. Nielsen Company's Potential Rating Index for ZIP Markets, New Evolution (PRIZM NE) Market Segmentation. Source: Nielsen Company (2013). 16

Figure 5. Computation Time by Number of Clusters for All Data Categories. 31

Figure 6. Average Distances from Medoids to Observations Assigned to Clusters for All Data Categories. 32

Figure 7. Pseudo R-squared Values by Number of Clusters from Poisson Generalized Linear Models. 33

Figure 8. Two-Dimensional Plots of ZIP Codes Using Education Data. 35

Figure 9. Average Distance by Number of Clusters for Demographic Data. 36

Figure 10. Total Residual Sum of Squares by Number of Clusters for Demographic Data. 37

Figure 11. Pseudo R-squared Values for Generalized Linear Models with PRIZM NE Segments and *treeClust* Cluster Assignments. 39

Figure 12. Map Showing Cluster Assignments with Six Clusters. 42

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. National Leads by Year from 2011—2014 (USAREC 2015a). 30

Table 2. Demographic Data Variable Names, Sources, and Descriptions..... 45

Table 3. Economic Data Variable Names, Sources, and Descriptions. 45

Table 4. Education Data Variable Names, Sources, and Descriptions. 50

Table 5. Health Data Variable Names, Sources, and Descriptions..... 51

Table 6. Military Data Variable Names, Sources, and Descriptions. 55

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AD	Active Duty
AR	Army Reserve
ACS	Army Custom Segments
CBSA	Core Based Statistical Area
CDC	Centers for Disease Control and Prevention
CHSI	Community Health Status Indicators
CONUS	contiguous United States
GLM	generalized linear model
HUD	Department of Housing and Urban Development
IRS	Internal Revenue Service
JAMRS	Joint Advertising, Market Research & Studies Group
OCONUS	outside the contiguous United States
PRIZM NE	Potential Rating Index for ZIP Markets New Evolution
QMA	qualified military available
RA	regular Army
SAMA	segmentation analysis and market assessment
USAREC	United States Army Recruiting Command
USPS	United States Postal Service
ZIP	Zone Improvement Plan

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

From 2007 to 2016, the United States Army Recruiting Command (USAREC) has been responsible for recruiting an average of 66,900 Active Duty enlisted soldiers per year (United States Army Recruiting Command 2016). To assist its recruiters in finding, attracting and signing enough youth to fill this positions, USAREC uses the Potential Rating Index for ZIP Markets New Evolution (PRIZM NE) market segmentation produced by the Nielsen Company (Joint Advertising, Market Research & Studies Group 2005). This market segmentation is used for market analysis and for tracking market penetration but it is not currently used in recruiter assignment and goal allocation models (Fleischmann and Nelson 2014). USAREC requires a market segmentation tool that will allow commanders to easily understand the different types of markets within their areas of responsibility.

The PRIZM NE estimates categorize each market into 66 different segments. These segments are based on different socio-economic and age categories (Nielsen Company 2016). Sixty-six segments for each Zone Improvement Plan (ZIP) code are too many for a company or battalion commander trying to understand their operational environment. With an average of 174 ZIP codes per company, commanders need a method to segment the ZIP codes themselves so they can more easily compare their potential for producing recruits (United States Army Recruiting Command 2015b).

USAREC assigns recruiters and goals to its recruiting stations based on a four year weighted average of past accessions and the number of qualified military available estimated to be in each ZIP code (Fleischmann and Nelson 2014). These goals are adjusted using a Recruiting Market Index that takes into account factors such as unemployment, the propensity for military service of an area, and economic factors (Fleischmann and Nelson 2014). USAREC does not currently use market segmentation to inform its recruiter assignment and goal allocation models.

In this study, we use 347 publicly available variables to cluster ZIP codes into groups along with those that have similar potential to produce recruits. We do this by

categorizing the data into demographics, economic activities, education opportunity, military influence, and health status. We then cluster the ZIP codes separately by category using the Tree Clustering algorithm developed by Buttrey and Whitaker (2015). We create 153 different cluster models for each category using different combinations of tuning parameters and numbers of final clusters between 2 and 18. We find that the economic data outperforms the other data categories and provides the highest predictive power with 18 clusters.

To test the predictive ability of our ZIP code clusters, we use the cluster assignments they provide to fit Poisson regression models using the number of national leads from 2011–2013 as a response variable. National leads are generated by potential recruits who seek out more information about joining the Army through various means, such as the Army's recruiting website (United States Army Recruiting Command 2015a). We use this metric because it is less influenced by the efforts of recruiters than other metrics, such as the number of past accessions or local leads (Gibson, Hermida, Luchman, Griepentrog, and Marsh 2011).

We then fit a Poisson model using percentages of PRIZM NE segments in each ZIP code as predictors. We find that the cluster assignments from the economic data explain 69% of the deviance in national leads, while the PRIZM NE segments only explain 60% of the deviance. We then attempt to predict the number of 2014 national leads using these two models. We find that the economic cluster assignments result in a median difference between the actual and predicted leads of 1.9. This outperforms the PRIZM NE models which have a median difference of 3.4.

The cluster assignments using the economic activity data allow us to replace the 66 PRIZM NE market segments per ZIP code with a single cluster assignment. This allows commanders to better understand the differences in the ZIP codes within their operating environment. With a higher predictive power of national leads compared to the PRIZM NE data, the cluster assignments could also be used to inform recruiter assignment and goal allocation models.

LIST OF REFERENCES

Buttrey S, Whitaker L (2015) treeClust: An R package for tree-based clustering dissimilarities. *The R Journal* 7(2) (December). Retrieved April 8, 2016, https://journal.r-project.org/archive/2015-2/buttrey-whitaker.pdf.

Fleischmann M, Nelson M (2014) Recruiting mission allocation using a recruiting market index. Presentation, Army Operations Research Symposium, November 4, Aberdeen Proving Ground, Maryland.

Gibson J, Hermida R, Luchman J, Griepentrog B, Marsh S (2011) ZIP code valuation study technical report. Joint Advertising, Market Research & Studies (JAMRS), Defense Human Resources Activity, Arlington, Virginia.

Joint Advertising, Market Research & Studies Group (2005) A national segmentation analysis of the joint services (FY00-FY05). Unpublished presentation.

United States Army Recruiting Command (2015a) National Leads Data. Unpublished dataset.

United States Army Recruiting Command (2015b) Recruiting station identification to ZIP code crosswalk. Unpublished dataset.

United States Army Recruiting Command (2016) Frequently asked questions about recruiting. Retrieved May 14, 2016, http://www.usarec.army.mil/support/faqs.htm.

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 thesis advisor and second reader, Dr. Lyn Whitaker and LTC Jeffrey House, for all of their advice, mentorship, and assistance on this project. Without you, I would not have been able to pull it off. Second, I would also like to thank Dr. Sam Buttrey and LTC Jonathan Alt; although you were not on my official advisory team, you have been contributed greatly to this process.

I would like to thank the Army (and Navy) crew that I have worked with and explored California with over the past two years. You have made this challenging program a lot of fun. Finally, I would like to thank my wife for supporting me through our time apart.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. PROBLEM DESCRIPTION

The United States Army Recruiting Command (USAREC) is responsible for all Regular Army (RA) and Army Reserve (AR) recruiting. Each year, the leadership of USAREC must determine where to assign its recruiters and how to allocate the total annual mission—the Army's term for a quota—to each target area (United States Army Recruiting Command 2013 p. 5–4). As military budgets decrease, it is becoming more important for military leaders to efficiently use the nation's resources. At the same time, the recovering economy is making it more difficult to gain new, highly qualified recruits (Vergun 2015). This means USAREC must effectively assign and task the Army's recruiters. In fiscal year 2012, USAREC spent an average of $22,300 on each recruit before they arrived at their initial training, not including the amount of money spent to lease recruiting offices throughout the country (United States Army Recruiting Command 2016).

USAREC is primarily made up of 5 brigades, responsible for 44 battalions, 262 recruiting companies and 948 recruiting stations spread across the world (United States Army Recruiting Command 2016). Each recruiting station is assigned several recruiters and a set of Zone Improvement Plan (ZIP) codes from which to recruit. Currently, USAREC uses a number of resources to inform its recruiting strategy, including Nielsen Company market segmentation, studies of population, and records of past accessions—the Army's term for enlistees who arrive at training. One resource USAREC lacks is a means of performing analysis at the ZIP code level to determine which ZIP codes are similar and therefore may have similar production rates and respond to similar recruiting strategies. This is important because USAREC uses ZIP codes, not counties, states or cities, to assign territories to the recruiters.

B. PURPOSE

The purpose of this study is to use publicly available data about United States ZIP codes to determine which geographically dispersed ZIP codes are similar to others. By determining which ZIP codes within a recruiting station's territory have a higher potential

for producing recruits, recruiting commanders will be able to better target their operations and better assign their recruiters. The recruiting headquarters will also be able to better allocate ZIP codes between their recruiting companies and better assign missions to each recruiting company.

Without this type of analysis, recruiters can only focus their efforts based on historic relationships with high schools, predictions of the number of qualified military available (QMA) in a ZIP code, and records of past recruits assessed from a ZIP code. This means that a ZIP code that does not produce recruits for a few years will not get any new attention in subsequent years, potentially leaving a large source untapped. Hojnowski (2005) shows that production per recruiter can be determined as a function of the number of recruiters assigned to a station and the goal assigned to that station. This indicates that using past production to predict potential for future production is an effective metric because recruiters will generally contract the number of recruits they are assigned.

The Army is currently in the midst of downsizing and has begun reducing the number of required recruits to shape the smaller Army. In fiscal year 2014, the Army only needed to recruit 57,000 Active Duty (AD) soldiers, compared to 69,000 in 2013 (United States Army Recruiting Command 2016). At the same time, the Army is trying to recruit higher-quality recruits and is maintaining its force of recruiters at its current size (Lopez 2014). With a better understanding of the recruiting terrain at the ZIP code level, the Army will be able to more efficiently assign its recruiters and could possibly reduce the recruiter force to better meet budgetary requirements.

C. RESEARCH QUESTIONS

Can a classification or categorization of United States ZIP codes determine which geographically dispersed ZIP codes have similar characteristics with similar recruiting potential? We answer this problem statement by answering two contributing questions:

1. Which factors are important for classifying similar United States ZIP codes?
2. Are the classifications useful for predicting the recruiting output of a specific ZIP code?

D. SCOPE OF STUDY

This study uses clustering techniques to classify United States ZIP codes for the purpose of United States Army recruiting. We use economic, demographic, and county health data, and the presence of higher education institutions and military bases to determine which geographically dispersed ZIP codes share similar characteristics. All datasets are from the past ten years and are publicly available. We focus only on United States ZIP codes that occur within the contiguous United States (CONUS). To assess the applicability of the cluster assignments, we use the numbers of United States Army national leads as a proxy for a ZIP code's potential to produce recruits.

1. Constraints

The largest constraint on this project is the use of only publicly available and already collected data. The sponsor for this study, USAREC, has several data sources they use to guide recruiting operations. They use population estimates from companies such as Woods and Poole and the Lewin Group to determine the likely number of QMA in any ZIP code (United States Army Recruiting Command 2016). Both the Lewin Group and Woods and Poole also provide demographic data along with forecasts for several years in the future. USAREC also has data at the ZIP code level for past Army and Department of Defense contracts signed.

Additionally, USAREC receives market data from the Joint Advertising, Market Research & Studies Group (JAMRS) (United States Department of Defense 2016). Since the Department of Defense already pays for proprietary data, this study is constrained to using data that has already been procured or is publicly available.

2. Limitations

We limit the scope of this study to ZIP codes inside CONUS, omitting data from Alaska, Hawaii, or any of the outlying territories. This limitation is necessary because the availability of data for some of those territories is far less than for CONUS. Also, since these territories are so far from the American mainland and have their own unique

populations and characteristics, they are likely to be outliers from the rest of our ZIP code observations.

A second limitation on this study is to only study the regular Army. Parker (2015) already concludes the ability of AR units to meet their manning requirements has a strong correlation to the location of Reserve unit basing. Since this study will not address changes in the location of basing, we leave out AR units and focus only on AD recruits.

Furthermore, we limit the study to enlisted recruiting. While officer recruiting is a large portion of annual recruiting, the requirements for officers are significantly different: they are targeted at the college or college graduate level instead of at the junior and senior years of high school. It is more effective to cluster ZIP codes to see which ones contribute enlisted recruiting alone compared to those that could contribute to enlisted or officer recruiting.

The final limitation is the timing of the data used in the study. Since all of the data is collected by different agencies, it was all collected with different methods at different times. The United States Census Bureau (2007) conducts the Census and the Economic Census every ten years and much of the data is collected at the county level instead of at the ZIP code level. Some of the data must be manipulated to be useful to our study. For example, the number of annual deaths in a ZIP code may not be as telling a statistic as the death rate. See Chapter III for a more detailed discussion of the data used.

3. Assumptions

Due to the limitations of our data, we make several assumptions. First, for the Community Health data and any other county level data, we assume the distribution of population throughout the county is homogenous. While this is most likely not true, we have no other methods available to distribute our data to the various ZIP codes throughout the county. As described in further detail in Chapter III, we assume the percentage of residential addresses in each ZIP code is a valid means of apportioning county level data.

Our second assumption is that the variables that we create throughout—such as distance to the nearest military base—are valid and contribute to classification of ZIP

codes. Since several of these variables do not exist intrinsically in ZIP codes, we must find various ways to attribute environmental characteristics in a unique way to each ZIP code to include these variables at all.

During the assessment phase of our analysis, we assume the number of national leads is a valid measure of a ZIP code's potential to produce recruits. Since the activity of recruiters highly influences local leads, national leads are more indicative of potential recruits who seek out the Army and request more information about joining. While the population count will be a strong influencer of the number of leads a ZIP code can generate, some ZIP codes have a much higher rate of leads per person than others.

The final assumption that we must make is that while the boundaries of a ZIP code do change periodically, data collected over a span of eight years is all generally representative of the same geographic area. Since we are dealing with data from over 3,000 counties and over 34,000 ZIP codes (Housing and Urban Development 2016), we can assume that while there will be some outlier ZIP codes where the population has changed significantly, these will not affect the overall classification process.

E. STUDY OUTLINE

In Chapter II, we discuss background information that pertains to recruiting and the ZIP code categorization problem. We first explore current efforts by USAREC to determine which markets have higher potential to produce recruits. We also discuss existing means of categorizing geographic areas. We then discuss past analyses of the factors that contribute to a market's potential to produce recruits. In Chapter III, we discuss the data we use in this study, where we obtain it, and how it relates to recruiting.

In Chapter IV, we discuss methods we use to assign our ZIP codes to clusters and how we determine which method is better. We also discuss the means we use to assess the applicability of the clusters that we create. Once we determine a good method for assigning ZIP codes to clusters, we use these cluster assignments to predict a ZIP code's potential to produce recruits. Finally, in Chapter V, we discuss our results, the implications and uses of these results and potential avenues for future research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

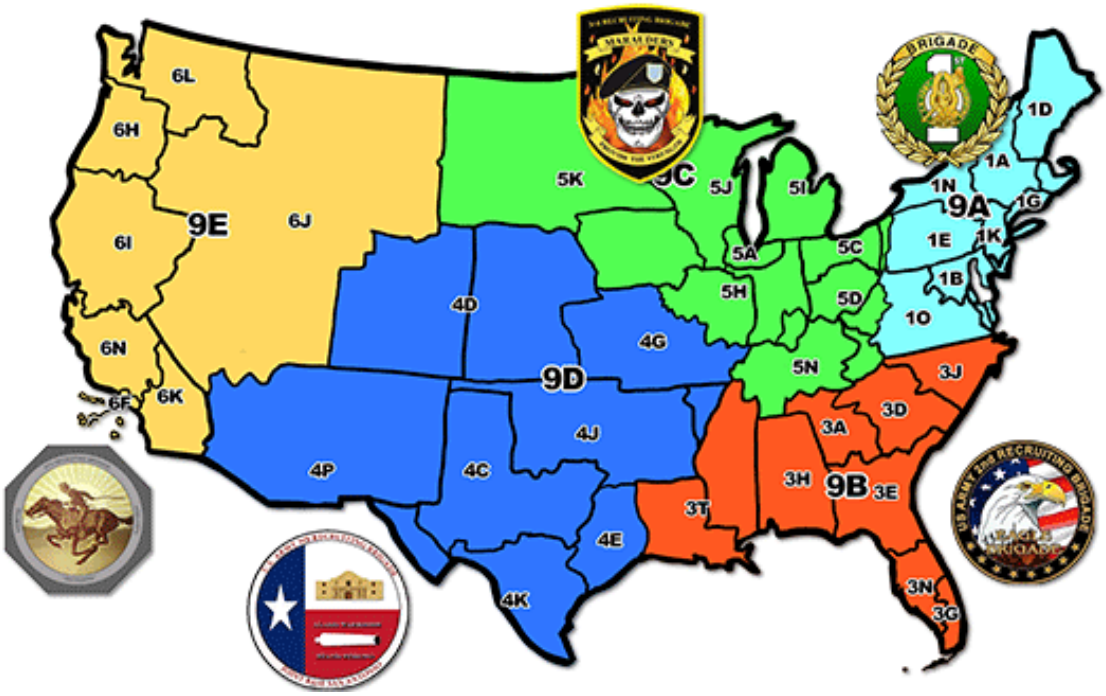
A. RECRUITING CONSIDERATIONS

USAREC is headquartered at Fort Knox, Kentucky. Its commander and staff are responsible for overseeing five recruiting brigades responsible for different regions of the United States. An additional brigade is responsible for recruiting outside the contiguous United States (OCONUS) and other units responsible for special operations recruiting and recruits with medical or legal degrees (United States Army Recruiting Command 2016).

Per USAREC, each recruiting brigade is made up of several battalions, with each of those made up of several recruiting companies. Each recruiting company oversees multiple recruiting centers. Each brigade is led by a colonel, and each battalion is led by a lieutenant colonel. They oversee over 250 recruiting companies and 948 recruiting centers (United States Army Recruiting Command 2016). Each recruiting station is staffed by the enlisted soldiers in the rank of sergeant and staff sergeant, and is responsible for visiting high schools, hosting community events and calling potential recruits to meet their assigned mission.

1. Recruiting Boundaries

The recruiting unit boundaries, from brigade down to company, cross state, county, and even city lines. USAREC assigns its brigades regionally, as depicted in Figure 1. The only United States government boundary that USAREC uses is the ZIP code boundary. Each Army recruiting station is responsible for anywhere between 4 and over 250 ZIP codes, with the average number being 11.2 ZIP codes (United States Army Recruiting Command 2015b). Because ZIP codes are assigned at the recruiting station level, the USAREC staff uses them to track the number of recruits in an area and to develop strategies for future recruiting.



The number and letter combinations refer to the brigades and battalions, respectively.

Figure 1. United States Army Recruiting Command Boundaries. Source: United States Army Recruiting Command (n.d.).

2. Recruiting Personnel

Recruiter assignments, like unit boundaries, are made one level up (United States Army Recruiting Command 2013 p. 9–2). A recruiting battalion is therefore responsible for assigning recruiters to the individual stations and the recruiting brigade is responsible for assigning boundaries for recruiting companies. Each recruiting station has between 2 and 19 recruiters (United States Army Recruiting Command 2015b). These recruiters are responsible for office administration, finding and contracting recruits, transporting recruits to pre-enlistment medical appointments, and ensuring they report to the Military Entrance Processing Station on time to ship to their initial training.

3. Recruiting Mission

The recruiting mission is a target number of contracts to sign. The mission is assigned by commanders down to the company level. The Commanding General of USAREC is responsible for assigning the overall annual mission for AD and AR

recruiting based on the total requirements that the Deputy Chief of Staff of the Army for Personnel assigns. The USAREC Commanding General then apportions the total mission down to the brigades.

To set a mission baseline, USAREC currently uses a four-year rolling model to calculate the future missions for each battalion. To determine the required future production, they weight the past four years of production, weighting the prior year at 40%, the year before at 30%, two years prior at 20% and three years prior at 10%. They multiply these weights by the past Department of Defense production (50%), past Army production (20%), and number of QMA (30%) (Fleischmann and Nelson 2014). USAREC makes these calculations at the ZIP code level then aggregates the results to assign the brigade level missions.

Once a battalion assigns a mission to a company, the company commander is responsible for dividing that mission among his recruiting stations, taking market factors and capabilities of personnel into account. The recruiters at a recruiting station share a mission among themselves, so there is an incentive for all of the recruiters in each office to work together (United States Army Recruiting Command 2013 p. 9–2)

B. EXISTING GEOGRAPHIC CLASSIFICATIONS

Geographic boundaries have a significant effect on this study and on past studies for two reasons. First, geographic boundaries affect how USAREC assigns its units and allocates its recruiters and their goals. Second, geographic boundaries affect how data is collected and how we can use that data to better understand the recruiting environment. In this section, we discuss the different levels are classification that are relevant to this study.

1. ZIP Codes

The United States Postal Service (USPS) defines ZIP code boundaries and changes them frequently based on supply and demand of postal services (United States Postal Service 2016). ZIP codes are especially useful because there are no intellectual property right claims on them; they are a public good (U.S. Postal Service Office of the

Inspector General 2013). This allows businesses and federal agencies alike to use them to develop products and to use them to track information.

There are currently over 42,000 ZIP codes. Some are solely for post office boxes and large commercial entities. These special ZIP codes do not have a permanent population and are not helpful to recruiters, although they are still assigned to a recruiting station. The Department of Housing and Urban Development (HUD) publishes quarterly a list of all ZIP codes with addresses that allows data translation from county, Core Based Statistical Area (CBSA) or census tract level to ZIP code level (Housing and Urban Development 2016) based on the percentage of each county's addresses in a ZIP code. There are ZIP codes dedicated to single businesses, which hold less than 1% of a county's addresses. We ignore these and include only those ZIP codes that have greater than .01% of a county's ZIP codes. This leaves us with 34,685 ZIP codes.

The challenge with using ZIP codes to define local regions is that they are not static and are not tied to any physical boundaries. All ZIP codes are assigned based on the generally central location of a physical Post Office and the proximity of the area to that office. This allows USPS employees to efficiently deliver mail each day from a central location (U.S. Postal Service Office of the Inspector General 2013). This means that when data is collected at the ZIP code level it quickly becomes out of date. Data collected at a different level, such as county, can always be translated to the ZIP code level using an updated HUD list, but this comes at the expense of accuracy (McDonald 2016 p. 116).

2. Recruiting Regions and Boundaries

As shown in Figure 1, the recruiting regions do not follow state or federal government boundaries. USAREC assigns these boundaries based on regions of the United States for command and control of subordinate units and to provide those units with strategic guidance. Because these regions are so large, knowing that a ZIP code is in a particular region does little to contribute to the ability to predict how well USAREC will recruit in that ZIP code precisely. Instead, the regions could be used to partition ZIP codes into regional groups. This gives each region its own unique model to predict recruiting potential. For example, Intrater (2015) modeled Navy recruiting regionally.

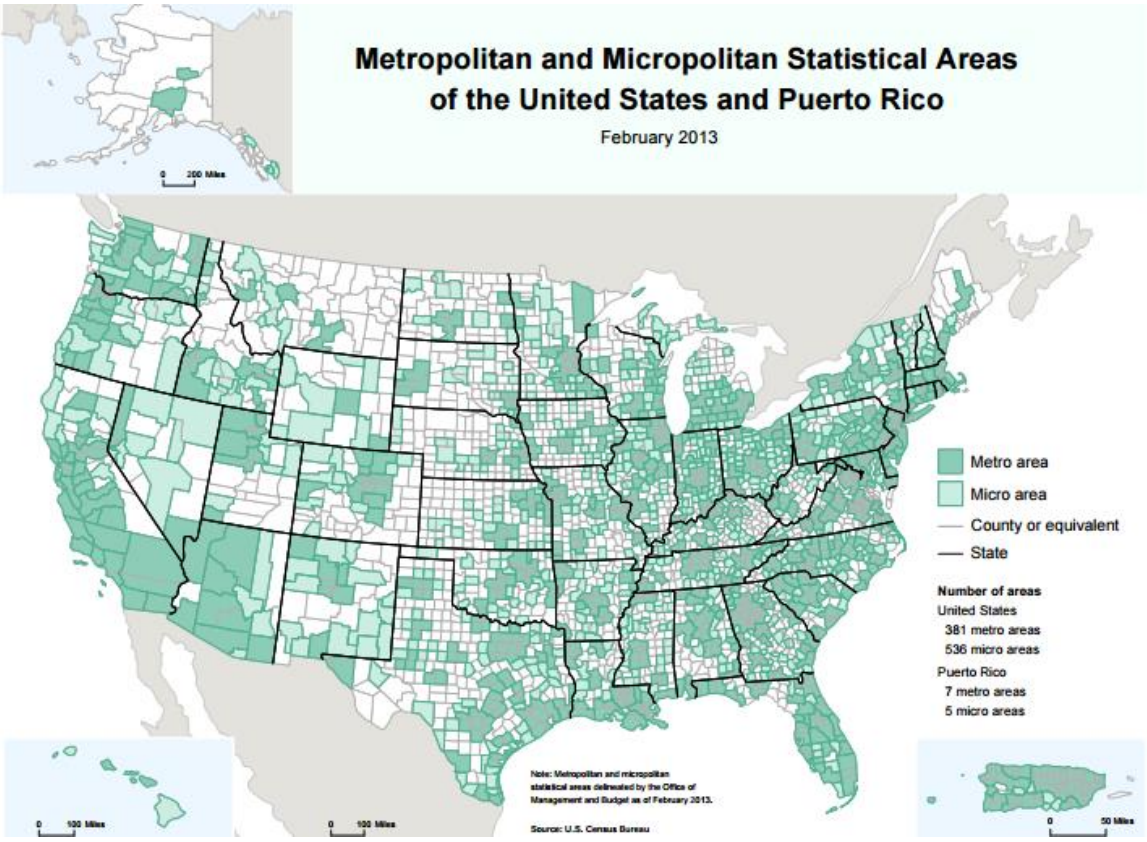
3. Counties

County boundaries are more specific than recruiting region boundaries, but are too broad for the level of analysis USAREC needs. With over 3,000 counties with an average population of over 25,000 people (Centers for Disease Control and Prevention 2010), individual recruiters are not able to target specifically based on a county. With the smallest recruiting stations being made up of only two recruiters, the commanders and staff at USAREC need to be able to provide information detailed enough for these individuals to act on. Because counties are a level of government, many federal and state agencies collect data at the county level. These include population data, health data, and home value data. This requires us to translate the data to the ZIP code level using a crosswalk between counties and ZIP codes.

4. U.S. Census Bureau

A geographic boundary more specific than counties is the CBSA. Although they are more specific, they do cross county and even state boundaries. This geographic demarcation is designed by the Office of Management and Budget and is based on U.S. Census Bureau data (United States Census Bureau 2013a). It consists of 31 metropolitan areas that each have a population of over 50,000 and 556 micropolitan areas that have populations between 10,000 and 50,000 (United States Office of Management and Budget 2015). CBSAs are defined for all U.S. states and their territories, as shown in Figure 2.

Because the Census Bureau is the agency whose sole mission is to collect data on the United States population, many reports have data available at the CBSA level. However, since these areas are focused on large concentrations of population, over 1/3 of ZIP codes are aggregated into CBSA 99999, which includes all rural areas. These are the unshaded areas in Figure 2. With no specificity for a large portion of the United States, this geographic level is not useful to USAREC for targeting recruits at the company or station level.

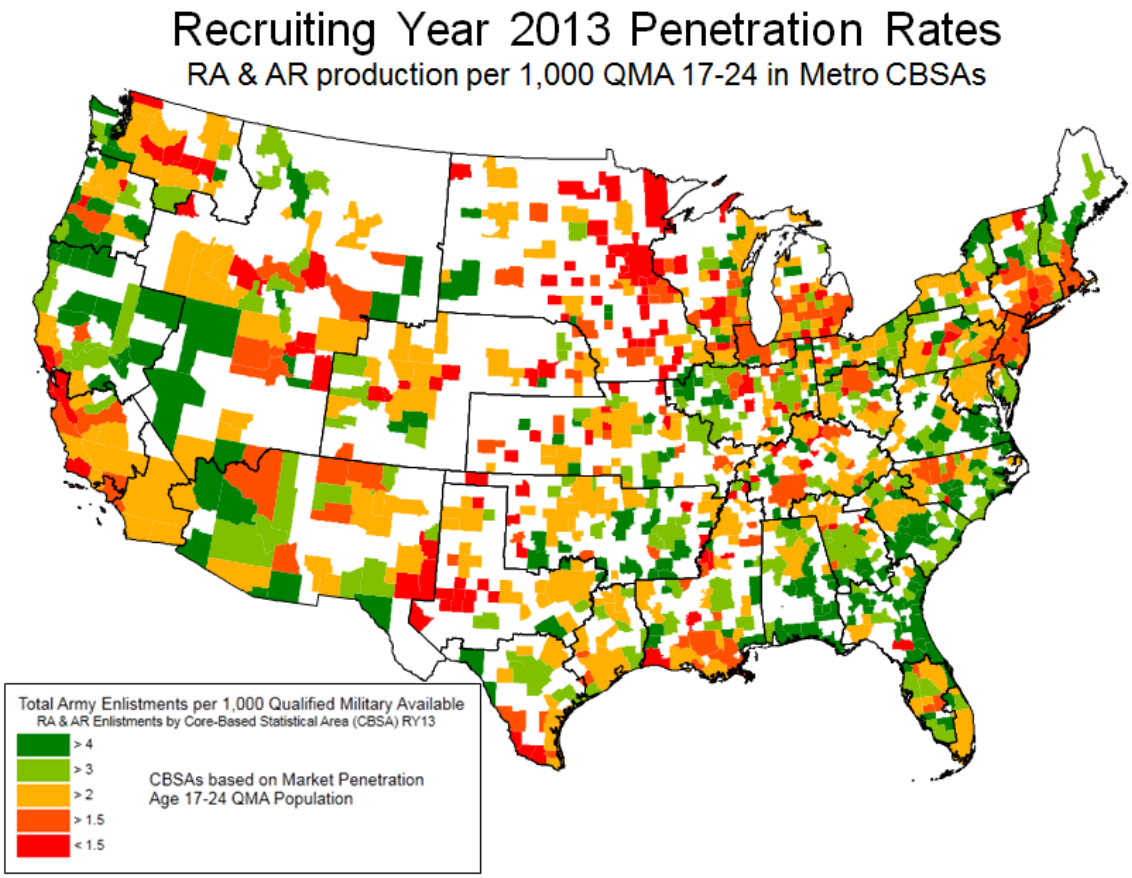


This map shows a breakdown of the Core Based Statistical Areas in the United States, provided by the United States Census Bureau. Unshaded areas are defined as CBSA 99999 or "rural."

Figure 2. Metropolitan and Micropolitan Statistical Areas of the United States and Puerto Rico. Source: United States Census Bureau (2013b).

C. PAST ANALYSIS OF FACTORS INFLUENCING RECRUITING

USAREC uses the numbers of QMA and the numbers of past production to assign brigade-recruiting missions, as described in Section A. Analysis shows that each geographic area does not produce recruits equally. USAREC uses market penetration to compare the production of different areas. Market penetration is defined as the number of recruits assessed as a percentage of the QMA population of an area (United States Army Recruiting Command 2013). Figure 3 shows the different levels of market penetration aggregated at the CBSA level. Since each area—or market—does not produce at an equal rate, USAREC strives to adjust those missions based on the potential of each individual market.



Regular Army (RA) and Army Reserve (AR) Penetration rates as defined by number of enlistments per 1,000 QMA in each CBSA. White areas are rural and are not assigned to any CBSA.

Figure 3. Recruiting Year 2013 Penetration Rates. Source: United States Army Recruiting Command (2013).

The USAREC recruiting manual calls for commanders at all levels to use mathematical modeling and analysis to better inform their decisions for recruiting operations (United States Army Recruiting Command 2013). The same manual assigns the intelligence section of the USAREC staff (USAREC G2) to lead the data gathering and analysis effort. In addition to annual and quarterly assessments of the market and geographic and personnel factors, several analysts have conducted lengthy studies to evaluate the impact of various factors on recruiting operations. For each study, the goal is the same: maximize the number of quality recruits assessed into service while minimizing the amount of resources expended.

For clarity of discussion, we divide this problem and past research into two aspects that can be studied individually. The first aspect to study is the supply of recruits itself. These studies focus on determining the number of potential recruits in a population and what types of populations can be expected to provide the most recruits. These studies can also analyze what motivations actually cause a person to enlist in the military. Factors involved in this first aspect generally change slowly because they are intrinsic to each area or population studied.

The second aspect focuses on the efforts of the recruiters themselves, and on outside influences that can change frequently. Analysts study the effects of different incentives or motivations on recruiting outcomes. These include positive incentives such as recruiting bonuses and disincentives such as a high amount of non-military job opportunity nearby. These dynamic factors also include the number of recruiters assigned to an area or the amount of money spent or methods used in advertising. While this study focuses on the more stable aspects of population to generate the ZIP code categorizations, it also draws insights from incentives and recruiter assignments to test the value of these categorizations.

1. Studies of Market Characteristics

Previous work by Oh (1998) shows that recruits from different ethnic and socioeconomic backgrounds enlist based on different motivations. USAREC uses the term "environmental factors" to describe factors affecting a population or geographic area (Fleischmann and Nelson 2014). These factors include economic conditions and an individual's propensity to join the service. We will use the term market characteristics to describe the environmental factors that USAREC already studies and to also include other intrinsic characteristics of a population such as its demographics, health, education, and proximity to military bases.

USAREC uses the Recruiting Market Index to adjust the baseline missions assigned to its recruiting battalions. This is based on a regression model that takes quarterly inputs such as number of recruiters, propensity, and unemployment to determine the following quarter's adjusted mission (Fleischmann and Nelson 2014). They

then use these predicted changes in production and the unit commander's predictions to increase or decrease the unit's mission by 0 to 19%. While these adjustments are helpful, they are based mostly upon past production and do not take into account significant characteristics of the populations.

USAREC's primary tool to understand the populations themselves is the Potential Rating Index for Zip Markets, New Evolution (PRIZM NE) market segmentation that Nielsen Company (formerly Claritas) created and owns. (Nielsen Company 2016). This analysis tool assigns each household in the United States into one of 66 proprietary segments based on over 150 variables that Nielsen Company extracts from census data, demographic factors, and surveys. Nielsen Company further categorizes these 66 segments into one of 14 social groups and 11 age categories. Figure 4 shows the 66 segments sorted by social groups. This data is used for marketing by hundreds of companies across the United States.

In 2007, USAREC created the 39 Army Custom Segments (ACS) (Dorminey 2007). These segments rely on the PRIZM data with the addition of attitudes, motivators, and barriers to service so that each potential recruit falls into one of the 39 ACS (Clingan 2007). USAREC then categorizes these 39 tactical segments into 10 strategic segments and determines which segments have a population with a higher propensity for joining the military. Recruiters can use this information to better inform decisions on who they should attempt to recruit, and how.

Nielsen Company's PRIZM NE market categorization shows how Nielsen divides the U.S. population into 66 market segments that can be grouped into 14 social groups based on wealth and urbanization.

Figure 4. Nielsen Company's Potential Rating Index for ZIP Markets, New Evolution (PRIZM NE) Market Segmentation. Source: Nielsen Company (2013).

Army recruiters have access to market segmentation and analysis through the Segmentation Analysis and Market Assessment (SAMA) tool, which is available to the entire USAREC organization. The Army currently uses the 39 custom segments to

analyze the recruiting market. SAMA allows recruiters to see the population of each of the 39 custom segments that make up the total population of each ZIP code in their area of responsibility. While there are certain segments that are known to outperform others, dividing a recruiter's 11.2 average ZIP codes into 39 ACS does not simplify the problem.

Additionally, the current SAMA tool does overestimate the production for over 96% of recruiting centers (Marmion 2015). Marmion (2015) finds that the reason for this over-estimation is that the model uses the higher of the past market penetration for the recruiting center and the company responsible for it. That is, if a company is able to achieve a penetration rate of 20 enlistments per 10,000 QMA within one of the market segments, all centers within that company are expected to achieve that same rate. Since the company penetration rate is a weighted average of the center rates, a single high-performing center raises the standard for all of the rest of the centers within that company (Market Analysis Division 2015). He further determined that it would be possible to more accurately estimate production if the SAMA tool relied on PRIZM NE segmentation instead of Army Custom Segments. For more information on the SAMA tool, how it is used, and recommendations for improving it, see Marmion (2015).

By using the SAMA tool and market segmentation, it is possible to predict the number of recruits from different segments. By aggregating the segment populations, it is also possible to predict the expected production from a geographic area. Further analysis indicates which environmental factors most contribute to the increased production of these populations and geographic areas (Marmion 2015).

2. Studies of Recruiting Incentives and Efforts

The second major area analysts study to build recruiting models is the effects of incentives and efforts that USAREC can control, such as bonuses and recruiter assignments (Fleischmann and Nelson 2014). ZIP code valuation studies by the Marsh Group for JAMRS has shown that locations of recruiting stations and the number of recruiters are consistently among the strongest predictors of accessions (Gibson, Hermida, Luchman, Griepentrog and Marsh 2011). These periodic Marsh Group studies include economic, demographic, crime, and military factors to fit zero-inflated Poisson

regression models to predict future recruiting. They also aggregate their models to the CBSA level and find that the models predict as well as models aggregated to the ZIP code level. This study also finds that proximity to a recruiting station or an increase in number of recruiters of any branch generally has a positive effect on Army accessions.

Williams (2014) creates a single index variable to determine the relative value of different recruiting markets for Navy recruiting. He tests 18 different factors to determine which 5 are the most significant to create a valuation index for monthly and 3-year models. Williams finds that 3 of the 5 most significant factors in the 3-year model are directly tied to recruiting efforts. The most significant factors are the number of Navy recruiters assigned to the area, the average number of national leads per population, the competition index—a measure of recruiting efforts of other services, the proximity of the ZIP code to the nearest Navy recruiting station, and the percentage of individuals in the ZIP code with a high school education. These five factors contribute to a model that is able to explain 55% of the variation in the number of Navy accessions in ZIP codes (Williams 2014).

Research shows that assigning a larger number of recruiters to an area generally results in a larger number of accessions. The Army then uses the number of accessions in its SAMA models to predict future expected accessions and potentially assigns more recruiters to the areas that are expected to have higher production. This circular logic presents the risk that there are ZIP codes in the United States that have high potential for producing accessions based on their intrinsic qualities, but that do not actually result in high production because they do not have a sufficient number of recruiters assigned. This study will determine the potential of an area to produce accessions independent from the number of recruiters assigned or the proximity to recruiting offices.

III. DATA SOURCES

All of the datasets used to build the ZIP code classifications are publicly available on government agency websites. The collecting agencies include the Internal Revenue Service (IRS), the Census Bureau, the Centers for Disease Control and Prevention (CDC), and HUD. Wherever possible, we use data originally collected at the ZIP code level; however, some of the data is at the county level and we convert it to ZIP code level using a crosswalk between counties and ZIP codes. More information on this conversion is available in Appendix A.

To include a ZIP code, it has to contain at least one residential address. Of the 41,306 CONUS ZIP codes assigned to recruiting stations, 34,007 contain residential addresses. The remaining 7,299 are ZIP codes that belong entirely to Post Office boxes, universities, or businesses. The list of United States ZIP codes with residential addresses is obtained from the HUD county to ZIP code crosswalk (Housing and Urban Development 2016).

Several of the economic datasets containing number of businesses or total salaries paid are collected for commercial ZIP codes with no residential population. Since the purpose of using this data is to take into account economic opportunity available to the population, we attribute this data to nearby residential ZIP codes, as explained in Section C of this chapter.

A. COMMUNITY HEALTH STATUS INDICATORS

The Community Health Status Indicators (CHSI) data was collected by the CDC in 2010. This dataset contains demographic information such as total population, population density, and race and age ratios. It also contains health data including leading causes of death, numbers of suicides, number of births, birth rates, and reports of certain diseases (Centers for Disease Control and Prevention 2010). A full listing of the 152 variables we use can be found in Appendix B. The dataset also contains many national average and confidence interval factors we do not use.

The CHSI data was collected at the county level so we convert it to the ZIP code level using the HUD county to ZIP code crosswalk. This crosswalk uses the portion of residential addresses in each ZIP code within the county to apportion the data. Although the data was collected in 2010, we use the 2015 crosswalk because we are analyzing ZIP codes as they are today and not as they were. County lines do not change as often as ZIP codes so we do not expect to lose much information by using the 2015 crosswalk. Only three counties from 2010 with a combined population of less than 2,500 no longer exist. More information on this crosswalk can be found in Appendix B.

Although the original CHSI data contains numbers of deaths and numbers of incidents of disease, we convert totals to the rates of these events based on the total ZIP code population sizes. Converting raw totals to rates comes at a price. Because the CHSI data is recorded at the county level all ZIP codes completely within a county have constant rates for 150 of its 152 CHSI variables. This causes the clustering based on CHSI variables to be at the county instead of the ZIP code level. For the 5,000 ZIP codes that span multiple counties, the rates and numbers are averaged based on population sizes.

B. INDIVIDUAL INCOME TAX RETURNS

The individual income tax return data was collected by the IRS in 2013 at the ZIP code level. It contains information such as the number of returns filed, average number of dependents per return, the total Adjusted Gross Income, and total Capital Gains for each ZIP code (Internal Revenue Service 2016). As with the CHSI dataset, we transform the total numbers per zip code to averages. Also, the numbers of returns by filing status are turned into percentages. More information on the 65 variables included in this data frame can be found in Appendix A.

To protect the privacy of individual taxpayers, only ZIP codes with more than 100 income tax returns are included in this dataset (Internal Revenue Service 2016). The IRS also excludes data for any ZIP codes where a high percentage of the data came from a small number of taxpayers. That leaves us with IRS data for 27,578 of our 34,685 ZIP codes. We code the observations without data as having no data available.

C. ECONOMIC CENSUS

The Economic Census is one of the few datasets collected by the United States Census Bureau at the ZIP code level (United States Census Bureau 2007). The most recent economic census with data available at the ZIP code level is from 2007.

The Economic Census contains the total number of business establishments in each ZIP code of 8 different establishment types such as retail, health, or education. An establishment is considered to be a single building or location, so a chain restaurant may have multiple establishments in one ZIP code. In addition to the total number of establishments, we construct a second variable that shows the percentage of establishments of the different types in each ZIP code.

With this dataset, we attempt to represent the availability of economic opportunity in the particular ZIP code. Since this dataset includes businesses, several thousand observations are for ZIP codes without residential addresses (United States Census Bureau 2007). To make use of this data, we use the R package *zipcode* to measure the distance between the commercial ZIP code and all residential ZIP codes that are already in our dataset (Breen 2015). We then assign the data for each commercial ZIP code to the nearest residential ZIP code. While it is true that other ZIP codes also benefit from the presence of these businesses, the same is true for every ZIP code with businesses nearby but not in their ZIP code. See Appendix B for more information on the methods used to generate this dataset's 33,430 observations.

D. COUNTY BUSINESS PATTERNS

The Census Bureau also collects the annual County Business Patterns survey. It contains the total number of establishments, employees, and quarterly and annual payrolls for each ZIP code (United States Census Bureau 2011). The most recent year for which data is available at the ZIP code level is 2011. This dataset also contains a large number of commercial ZIP codes that are attributed to residential ZIP codes in the same manner as the economic census data. For more information on the modifying of this dataset, see Appendix B.

E. MILITARY BASES

Brown and Rana (2005) show that prior exposure to military increase the propensity for military service for American youth. To capture the increased percentage of military children near bases along with the other impact of nearby basing we use two datasets to determine the distances to the nearest military base and the size and type of the largest bases within 10 and 50 miles.

We collect this data from two different datasets and manually combine the base population data with the base location data. The first dataset from the Census Bureau contains the locations of over 700 different U.S. bases (United States Census Bureau 2012). Because the dataset includes multiple entries for several bases, we manually remove duplicates, ranges, bases that are now closed, and bases outside CONUS.

The second data file contains a list of U.S. military and Coast Guard Installations and the number of each branch of service and number of civilians employed by those bases (Defense Manpower Data Center 2009). The same base can be listed in several ways, such as Fort Meade, Fort George Meade, or Ft. Meade. To ensure we capture the correct information, we manually combine the datasets. We also add a column denoting the type of base, which includes categories such as Army, Air Force Reserve, or Depot. The latter category includes plants and maintenance facilities that are primarily staffed by civilians. The Coast Guard stations and many smaller bases do not have a recorded population size (Defense Manpower Data Center 2009). These we assign a zero population size in the combined dataset we attempt to minimize the effect of the zeros by using a type variable for each base.

We then calculate distances and create nine variables. These are the distance, type, and population of the nearest base to each ZIP code and the distance, type, and population of the largest base within 10 and 50 miles of the ZIP code. If there are no bases within 10 miles, the type is listed as “none” and the population and distances are coded as blanks.

F. VETERAN POPULATIONS

As another measure of previous exposure to military, we use several indicators of veteran population in a ZIP code. Intrater (2015) shows in his analysis that the total number of veterans in a population is a significant indicator of a ZIP code's recruit production. To account for the veteran population in a ZIP code we use data from the United States Census Bureau's American Community Survey from 2014 that (United States Census Bureau 2014).

This dataset contains the numbers of veterans living in each county by age, gender, and which conflict period they served in. Because this data was originally collected at the county level we transform it using the same county to ZIP code crosswalk that we used for the CHSI dataset in Section A (Housing and Urban Development 2016). We further transform this dataset to include the estimated percentages of total population in a ZIP code that are veterans. For example, the original dataset includes the estimated number of female veterans over 18 and total females over 18 in an area. We include the estimate total number of female veterans over 18 and the percentage of females over 18 that are veterans. For more information on the variables in this dataset, see Annex A.

G. SECONDARY EDUCATION INSTITUTIONS

Intrater (2015 p. 50) shows that the presence of large universities increases the likelihood that a ZIP code will yield zero recruits. Since the presence of a major university indicates the presence of an opportunity or alternative to military service, we will include these, along with all 7,300 institutions in the Integrated Postsecondary Education Data System (National Center for Education Statistics 2016). This dataset contains latitude and longitude for each registered institution, along with a size category between 1 and 5. A size-category 5 institution is a major university with thousands of students, while a size category 1 institution could be a small university, an auto mechanic school, or a beauty school.

To include all levels of education opportunities in our dataset, we calculate the distance from the centroid of each ZIP code to each education institution. This is similar to the method that Pinelis (2011) uses to show that ZIP codes with a larger distance to the

nearest school have a higher likelihood of producing recruits. We then total, for each ZIP code, the number of institutions of each size within 10 miles and the number within 50 miles. This allows us to capture the immediacy of institutions just down the street along with the reduced influence of a college several miles away. For more information on these calculations see Appendix B.

H. NATIONAL LEADS

To check the applicability of our clusters, we use the national leads data provided by USAREC (2015a). This dataset contains 2,026,747 recruiting leads from January 2011 to September 2015. These leads are generated by potential recruits who are interested in joining the Army who call 1–800-USA-ARMY, who request more information from www.goarmy.com, or who sign up for information about joining the Army at a career fair or sporting event. We use national leads instead of local leads because the national leads are less influenced by the efforts of local recruiters and are more representative of the ZIP codes themselves.

Of these leads, 1,984,430 (97.9%) are attributed to one of our 34,007 ZIP codes. We total the number of leads for each ZIP code across that entire time span. There is a possibility that the same person generated more than one national lead, such as someone who signed up on a website and at a sporting event. Since we have no means of accurately evaluating and correcting for this effect we assume this effect is minimal or is indicative of an increased quality of lead. Even by keeping all data for the 57 month period there are no recorded national leads for 3,317 (9.8%) of our ZIP codes. The ZIP codes with no recorded national leads are mostly smaller markets with a median population of 273 people. This is small compared to the median population of all ZIP codes we study, which is 2,806.

IV. CLUSTERING AND ANALYSIS

In this chapter, we discuss how we cluster ZIP codes, how we analyze the cluster assignments, and the how we check for the applicability of those results. We first determine how to categorize the 347 variables into five types or categories to ensure no set of variables has more influence on the clustering assignments. We then cluster each of the five sets of variables separately to construct five sets of ZIP code assignments, one per variable category. Finally, we assess the cluster assignments for applicability.

A. CLUSTER BUILDING

We are comparing and clustering ZIP codes based on characteristics that are not affected by recruiting. That is, we are not comparing them based on the number of recruits they produce or the propensities of people in ZIP codes to join the military. Since this means we do not have any response variables we need to use a method that allows for unsupervised clustering. The *treeClust* package for R of Buttrey (2015) accomplishes this task. For a description of the algorithms used in *treeClust* see Buttrey and Whitaker (2015).

This algorithm, which we call Tree Clustering, allows the clustering of observations based on both numeric and categorical variables, variables with missing values, and is invariant to how numeric variables are scaled. Clustering based on Gower dissimilarities (Gower 1971) is a well-known approach that also has these properties. However, Tree Clustering is less influenced by extreme values in numeric variables and is particularly good at clustering in the presence of noise variables (Buttrey and Whitaker 2015).

The Tree Clustering algorithm takes as inputs a set of variables, a method of calculating distances or dissimilarities, an algorithm to compute final clusters, and a value, k, for the number of clusters to create. We determine how to categorize our variables then we create a design of experiments to create clusters with different methods for computing dissimilarities, different clustering algorithms, and different values of k.

1. Variables

The first decision we make for the clustering is to determine which combinations of variables to use to cluster ZIP codes. With the seven data sources described in Chapter III, we have a total of 347 variables for our 34,007 ZIP codes. Since the CHSI dataset and the individual income tax return datasets together contain 269—over 77%—of those variables we need to find a method that will prevent any one dataset from wielding an overwhelming influence on our analysis. This is especially important since the CHSI and veteran data was collected at the county level. If we attempt to cluster ZIP codes based on all 347 variables, ZIP codes in the same county would tend to end up in the same cluster, yielding essentially county-level clusters.

To prevent datasets with large numbers of variables from masking the impact of the other datasets, we divide our variables into categories, cluster based on those categories then attempt to create final clusters based on cluster assignments from the within-category clustering. This allows us, for example, to transform 138 health-related variables into a single categorical variable indicating cluster assignment of each ZIP code based on the health data alone.

The variable categorization developed by Intrater (2015) informs our own variable categorization. Intrater uses military influence and recruiter workload, crime, population characteristics, economic stability, education opportunities, and veteran population (Intrater 2015). Because we are conducting unsupervised clustering and are not including recruiter data, we group veteran data with the size and type of the nearest military bases. This serves as a proxy for military influence on a ZIP code. We exclude the crime data because this data is not collected uniformly and is only collected when local governments elect to report it (Intrater 2015). See Intrater (2015) for further discussion on the collection of crime data.

We use the categories of demographics, health, education, economic, and military to group our 347 variables into five sets. Some variables from the CHSI dataset are assigned to four of the five categories. Appendix A shows which variables we assign to each of the five categories.

2. Dissimilarity Calculation

The first step of the Tree Clustering algorithm is to create trees to determine the distance between each observation and each of the other observations, which results in an *n by n* dissimilarity matrix, where *n* is the number of observations (Buttrey and Whitaker 2015). To determine the distances, the algorithm fits a series of trees that use each variable as the response variable in turn. These trees are then pruned to an optimal size and the distances of the observations are calculated based on which leaves they fall in in each of the trees. The Tree Clustering algorithm offers four methods of calculating these distances (Buttrey and Whitaker 2015).

The first method, *d1*, returns the proportion of trees in which the observations land in different leaves. That is, the "distance" between two ZIP codes will be the proportion of the final set of trees in which they are not in the same end leaf. The second method, *d2*, includes a measure of quality for each tree, which then affects the final distances. The quality of each tree is related to how well the response variable is predicted by the other variables. Distances between observations in higher quality trees count for more. The third method, *d3*, measures the distances depending on how far apart the observations fall on the tree. Instead of just counting whether or not the observations fall in different leaves, this method calculates a distance as farther if observations are several nodes apart (Buttrey and Whitaker 2015).

Method *d4* combines methods *d2* and *d3*, creating distances based on the different node separations of observations and on the qualities of the trees (Buttrey and Whitaker 2015). See Buttrey and Whitaker (2015) for a more detailed description of the methods to calculate the dissimilarity matrices. In this study we use methods *d1*, *d3*, and *d4*, to determine the optimal method to cluster ZIP codes for recruiting. We do not use *d2* because due to the high degree of dependence among variables of each category, *d2* results in clusters that are very similar to those constructed using *d1*.

3. Clustering Algorithm

Once we calculate the dissimilarity matrices we select which clustering algorithm to use. The *treeClust* package supports "pam," "agnes," "clara," or "k-means" (Buttrey

2015). The algorithms "pam" and "agnes" require a lot of computation time, especially for large datasets. "Agnes" is an agglomerative approach that starts with each observation in its own cluster and combines nearest pairs of clusters (Kaufman and Rousseeuw 1990). "Agnes" calculates at a much slower rate because it uses a bottom-up approach with each observation starting as a separate cluster.

"K-means," "pam," and "clara" are partitioning methods. "Pam" uses medoids as cluster centers with a user specified dissimilarity matrix to measure distances between observations. "Clara" is a version of "pam" that uses a randomly sampled subset of the observations to speed up computations. "K-means" uses centroids as cluster centers and calculates the Euclidian distances between each observation, clustering nearest observations together (MacQueen 1967). This method normally does not work with categorical data but *treeClust* provides a version of the data whose Euclidian distances are approximately the same as those created by *treeClust*. This method computes quickly compared to "pam" and "agnes" and at a similar rate as "clara." For further explanation of "k-means," see MacQueen (1967).

"Agnes" is too computationally intensive to use on a dataset as large as ours, but "pam" is tenable with the use of a high performance computing node. "Clara" uses the same approach as "pam" but works a much faster and with a constant rate because it uses only a sample of the data to create the clusters. See Kaufman and Rousseeuw (1990) for a more detailed explanation of these clustering algorithms. We use "pam," "clara," and "k-means" to create final cluster assignments.

4. Number of Clusters

After choosing a method of calculating dissimilarities and a clustering algorithm, we need to determine the number of clusters to assign the observations to. Since we do not know how many groups of ZIP codes exist in the United States, we cluster our ZIP codes with different values of k and choose the best one based on predictive ability.

As described in Chapter II, the Nielsen Company uses 66 different segments to cluster markets. We treat this as an extreme case where each ZIP code belongs to one of 66 different segments of ZIP codes. However, JAMRS then converts these 66 segments

into 39 Army custom segments (Clingan 2007). This reduction shows the need to simplify the number of segments so commanders can better understand and group the different markets they are responsible for. Further research by JAMRS shows that of the 66 PRIZM NE segments, only 8 are considered "high-performing" segments (Joint Advertising, Market Research and Studies Group 2005). At the other extreme this could mean that the only information needed is whether a ZIP code is high performing or not. This could lead us to use k = 2.

We use 18 as a maximum number of clusters because it is between the mean and median number of ZIP codes per recruiting station—11 and 22, respectively (USAREC 2015a). If USAREC is attempting to assign twenty different ZIP codes per single recruiting station, having twenty different categorizations of ZIP code makes it difficult to distribute all of the ZIP codes evenly based on segmentation.

We use a minimum k of 2 because it is the minimum number of clusters possible. This shows us results using the low value, which is more useful for understanding or visually representing the differences between ZIP codes and a high that allows for more nuance and possibly more predictive power.

With our five variable categories, three dissimilarity calculation methods, three clustering algorithms and seventeen variables sizes we create our models. With a full factorial design of experiments method, we fit 765 different initial models. This is 153 different models for each of the five variable categories. Once we fit our models using these methods, we assess their quality.

B. ASSESSING CLUSTERS

To identify a good model and a good cluster building method for our data, we assess the predictive power of each of our 765 models by fitting a generalized linear model (GLM). We fit a Poisson model for count data as described by Faraway (2006). We use the number of national leads, as discussed in Chapter III as our response variable. Since we have four complete years of national leads data, we use the number of 2011–2013 leads to fit GLM models. We then assess those models by attempting to predict the

number of 2014 national leads. For the training models and test models, we use the different categories of final cluster assignments as factors for independent variables.

We have 1,677,508 national leads for our 34,007 ZIP codes between 2011 and 2014 (USAREC 2015a). Each of these years had a significantly different number of leads as shown in Table 1. While the numbers change from year to year, the average number of leads from 2011 to 2013 is 422,906. This is only 5.1% less than the leads for 2014, which we use to test our models' predictive capabilities.

Table 1. National Leads by Year from 2011—2014 (USAREC 2015a).

Year	2011	2012	2013	2014
Leads	171,819	621,263	475,637	444,667

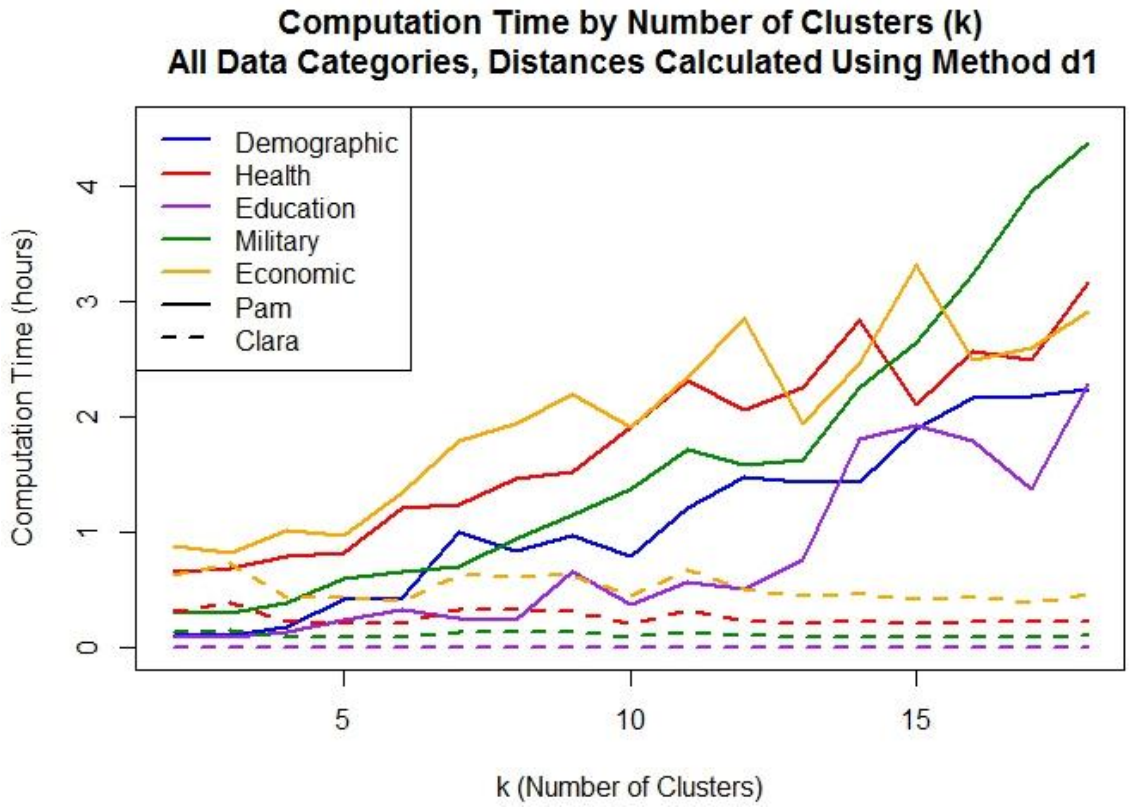
We use the pseudo R-squared value described by Faraway (2006) to assess the quality of our models. This value is calculated by dividing the model's residual deviance by its null deviance and subtracting this number from 1. We do not cross-validate the pseudo R-squared values because with so few independent factors and so many ZIP codes, overfitting is unlikely to be a problem. Once we determine which Tree Clustering methods give the better models, we use those models to predict 2014 national leads and compare the results to those predicted by models fit using the PRIZM NE data. We find a median difference of 1.9 between the actual 2014 leads and predictions from Poisson regression models fit with economic cluster assignments. Models fit with economic data outperform those fit with PRIZM NE data, which have a median difference of 3.4 between actual and predicted 2014 national leads.

C. COMPARING TREECLUST PARAMETERS

The next step is to determine which dissimilarity calculation method to use, which clustering algorithm and which k-value—number of clusters—for each of the five variable type and nine dissimilarity calculation-final algorithm combinations. We do this in part by fitting generalized linear models and comparing the pseudo R-squared values but we also examine the differences in the cluster models themselves.

1. Comparing Clustering Algorithms

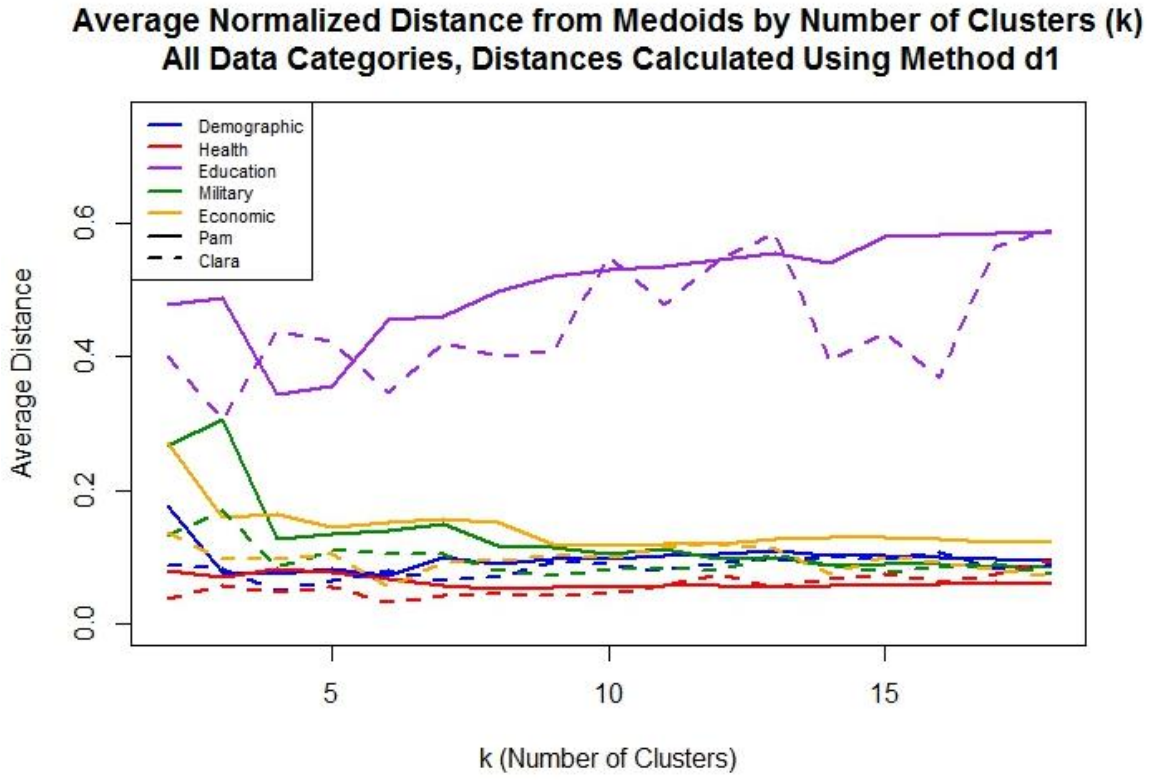
We first determine which clustering algorithm to use. As a first step, we determine the difference between "pam" and "clara." Both methods use the same algorithm but "clara" uses only a sample of the data to create the final cluster assignments. This is helpful with large datasets since the computation time for "clara" does not increase with more observations, but the computation time for "pam" does. Figure 5 shows the difference in computation time for clustering methods "pam" and "clara" for each data category and k-value using dissimilarity matrix calculation method *d1* (*d3* and *d4* yield similar results). With a high performance computing cluster we are able to perform these operations in parallel so the total computation time is equal to the longest time required by the slowest of the 170 models.



TreeClust models are fit using dissimilarity calculation method *d1*, and clustering algorithms "pam" and "clara."

Figure 5. Computation Time by Number of Clusters for All Data Categories.

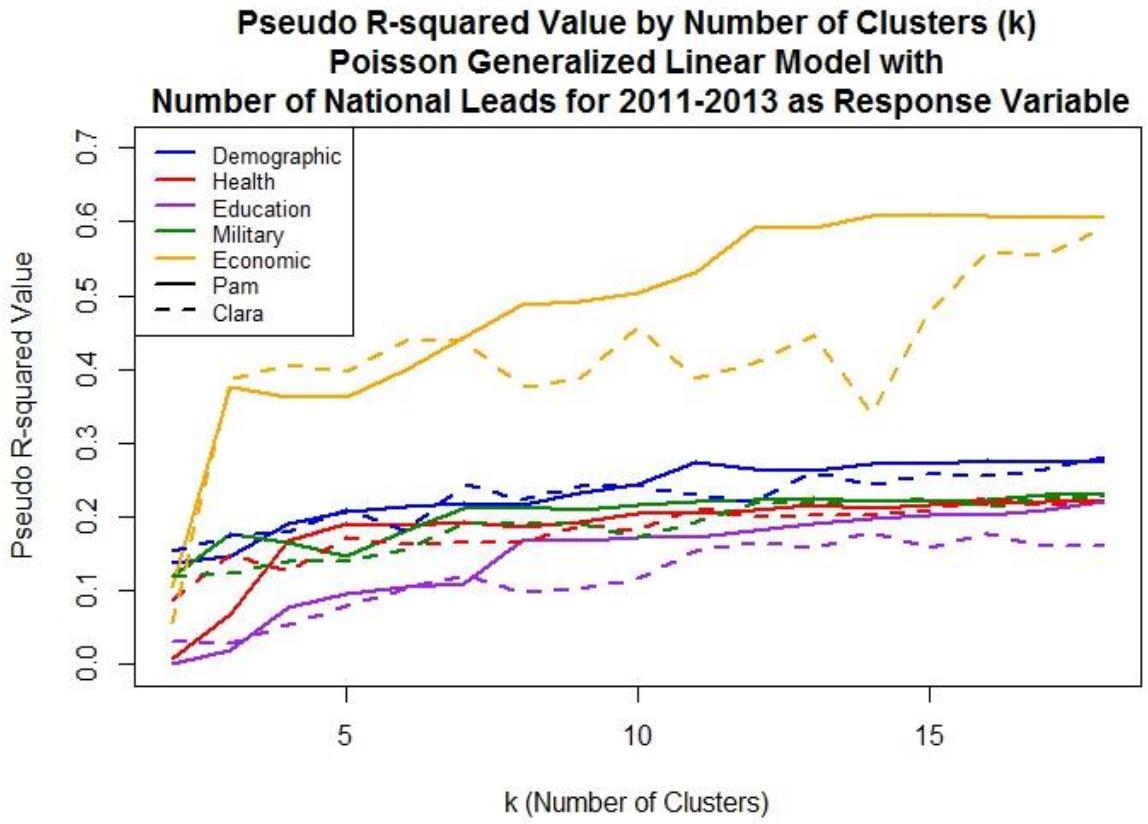
Since we see the increased cost in computation time, we need to determine if we lose anything by using "clara" instead of "pam." One metric is the average distances between the observations and the medoid of the cluster they are assigned to. This metric is returned as a part of the *treeClust* object. Figure 6 shows that the average distances between observations and medoids are similar but not equal when final clusters are assigned by "pam" or "clara." Smaller average distances would indicate more similar clustering of ZIP codes in multi-dimensional space.



All *treeClust* models are fit using dissimilarity calculation method *d1*.

Figure 6. Average Distances from Medoids to Observations Assigned to Clusters for All Data Categories.

Because average distances between observations and cluster medoids differ when using different clustering algorithms, we favor "pam," which uses all of the data, over "clara" at the expense of the extra computational burden.



All *treeClust* cluster assignments are fit using dissimilarity calculation method *d1*.

Figure 7. Pseudo R-squared Values by Number of Clusters from Poisson Generalized Linear Models.

Since high performance computing is not universally available, it is important to note that there could be some loss in predictive ability by using "clara" as a clustering algorithm. Figure 7 shows the pseudo R-squared models fit to predict the number of leads. From Figure 7, we see that "pam" generally has a higher pseudo R-squared value. We compare the "pam" models in a similar way with models fit with "k-means" and find that "k-means" outperforms "pam," although the difference is not great.

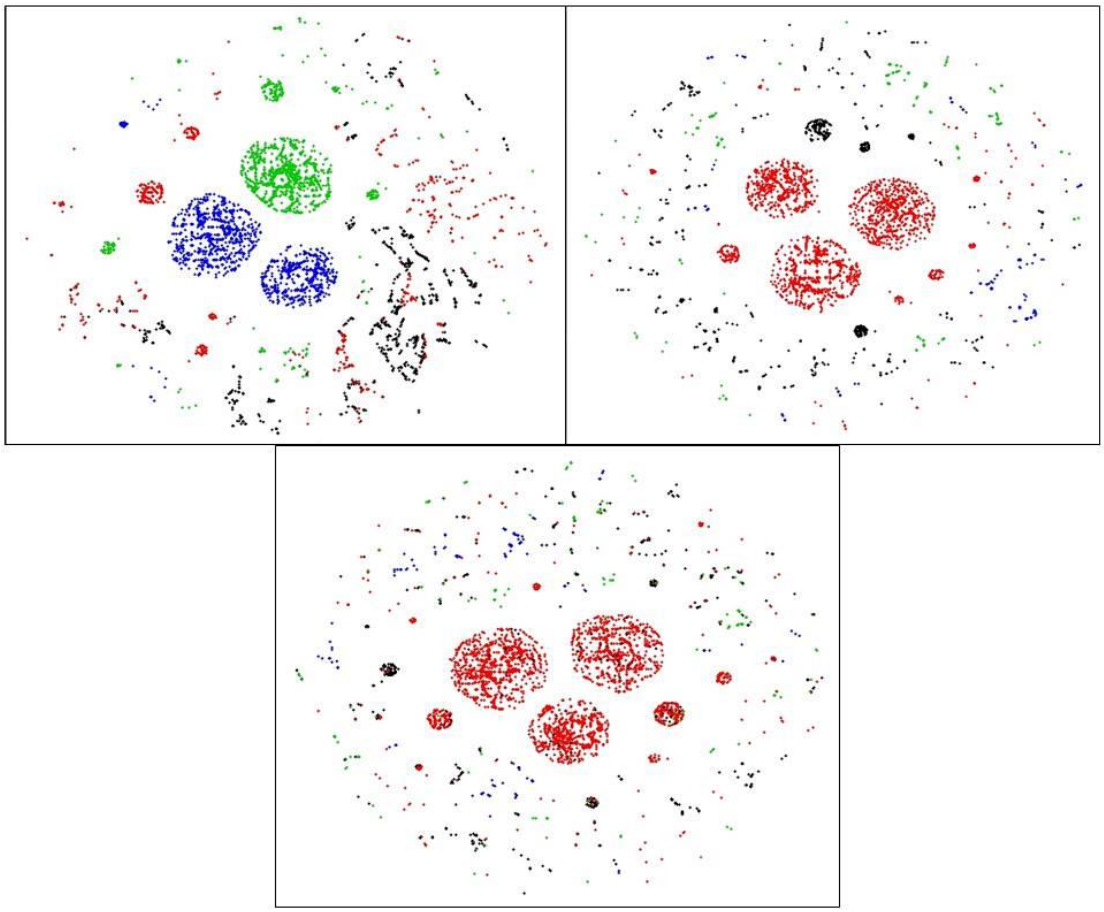
2. Comparing Dissimilarity Calculation Methods

Now that we have determined that clusters created by "k-means" outperform those created by "pam" and "clara," we need to explore the differences between the dissimilarity calculation methods. Comparing pseudo R-squared values from all models shows that *d3* consistently outperforms *d1* and *d4*. The *d3* method is the method that does

not weight the quality of each variable tree but does not treat the distances between leaves the same. Instead it rates the distances between leaves differently depending on how far apart they are on the tree.

In addition to using the pseudo R-squared values, we can visualize the multi-dimensional distances between observations in two or three dimensions using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm as described by Van der Maaten and Hinton (2008). We compare the distances between observations using education data by mapping the multi-dimensional space to two dimensions. We use the education category for our comparison because it has the fewest variables and is the easiest to compare when viewed in two dimensions.

Figure 8 shows ZIP code distances in two dimensions where the multi-dimensional distances are computed using methods *d1*, *d3*, and *d4*. Colors represent clusters of ZIP codes found using the algorithm "pam" to cluster ZIP codes into four clusters. They are plotted using the *Rtsne* package for R (Krijthe 2015). They show that when using *d3*, the distances are close enough that ZIP codes are classified into the same category. With *d1*, the visually close observations in the center are categorized into two different clusters and in *d4*, they are all categorized the same, but there are several observations interspersed that are categorized separately.



These plots are created using visualization methods described by Maaten and Hinton (2008). We map the *d1*, *d3*, and *d4* multi-dimensional distances to two dimensions using R Package *Rtsne* in the first, second, and third plots respectively (Krijthe 2015). Colors indicate *treeClust* cluster assignments created from these distances using "pam" and four clusters.

Figure 8. Two-Dimensional Plots of ZIP Codes Using Education Data.

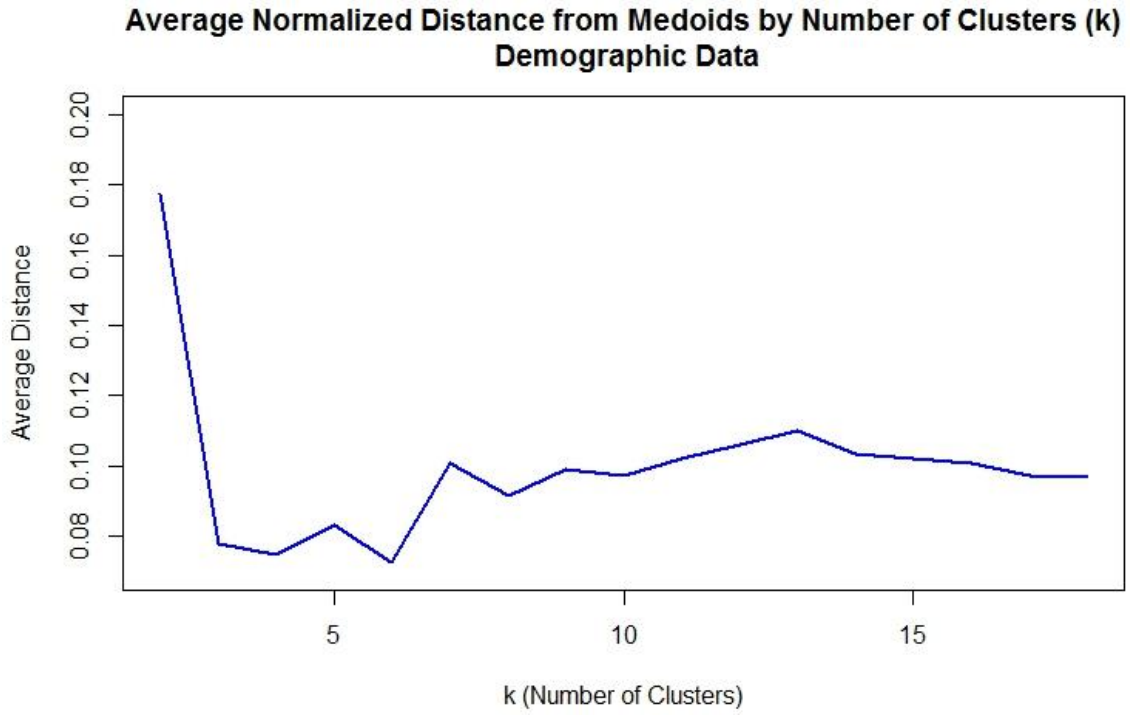
Now that we have selected the highest performing distance calculation method and clustering algorithm, we explore the performance of the different number of clusters.

3. Comparing K-Values

For a market analysis tool, a number of clusters between 4 and 8 may be ideal. With only three clusters, a commander will see the difference between cities, suburbs and rural areas, which is information that they already know and is therefore not helpful. A large number of clusters is also not useful for comparing ZIP codes, due to the complexity. We explore the use of internal *treeClust* metrics and of using the pseudo R-

squared values calculated from Poisson regressions fit using the cluster assignments to determine a reasonable number of clusters.

Using the internal *treeClust* metrics for clustering algorithms "pam" and "clara," the optimal number of clusters is the one with the smallest average distance between the cluster medoid—the multi-dimensional center of the cluster—and the observations assigned to that cluster. These distances are normalized on a scale of 0 to 1. Figure 9 shows the average distances from the medoid of each cluster to the observations assigned to that cluster for each k-value for the demographic data. This figure shows that six clusters give the smallest average distances for the demographic data when using dissimilarity calculation method *d1* and clustering algorithm "pam."

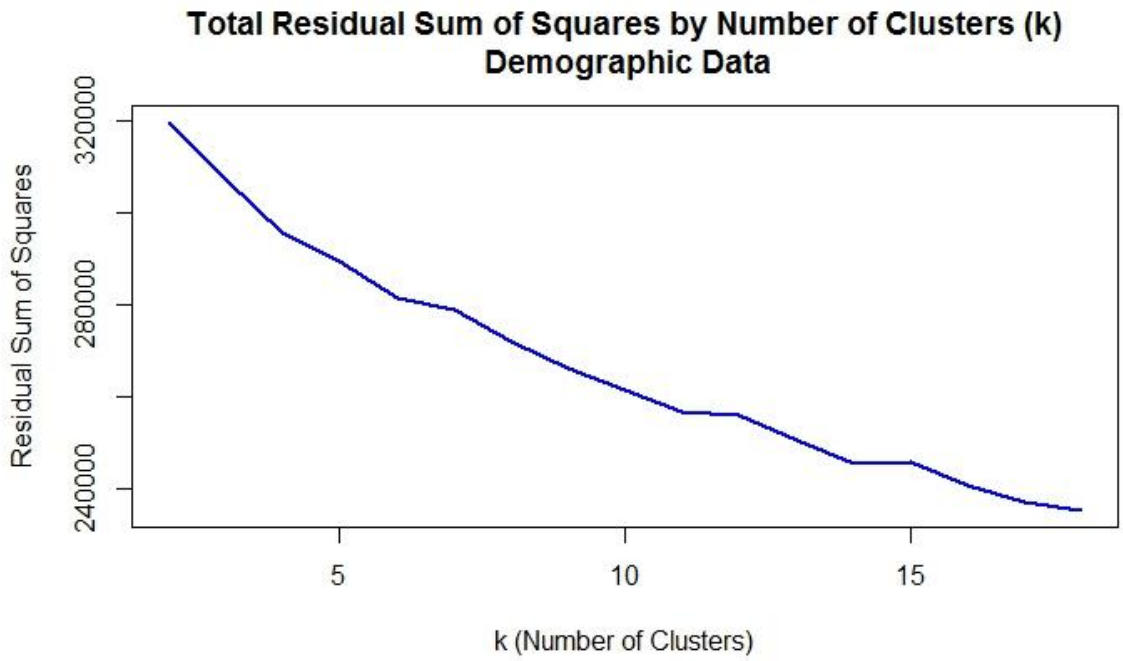


TreeClust models are fit using dissimilarity calculation method *d1* and clustering algorithm "pam."

Figure 9. Average Distance by Number of Clusters for Demographic Data.

"K-means" calls for a similar method to determine the optimal cluster size. The only difference is that instead of using smallest average distance, we find the "knee in the

curve," the value of k when the residual sum of squares tends to level off. The effect is the same, although the scale is different. Figure 10 shows the total residual sum of squares for each k-value using the same dissimilarity calculation method and data type as Figure 5, but after using "k-means" for the clustering algorithm. This graph is representative of all of the graphs that use "k-means" for the clustering algorithm in that they universally favor larger k-values. That is, the larger the k-value, the smaller the residual sum of squares.



TreeClust models are fit using dissimilarity calculation method *d1* and clustering algorithm "k-means."

Figure 10. Total Residual Sum of Squares by Number of Clusters for Demographic Data.

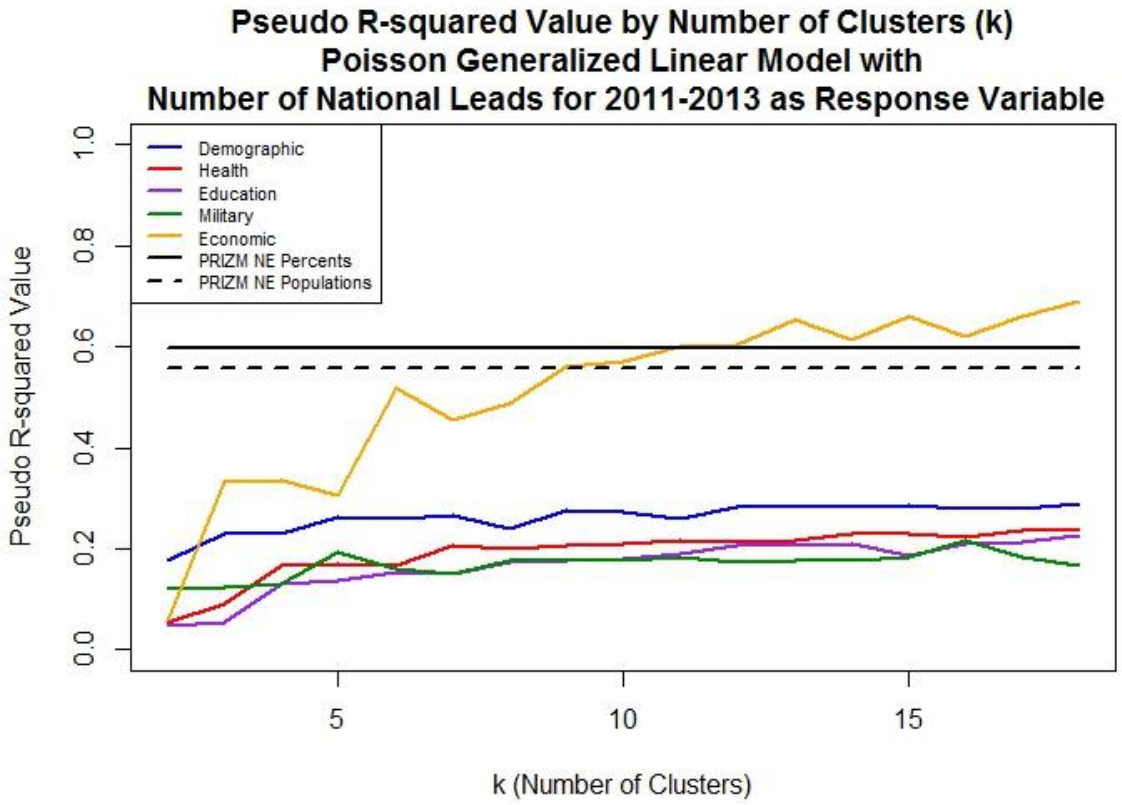
Since this first metric does not allow for direct comparisons between "k-means" and "pam" we also use the pseudo R-squared value of Poisson regression models fit with the cluster assignments created by the various combinations of distance calculation methods and clustering algorithms. Since a smaller number of clusters is more useful for market analysis, we compare the different k-values using the best performing combination of distance calculation methods and clustering algorithms, *d3* and "k-

means." We find that we can achieve a pseudo R-squared value of .6884. In the next section, we compare the pseudo R-squared models for Poisson regressions fit with these models and with models fit using the PRIZM NE data. We compare the regression models fit with the PRIZM NE data to models fit with different data categories and numbers of final clusters.

D. COMPARING CLUSTERS TO PRIZM NE

We use two different methods to fit models using the Nielsen PRIZM NE data. For the first method, we fit a Poisson generalized linear model using the estimated PRIZM NE population size of each segment in each ZIP code as predictors and again use the number of national leads from 2011–2013 as the response variable (USAREC 2015c). For the second model we replace the segment population sizes for each ZIP code with percentages of that ZIP code's population in each PRIZM NE segment.

Using the PRIZM NE population estimates as predictors yields a pseudo R-squared value of .5569, while the percentages yield a pseudo R-squared value of .5976. Since the percentages yield a higher R-squared value, we use this model to predict the 2014 test set's national leads. Figure 11 shows the different pseudo R-squared values for models fit with dissimilarity calculation method *d3* and clustering algorithm "k-means." With 11 clusters, the GLM using the economic data results in a pseudo R-squared value slightly better than GLMs using the PRIZM NE segment data.



TreeClust cluster assignments are fit with dissimilarity calculation method *d3* and clustering algorithm "k-means."

Figure 11. Pseudo R-squared Values for Generalized Linear Models with PRIZM NE Segments and *treeClust* Cluster Assignments.

We find that we can increase the pseudo R-squared value from .6884 to .7217 by including the cluster assignments created by both the economic and health data or to .7570 by including cluster assignments for all five data types as predictors. However, by increasing the number of cluster categories, we are also increasing the number of levels which makes the ZIP code clusters less interpretable. For example, if we attempt to use the 11 different clusters of economic data and 11 different clusters of health data, we now have 121 different market segments for our ZIP codes. This would have a similar result for recruiting decision makers as increasing the total number of clusters from our maximum of 18. To test predictive ability of the GLMs we use the cluster assignments created from economic data with k-values of 11 and 18 to see if the increased pseudo R-squared value results in increased predictive ability.

When predicting with the Poisson regression models fit from cluster assignments created using dissimilarity calculation method *d3*, final algorithm "k-means," and 18 clusters there is a median difference of 1.9 between the predictions and the actual 2014 leads. This outperforms the PRIZM NE model predictions which have a median difference of 3.4 from the actual 2014 national leads.

E. COMBINING MODELS

The ideal Tree Clustering model would return a single set of cluster assignments that are informed by all datasets. Ideally this model would have between 4 and 6 different clusters to maximize interpretability while still having high predictive power. We use four different methods to attempt to create a single set of cluster assignments that incorporate all data. For the first method, we fit Tree Clustering models with all 347 variables at the same time. We next attempt to use the five category cluster assignments with 18 clusters each and cluster again using intermediate cluster assignments as factors. We use the cluster assignments from each category with whichever k-value has the smallest average distance from medoids as factors to create a final cluster.

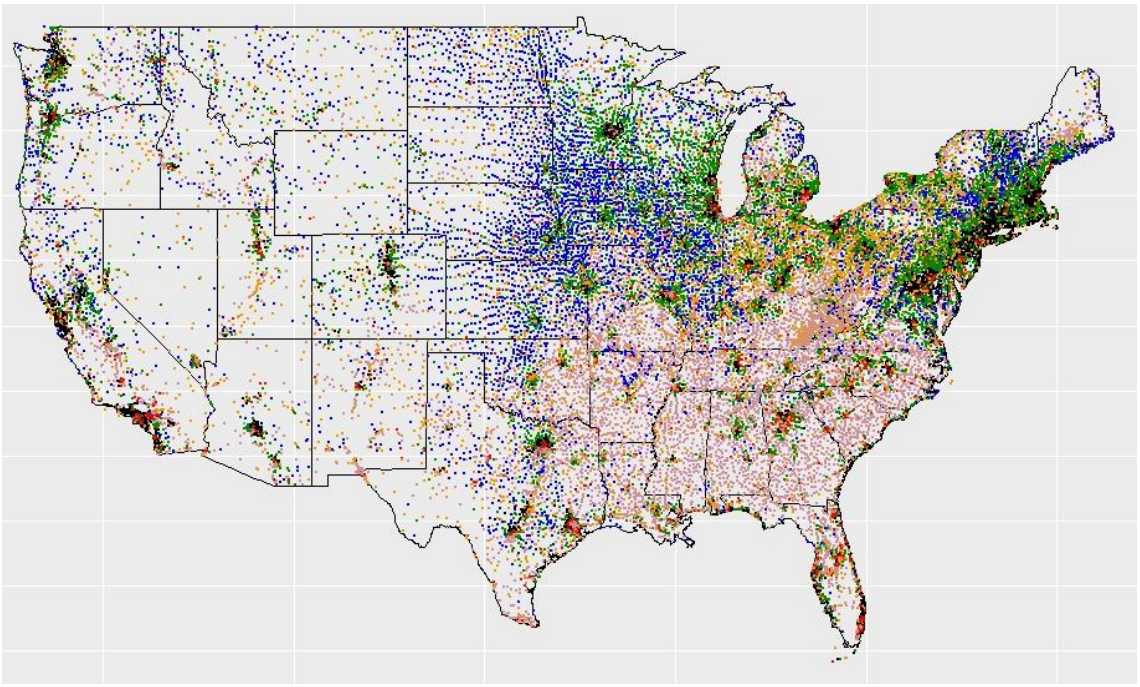
Finally, we attempt to total the distance matrices for all five categories, then use "pam" to cluster the total distances. We use these cluster assignments created by these four methods in separate GLMs using the number of 2011–2013 national leads as response variables. The highest pseudo R-squared value achieved by these methods is found by totaling the distance matrices and clustering with "pam." This results in a pseudo R-squared value of .341, which is far below .6884, the value achieved by only using the cluster assignments generated with economic data.

V. CONCLUSION AND RECOMMENDATIONS

A. SUMMARY OF RESULTS

This study concludes that models built with cluster assignments created by using the R package *treeClust* from publicly available economic data out-predict those created with PRIZM NE data. Tree Clustering models based on economic data have more predictive power than models based on the other data categories or from using a combination of all categories in a single model. This allows USAREC to replace the 66 variables of the PRIZM NE data with a single cluster assignment for each ZIP code for use in market analysis. Using a single cluster assignment allows recruiters to gain a better understanding of their operational environment.

Figure 12 shows ZIP code cluster assignments from economic data, using dissimilarity calculation method *d3*, clustering algorithm "k-means," and six clusters. Each dot represents one of 34,007 ZIP codes. The colors represent cluster assignments. We do not attempt to characterize the clusters at this point. We use dots to represent ZIP codes instead of shading so the geographically larger ZIP codes in the western United States do not appear to represent a larger proportion of the nation. This figure shows that cities, suburbs, and rural areas tend to be similar. It also shows significant regional effects with ZIP codes tending to be similar across the South, the West, or the Midwest. In larger cities—such as San Francisco—there are many clusters represented in a small area. This is where the clusters can be more useful for market analysis since they allow recruiting commanders to see the different types of ZIP codes within their area of operations.



TreeClust cluster assignments are fit with dissimilarity calculation method *d3* and clustering algorithm "k-means." Map created using R Packages *ggplot2* (Wickham and Chang 2016) and *ggmap* (Kahle and Wickham 2016).

Figure 12. Map Showing Cluster Assignments with Six Clusters.

To use these cluster assignments for prediction, a larger number of clusters and a combination of clusters for multiple data categories increases predictive power, although it also increases the number of variables per ZIP code. We find that combining the data categories into a single set of cluster assignments did not result in clusters that are as predictive as using the economic data by itself. Additionally, we only test the predictive power of the clusters against a single response variable, the number of national leads. Testing these cluster assignments against different recruiting metrics could provide more insight into their suitability for use in recruiter assignment and goal allocation models.

B. RECOMMENDATIONS FOR FUTURE WORK

The first priority for future work is testing the predictive ability of the economic and other clusters against other response variables. The second priority is finding a method to combine all data types to create clusters that still outperform the commercially

sourced data. Other avenues of exploration are clustering the rest of the recruiting ZIP codes and incorporating additional or updated data sources into the cluster models.

As discussed in Chapter II, there are several other metrics used by USAREC to judge a market's ability to produce recruits. These include the number of recruits produced in previous years, the propensity of the population to military recruiting, and the number of local leads (Fleischmann and Nelson 2014). However, all of these metrics are influenced by the past efforts and numbers of recruiters (Pinelas 2011). This means that models attempting to determine the impact of the clusters are more complex because they have to adjust for the influence of past recruiting efforts. Any model that can predict the future number of accessions produced by a ZIP code independent of the effort of recruiters is especially valuable.

We use 347 different variables but find that we can only achieve a pseudo R-squared value of .3411 by any of the four methods we use to combine all of our data into a single set of cluster assignments to predict national leads. By using cluster assignments for the economic data alone, we achieve a pseudo R-squared value of .7176 when using cluster assignments from the 137 variables of economic data. There may be other methods or combinations of variables that allow for better prediction when incorporating more variables into our final models.

We only cluster 34,007 ZIP codes because these are the ZIP codes that are CONUS and have sufficient population data available. USAREC assigns 41,699 ZIP codes to its recruiting stations (United States Army Recruiting Command 2015b). Many of these additional ZIP codes belong to businesses and national parks, or post office boxes with no recruiting populations. However, some of them belong to universities and to communities in Alaska, Hawaii, and United States Territories that do have recruiting populations. A future researcher can collect more data to include these additional ZIP codes to also determine which ZIP codes have no recruiting populations. This allows USAREC to only attempt to recruit from ZIP codes that do have recruiting populations.

A future researcher can also fit models with additional data such as the FBI crime data that Intrater (2015) uses in his study. By using the county to ZIP translation method

that we use here, a future researcher incorporates additional county-level data into the model, such as more detailed unemployment rates, literacy rates, and votes by political party in national elections. These additional datasets allow a researcher to further refine these models to result in better prediction.

C. CONCLUSION

This study shows that it is possible to use publicly available economic data to create clusters of ZIP codes that are able to predict much of the deviance in national leads between ZIP codes. Using cluster assignments created from this data using *treeClust*, we fit predictive models that outperform models that use proprietary data. When we fit a Poisson regression model using the economic cluster assignments created with distance calculation method *d3*, clustering algorithm "k-means," and 18 final clusters, we can account for 69% of the deviance in national leads. This Poisson regression model can predict leads with more accuracy than regression models created using proprietary data.

USAREC currently uses market segmentation with 66 different clusters per ZIP code. The single cluster assignments we create provide a more understandable—and more useful—tool for recruiting commanders to use for market analysis. By replacing 66 variables per ZIP code with a single cluster assignment, recruiting commanders can better understand the different types of markets that they are responsible for recruiting from. USAREC can use these cluster assignments to predict the each ZIP code's potential to produce recruits, allowing them to better assign and task recruiters.

APPENDIX A. VARIABLES USED

We translate all of the datasets described in Chapter III, using methods described in Appendix B. The results are the 347 variables listed in Tables 2 through 6.

Table 2. Demographic Data Variable Names, Sources, and Descriptions.

Variable Name	Source	Description
Population_Size	Community Health Status Indicators (CHSI)	Estimated population
Population_Density	CHSI	Population density
Age_19_Under	CHSI	Percentage of population under 19
Age_19_64	CHSI	Percentage of population between 19 and 64
Age_65_84	CHSI	Percentage of population between 65 and 84
Age_85_and_Over	CHSI	Percentage of population 85 and older
White	CHSI	Percentage of population classified as White
Black	CHSI	Percentage of population classified as Black
Native_American	CHSI	Percentage of population classified as Native American
Asian	CHSI	Percentage of population classified as Asian
Hispanic	CHSI	Percentage of population classified as Hispanic
MARS1	Income Tax	Percentage of returns filed single
MARS2	Income Tax	Percentage of returns filed joint
MARS4	Income Tax	Percentage of returns filed head of household

Table 3. Economic Data Variable Names, Sources, and Descriptions.

Variable Name	Source	Description
N1	Income Tax	Number of returns
PREP	Income Tax	Percentage of returns with paid preparer's signature
N2	Income Tax	Average number of exemptions
NUMDEP	Income Tax	Average number of dependents
A00100	Income Tax	Average adjusted gross income (AGI)
N02650	Income Tax	Percentage of returns with total income
A02650	Income Tax	Average Total income amount
N00200	Income Tax	Percentage of returns with salaries and wages
A00200	Income Tax	Average salaries and wages amount
N00300	Income Tax	Percentage of returns with taxable interest

Variable Name	Source	Description
A00300	Income Tax	Average Taxable interest amount
N00600	Income Tax	Percentage of returns with ordinary dividends
A00600	Income Tax	Average Ordinary dividends amount
N00650	Income Tax	Percentage of returns with qualified dividends
A00650	Income Tax	Average Qualified dividends amount [3]
N00700	Income Tax	Percentage of returns with state and local income tax refunds
A00700	Income Tax	Average State and local income tax refunds amount
N00900	Income Tax	Percentage of returns with business or professional net income (less loss)
A00900	Income Tax	Average Business or professional net income (less loss) amount
N01000	Income Tax	Percentage of returns with net capital gain (less loss)
A01000	Income Tax	Average Net capital gain (less loss) amount
N01400	Income Tax	Percentage of returns with taxable individual retirement arrangements distributions
A01400	Income Tax	Average Taxable individual retirement arrangements distributions amount
N01700	Income Tax	Percentage of returns with taxable pensions and annuities
A01700	Income Tax	Average Taxable pensions and annuities amount
SCHF	Income Tax	Number of farm returns
N02300	Income Tax	Percentage of returns with unemployment compensation
A02300	Income Tax	Average Unemployment compensation amount [4]
N02500	Income Tax	Percentage of returns with taxable Social Security benefits
A02500	Income Tax	Average Taxable Social Security benefits amount
N26270	Income Tax	Percentage of returns with partnership/S-corp net income (less loss)
A26270	Income Tax	Average Partnership/S-corp net income (less loss) amount
N02900	Income Tax	Percentage of returns with total statutory adjustments
A02900	Income Tax	Average Total statutory adjustments amount
N03220	Income Tax	Percentage of returns with educator expenses
A03220	Income Tax	Average Educator expenses amount

Variable Name	Source	Description
N03300	Income Tax	Percentage of returns with self-employment retirement plans
A03300	Income Tax	Average Self-employment retirement plans amount
N03270	Income Tax	Percentage of returns with self-employment health insurance deduction
A03270	Income Tax	Average Self-employment health insurance deduction amount
N03150	Income Tax	Percentage of returns with IRA payments
A03150	Income Tax	Average IRA payments amount
N03210	Income Tax	Percentage of returns with student loan interest deduction
A03210	Income Tax	Average Student loan interest deduction amount
N03230	Income Tax	Percentage of returns with tuition and fees deduction
A03230	Income Tax	Average Tuition and fees deduction amount
N03240	Income Tax	Returns with domestic production activities deduction
A03240	Income Tax	Average Domestic production activities deduction amount
N04470	Income Tax	Percentage of returns with itemized deductions
A04470	Income Tax	Average Total itemized deductions amount
N00101	Income Tax	Percentage of returns itemized
A00101	Income Tax	Average amount of AGI for itemized returns
N18425	Income Tax	Percentage of returns with State and local income taxes
A18425	Income Tax	Average state and local income taxes amount
N18450	Income Tax	Percentage of returns with State and local general sales tax
A18450	Income Tax	Average state and local general sales tax amount
N18500	Income Tax	Percentage of returns with real estate taxes
A18500	Income Tax	Average real estate taxes amount
N18300	Income Tax	Percentage of returns with taxes paid
A18300	Income Tax	Average taxes paid amount
N19300	Income Tax	Percentage of returns with mortgage interest paid
A19300	Income Tax	Average mortgage interest paid amount
N19700	Income Tax	Percentage of returns with contributions
A19700	Income Tax	Average contributions amount

Variable Name	Source	Description
N04800	Income Tax	Percentage of returns with taxable income
A04800	Income Tax	Average taxable income amount
N05800	Income Tax	Percentage of returns with income tax before credits
A05800	Income Tax	Average income tax before credits amount
N09600	Income Tax	Percentage of returns with alternative minimum tax
A09600	Income Tax	Average alternative minimum tax amount
N07100	Income Tax	Percentage of returns with total tax credits
A07100	Income Tax	Average tax credits amount
N07300	Income Tax	Percentage of returns with foreign tax credit
A07300	Income Tax	Average foreign tax credit amount
N07180	Income Tax	Percentage of returns with child and dependent care credit
A07180	Income Tax	Average child and dependent care credit amount
N07230	Income Tax	Percentage of returns with nonrefundable education credit
A07230	Income Tax	Average nonrefundable education credit amount
N07240	Income Tax	Percentage of returns with retirement savings contribution credit
A07240	Income Tax	Average retirement savings contribution credit amount
N07220	Income Tax	Percentage of returns with child tax credit
A07220	Income Tax	Average child tax credit amount
N07260	Income Tax	Percentage of returns with residential energy tax credit
A07260	Income Tax	Average residential energy tax credit amount
N09400	Income Tax	Percentage of returns with self-employment tax
A09400	Income Tax	Average self-employment tax amount
N10600	Income Tax	Percentage of returns with total tax payments
A10600	Income Tax	Average total tax payments amount
N59660	Income Tax	Percentage of returns with earned income credit
A59660	Income Tax	Average earned income credit amount
N59720	Income Tax	Percentage of returns with excess earned income credit
A59720	Income Tax	Average excess earned income credit (refundable) amount
N11070	Income Tax	Percentage of returns with additional child tax

Variable Name	Source	Description
		credit
A11070	Income Tax	Average additional child tax credit amount
N10960	Income Tax	Percentage of returns with refundable education credit
A10960	Income Tax	Average refundable education credit amount
N06500	Income Tax	Percentage of returns with income tax
A06500	Income Tax	Average income tax amount
N10300	Income Tax	Percentage of returns with tax liability
A10300	Income Tax	Average total tax liability amount
N85530	Income Tax	Percentage of returns with Additional Medicare tax
A85530	Income Tax	Average additional Medicare tax
N85300	Income Tax	Percentage of returns with net investment income tax
A85300	Income Tax	Average net investment income tax
N11901	Income Tax	Percentage of returns with tax due at time of filing
A11901	Income Tax	Average tax due at time of filing amount
N11902	Income Tax	Percentage of returns with overpayments refunded
A11902	Income Tax	Average overpayments refunded amount
under25K	Income Tax	Percent of returns under $25,000 AGI
twentyfiveto50	Income Tax	Percent of returns between $25K and 50K AGI
fiftyto75k	Income Tax	Percent of returns between $50K and 75K AGI
seventyfiveto100k	Income Tax	Percent of returns between $75K and 100K AGI
hundredto200k	Income Tax	Percent of returns between $100K and 200K AGI
over200k	Income Tax	Percent of returns over $200,000 AGI
ESTAB	County Business Patterns	Number of establishments
EMP	County Business Patterns	Paid employees for pay period including March 12
PAYQTR1	County Business Patterns	Total first-quarter payroll ($1,000)
PAYANN	County Business	Total annual payroll ($1,000)

Variable Name	Source	Description
	Patterns	
Retail	Economic Census	Number of retail establishments in ZIP code
ProfSciTech	Economic Census	Number of science and technology establishments in ZIP code
AdminWasteMgmt	Economic Census	Number of administration and waste management establishments in ZIP code
Education	Economic Census	Number of education establishments in ZIP code
HealthCare	Economic Census	Number of health care establishments in ZIP code
ArtsRec	Economic Census	Number of arts and recreation establishments in ZIP code
Hospitality	Economic Census	Number of hospitality establishments in ZIP code
Other	Economic Census	Number of other establishments in ZIP code
Total	Economic Census	Total number of establishments in ZIP code
RetailPerc	Economic Census	Percentage of retail establishments in ZIP code
ProfSciTechPerc	Economic Census	Percentage of science and technology establishments in ZIP code
AdminWasteMgmtPerc	Economic Census	Percentage of administration and waste management establishments in ZIP code
EducationPerc	Economic Census	Percentage of education establishments in ZIP code
HealthCarePerc	Economic Census	Percentage of health care establishments in ZIP code
ArtsRecPerc	Economic Census	Percentage of arts and recreation establishments in ZIP code
HospitalityPerc	Economic Census	Percentage of hospitality establishments in ZIP code
OtherPerc	Economic Census	Percentage of other establishments in ZIP code

Table 4. Education Data Variable Names, Sources, and Descriptions.

Variable Name	Source	Description
No_HS_Diploma	CHSI	Rate, no high school diploma (among adults age 25 and older)
Size1Dist1	Secondary Education	Number of education institutions of size 1 within 10 miles of ZIP code

Size2Dist1	Secondary Education	Number of education institutions of size 2 within 10 miles of ZIP code
Size3Dist1	Secondary Education	Number of education institutions of size 3 within 10 miles of ZIP code
Size4Dist1	Secondary Education	Number of education institutions of size 4 within 10 miles of ZIP code
Size5Dist1	Secondary Education	Number of education institutions of size 5 within 10 miles of ZIP code
Size1Dist2	Secondary Education	Number of education institutions of size 1 within 50 miles of ZIP code
Size2Dist2	Secondary Education	Number of education institutions of size 2 within 50 miles of ZIP code
Size3Dist2	Secondary Education	Number of education institutions of size 3 within 50 miles of ZIP code
Size4Dist2	Secondary Education	Number of education institutions of size 4 within 50 miles of ZIP code
Size5Dist2	Secondary Education	Number of education institutions of size 5 within 50 miles of ZIP code

Table 5. Health Data Variable Names, Sources, and Descriptions.

Variable Name	**Source**	**Description**
A_Wh_Comp	Community Health Status Indicators (CHSI)	Rate, under age 1, complications of pregnancy/birth, White
A_Bl_Comp	CHSI	Rate, under age 1, complications of pregnancy/birth, Black
A_Ot_Comp	CHSI	Rate, under age 1, complications of pregnancy/birth, other
A_Hi_Comp	CHSI	Rate, under age 1, complications of pregnancy/birth, Hispanic
A_Wh_BirthDef	CHSI	Rate, under age 1, birth defects, White
A_Bl_BirthDef	CHSI	Rate, under age 1, birth defects, Black
A_Ot_BirthDef	CHSI	Rate, under age 1, birth defects, other
A_Hi_BirthDef	CHSI	Rate, under age 1, birth defects, Hispanic
B_Wh_Injury	CHSI	Rate, ages 1–14, injuries, White
B_Bl_Injury	CHSI	Rate, ages 1–14, injuries, Black
B_Ot_Injury	CHSI	Rate, ages 1–14, injuries, other
B_Hi_Injury	CHSI	Rate, ages 1–14, injuries, Hispanic
B_Wh_Cancer	CHSI	Rate, ages 1–14, cancer, White
B_Bl_Cancer	CHSI	Rate, ages 1–14, cancer, Black
B_Ot_Cancer	CHSI	Rate, ages 1–14, cancer, other

Variable Name	Source	Description
B_Hi_Cancer	CHSI	Rate, ages 1–14, cancer, Hispanic
B_Wh_Homicide	CHSI	Rate, ages 1–14, homicide, White
B_Bl_Homicide	CHSI	Rate, ages 1–14, homicide, Black
B_Ot_Homicide	CHSI	Rate, ages 1–14, homicide, other
B_Hi_Homicide	CHSI	Rate, ages 1–14, homicide, Hispanic
C_Wh_Injury	CHSI	Rate, ages 15–24, injuries, White
C_Bl_Injury	CHSI	Rate, ages 15–24, injuries, Black
C_Ot_Injury	CHSI	Rate, ages 15–24, injuries, other
C_Hi_Injury	CHSI	Rate, ages 15–24, injuries, Hispanic
C_Wh_Homicide	CHSI	Rate, ages 15–24, homicide, White
C_Bl_Homicide	CHSI	Rate, ages 15–24, homicide, Black
C_Ot_homicide	CHSI	Rate, ages 15–24, homicide, other
C_Hi_Homicide	CHSI	Rate, ages 15–24, homicide, Hispanic
C_Wh_Suicide	CHSI	Rate, ages 15–24, suicide, White
C_Bl_Suicide	CHSI	Rate, ages 15–24, suicide, Black
C_Ot_Suicide	CHSI	Rate, ages 15–24, suicide, other
C_Hi_Suicide	CHSI	Rate, ages 15–24, suicide, Hispanic
C_Wh_Cancer	CHSI	Rate, ages 15–24, cancer, White
C_Bl_Cancer	CHSI	Rate, ages 15–24, cancer, Black
C_Ot_Cancer	CHSI	Rate, ages 15–24, cancer, other
C_Hi_Cancer	CHSI	Rate, ages 15–24, cancer, Hispanic
D_Wh_Injury	CHSI	Rate, ages 25–44, injuries, White
D_Bl_Injury	CHSI	Rate, ages 25–44, injuries, Black
D_Ot_Injury	CHSI	Rate, ages 25–44, injuries, other
D_Hi_Injury	CHSI	Rate, ages 25–44, injuries, Hispanic
D_Wh_Cancer	CHSI	Rate, ages 25–44, cancer, White
D_Bl_Cancer	CHSI	Rate, ages 25–44, cancer, Black
D_Ot_Cancer	CHSI	Rate, ages 25–44, cancer, other
D_Hi_Cancer	CHSI	Rate, ages 25–44, cancer, Hispanic
D_Wh_HeartDis	CHSI	Rate, ages 25–44, heart disease, White
D_Bl_HeartDis	CHSI	Rate, ages 25–44, heart disease, Black
D_Ot_HeartDis	CHSI	Rate, ages 25–44, heart disease, other
D_Hi_HeartDis	CHSI	Rate, ages 25–44, heart disease, Hispanic
D_Wh_Suicide	CHSI	Rate, ages 25–44, suicide, White
D_Bl_Suicide	CHSI	Rate, ages 25–44, suicide, Black
D_Ot_Suicide	CHSI	Rate, ages 25–44, suicide, other
D_Hi_Suicide	CHSI	Rate, ages 25–44, suicide, Hispanic
D_Wh_HIV	CHSI	Rate, ages 25–44, hiv/aids, White
D_Bl_HIV	CHSI	Rate, ages 25–44, hiv/aids, Black

Variable Name	Source	Description
D_Hi_HIV	CHSI	Rate, ages 25–44, hiv/aids, Hispanic
D_Wh_Homicide	CHSI	Rate, ages 25–44, homicide, White
D_Bl_Homicide	CHSI	Rate, ages 25–44, homicide, Black
D_Ot_Homicide	CHSI	Rate, ages 25–44, homicide, other
D_Hi_Homicide	CHSI	Rate, ages 25–44, homicide, Hispanic
E_Wh_Cancer	CHSI	Rate, ages 45–64, cancer, White
E_Bl_Cancer	CHSI	Rate, ages 45–64, cancer, Black
E_Ot_Cancer	CHSI	Rate, ages 45–64, cancer, other
E_Hi_Cancer	CHSI	Rate, ages 45–64, cancer, Hispanic
E_Wh_HeartDis	CHSI	Rate, ages 45–64, heart disease, White
E_Bl_HeartDis	CHSI	Rate, ages 45–64, heart disease, Black
E_Ot_HeartDis	CHSI	Rate, ages 45–64, heart disease, other
E_Hi_HeartDis	CHSI	Rate, ages 45–64, heart disease, Hispanic
F_Wh_HeartDis	CHSI	Rate, ages 65+, heart disease, White
F_Bl_HeartDis	CHSI	Rate, ages 65+, heart disease, Black
F_Ot_HeartDis	CHSI	Rate, ages 65+, heart disease, other
F_Hi_HeartDis	CHSI	Rate, ages 65+, heart disease, Hispanic
F_Wh_Cancer	CHSI	Rate, ages 65+, cancer, White
F_Bl_Cancer	CHSI	Rate, ages 65+, cancer, Black
F_Ot_Cancer	CHSI	Rate, ages 65+, cancer, other
F_Hi_Cancer	CHSI	Rate, ages 65+, cancer, Hispanic
LBW	CHSI	Rate, birth measures, low birth wt. (<2500 g)
VLBW	CHSI	Rate, birth measures, very low birth wt. (<1500 g)
Premature	CHSI	Rate, birth measures, premature births (<37 weeks)
Under_18	CHSI	Rate, birth measures, births to women under 18
Over_40	CHSI	Rate, birth measures, births to women over 40
Unmarried	CHSI	Rate, birth measures, births to unmarried women
Late_Care	CHSI	Rate, birth measures, no care in first trimester
Infant_Mortality	CHSI	Rate, infant mortality
IM_Wh_Non_Hisp	CHSI	Rate, infant mortality, White non Hispanic
IM_Bl_Non_Hisp	CHSI	Rate, infant mortality, Black non Hispanic

Variable Name	Source	Description
IM_Hisp	CHSI	Rate, infant mortality, Hispanic
IM_Neonatal	CHSI	Rate, infant mortality, neonatal
IM_Postneonatal	CHSI	Rate, infant mortality, post-neonatal
Brst_Cancer	CHSI	Rate, death measures, breast cancer (female)
Col_Cancer	CHSI	Rate, death measures, colon cancer
CHD	CHSI	Rate, death measures, coronary heart disease
Homicide	CHSI	Rate, death measures, homicide
Lung_Cancer	CHSI	Rate, death measures, lung cancer
MVA	CHSI	Rate, death measures, motor vehicle injuries
Stroke	CHSI	Rate, death measures, stroke
Suicide	CHSI	Rate, death measures, suicide
Injury	CHSI	Rate, death measures, unintentional injury
Total_Births	CHSI	Rate, total number of births
Total_Deaths	CHSI	Rate, total number of deaths
FluB_Rpt	CHSI	Rate, Haemophilus Influenzae B reported cases
HepA_Rpt	CHSI	Rate, Hepatitis A reported cases
HepB_Rpt	CHSI	Rate, Hepatitis B reported cases
Meas_Rpt	CHSI	Rate, Measles reported cases
Pert_Rpt	CHSI	Rate, Pertussis reported cases
CRS_Rpt	CHSI	Rate, Congenital Rubella Syndrome reported cases
Syphilis_Rpt	CHSI	Rate, Syphilis reported cases
Pap_Smear	CHSI	Rate, pap smears (18+)
Mammogram	CHSI	Rate, mammography (50+)
Proctoscopy	CHSI	Rate, sigmoidoscopy (50+)
Pneumo_Vax	CHSI	Rate, pneumonia vaccine (65+)
Flu_Vac	CHSI	Rate, flu vaccine (65+)
No_Exercise	CHSI	Rate, no exercise
Few_Fruit_Veg	CHSI	Rate, few fruits/vegetables
Obesity	CHSI	Rate, obesity
High_Blood_Pres	CHSI	Rate, high blood pressure
Smoker	CHSI	Rate, smoker
Diabetes	CHSI	Rate, diabetes
Uninsured	CHSI	Rate, uninsured individuals
Elderly_Medicare	CHSI	Rate, medicare beneficiaries, elderly (age 65+)

Variable Name	Source	Description
Disabled_Medicare	CHSI	Rate, medicare beneficiaries, disabled
Prim_Care_Phys_Rate	CHSI	Rate, primary care physicians per 100,000 pop.
Dentist_Rate	CHSI	Rate, dentists per 100,000 pop.
ALE	CHSI	Rate, average life expectancy
All_Death	CHSI	Rate, all causes of death
Health_Status	CHSI	Rate, self-rated health status
Unhealthy_Days	CHSI	Rate, average number of unhealthy days in past month
Sev_Work_Disabled	CHSI	Rate, severely work disabled
Major_Depression	CHSI	Rate, major depression
Recent_Drug_Use	CHSI	Rate, recent drug users (within past month)
Ecol_Rpt	CHSI	Rate, E.coli reported cases
Salm_Rpt	CHSI	Rate, Salmonella reported cases
Shig_Rpt	CHSI	Rate, Shigella reported cases
Toxic_Chem	CHSI	Rate, toxic chemicals released annually
Carbon_Monoxide_Ind	CHSI	Air quality standard indicator, carbon monoxide
Ozone_Ind	CHSI	Air quality standard indicator, ozone
Particulate_Matter_Ind	CHSI	Air quality standard indicator, particulate matter
Lead_Ind	CHSI	Air quality standard indicator, lead

Table 6. Military Data Variable Names, Sources, and Descriptions.

Variable Name	Description	Type
NearestType	Military Bases	Type of nearest military base
DistToNearest	Military Bases	Distance to nearest military base
PopNearest	Military Bases	Population of nearest military base
Within10	Military Bases	Type of largest military base within 10 miles
Pop10	Military Bases	Population of largest military base within 10 miles
Within50	Military Bases	Type of largest military base within 50 miles
Pop50	Military Bases	Population of largest military base within 50 miles
HC01_EST_VC01	American Community Survey (ACS)	Total; Estimate; Civilian population 18 years and over
HC02_EST_VC01	ACS	Veterans; Estimate; Civilian population

Variable Name	Description	Type
		18 years and over
HC02_EST_VC03	ACS	Veterans; Estimate; PERIOD OF SERVICE—Gulf War (9/2001 or later) veterans
HC02_EST_VC04	ACS	Veterans; Estimate; PERIOD OF SERVICE—Gulf War (8/1990 to 8/2001) veterans
HC02_EST_VC05	ACS	Veterans; Estimate; PERIOD OF SERVICE—Vietnam era veterans
HC02_EST_VC06	ACS	Veterans; Estimate; PERIOD OF SERVICE—Korean War veterans
HC02_EST_VC07	ACS	Veterans; Estimate; PERIOD OF SERVICE—World War II veterans
HC01_EST_VC10	ACS	Estimated percentage of total that are veterans; SEX—Male
HC02_EST_VC10	ACS	Veterans; Estimate; SEX—Male
HC01_EST_VC11	ACS	Estimated percentage of total that are veterans; SEX—Female
HC02_EST_VC11	ACS	Veterans; Estimate; SEX—Female
HC01_EST_VC14	ACS	Estimated percentage of total that are veterans; AGE—18 to 34 years
HC02_EST_VC14	ACS	Veterans; Estimate; AGE—18 to 34 years
HC01_EST_VC15	ACS	Estimated percentage of total that are veterans; AGE—35 to 54 years
HC02_EST_VC15	ACS	Veterans; Estimate; AGE—35 to 54 years
HC01_EST_VC16	ACS	Estimated percentage of total that are veterans; AGE—55 to 64 years
HC02_EST_VC16	ACS	Veterans; Estimate; AGE—55 to 64 years
HC01_EST_VC17	ACS	Estimated percentage of total that are veterans; AGE—65 to 74 years
HC02_EST_VC17	ACS	Veterans; Estimate; AGE—65 to 74 years
HC01_EST_VC18	ACS	Estimated percentage of total that are veterans; AGE—75 years and over
HC02_EST_VC18	ACS	Veterans; Estimate; AGE—75 years and over
HC01_EST_VC21	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—One race
HC02_EST_VC21	ACS	Veterans; Estimate; RACE AND

Variable Name	Description	Type
		HISPANIC OR LATINO ORIGIN—One race
HC01_EST_VC22	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—One race—White
HC02_EST_VC22	ACS	Veterans; Estimate; RACE AND HISPANIC OR LATINO ORIGIN—One race—White
HC01_EST_VC23	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—One race—Black or African American
HC02_EST_VC23	ACS	Veterans; Estimate; RACE AND HISPANIC OR LATINO ORIGIN—One race—Black or African American
HC01_EST_VC24	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—One race—American Indian and Alaska Native
HC02_EST_VC24	ACS	Veterans; Estimate; RACE AND HISPANIC OR LATINO ORIGIN—One race—American Indian and Alaska Native
HC01_EST_VC25	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—One race—Asian
HC02_EST_VC25	ACS	Veterans; Estimate; RACE AND HISPANIC OR LATINO ORIGIN—One race—Asian
HC01_EST_VC26	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—One race—Native Hawaiian and Other Pacific Islander
HC02_EST_VC26	ACS	Veterans; Estimate; RACE AND HISPANIC OR LATINO ORIGIN—One race—Native Hawaiian and Other Pacific Islander
HC01_EST_VC27	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—One race—Some other race
HC02_EST_VC27	ACS	Veterans; Estimate; RACE AND HISPANIC OR LATINO ORIGIN—One race—Some other race

Variable Name	Description	Type
HC01_EST_VC28	ACS	Estimated percentage of total that are veterans; RACE AND HISPANIC OR LATINO ORIGIN—Two or more races
HC02_EST_VC28	ACS	Veterans; Estimate; RACE AND HISPANIC OR LATINO ORIGIN—Two or more races
HC01_EST_VC30	ACS	Estimated percentage of total that are veterans; Hispanic or Latino (of any race)
HC02_EST_VC30	ACS	Veterans; Estimate; Hispanic or Latino (of any race)
HC01_EST_VC31	ACS	Estimated percentage of total that are veterans; White alone
HC02_EST_VC31	ACS	Veterans; Estimate; White alone

APPENDIX B. DATASET TRANSFORMATIONS

We transform each publicly available dataset into a set of unique variables that allow us to determine which ZIP codes are similar to other ZIP codes. In some cases, we only change the datasets by removing ZIP codes that we do not include in our study. In other cases, we alter variables to better represent the characteristics of a ZIP code that relate to recruiting potential. In extreme cases, we create variables that we expect to represent a ZIP code's potential to produce recruits.

In these datasets we address several issues that occur frequently. There are ZIP codes that do not have data available, data that we do not have ZIP codes to assign it to, and there are variables that do not directly represent factors that indicate a ZIP code's potential to produce recruits. Wherever possible, we use similar methods to address issues uniformly.

A. COMMUNITY HEALTH STATUS INDICATORS

The Community Health Status Indicators dataset consists of 578 variables that the CDC collects at the county level. These variables describe health indicators that include demographic data, incidence of diseases, leading causes of death, and some user generated ratings of health (Centers for Disease Control and Prevention 2010). This dataset also contains health strata assignments determined by the CDC, confidence intervals, and comparisons of variables to the other counties assigned to the same strata.

Since we are using this data to cluster similar ZIP codes, we only use variables that represent ZIP codes themselves, not variables comparing ZIP codes to others. We remove variables that indicate relations to other ZIP codes, CDC assigned strata, and self-reported indicators of health. We also remove variables that either have no data or the same data for all ZIP codes and variables that indicate expected incidences of diseases. This leaves us with 154 variables.

We then remove OCONUS counties from this dataset by checking the state Federal Information Processing Standard code in the HUD county to ZIP code crosswalk from 2105 (Housing and Urban Development 2015). We find that the county containing

Batesland, South Dakota, did not exist in 2010, when the CDC created the CHSI dataset. We do not have a means to recreate this data so we remove the ZIP code corresponding to Batesland, South Dakota, from the list of ZIP codes that we study.

We also find that three counties that exist in the CHSI dataset no longer have ZIP codes with residential addresses. These are Kenedy and Loving counties in Texas, and Bedford City County in Virginia. Since we do not have a means to assign the CHSI data from these counties to current ZIP codes, we are not able to use this data to create clusters.

We now have data for 3,141 counties that correspond to 34,007 ZIP codes with residential addresses. We use the proportions of each county's residential addresses in ZIP codes to apportion the count data. The HUD (2015) county to ZIP code crosswalk contains these proportions. Count data includes reported cases of diseases, the total county populations, the numbers reporting access to care, and the amount of toxins released into the air each year.

Once we expand the county data into ZIP codes we have to combine the instances of ZIP codes that cross county lines. There are 9,422 ZIP codes that occur in 2 to 6 different counties. To combine this data, we total the numeric variables, and use weighted averages based on population for all rate variables. Rate variables include rates of death, and percentages of population of different races and genders. For binomial variables that indicate the presence of lead or other chemicals in the air, we use the maximum value. If a toxin occurs in the air in any county that contains a portion of a ZIP code, we treat the entire ZIP code as testing positive for that toxin.

After we combine the data for ZIP codes that occur in two or more counties, we transform some of the variables to make them more useful for our study. Since we are using total population of ZIP codes as one of our variables, we attempt to remove the effect of population from the other variables. We do this by changing other count data to rate data. For example, instead of using the number of people in a ZIP code who are unemployed, we use the percentage of people who are unemployed. We do this by dividing the ZIP code count data by the ZIP code populations. Of the 154 variables,

two—population and tons of toxic chemicals released annually—are count data. The four variables indicating presence of toxins in the air are binomial and the remaining 148 are rate data.

B. INDIVIDUAL INCOME TAX RETURNS

In the individual income tax dataset, we find that we have four CONUS ZIP codes of individual income tax returns that are not in the list of ZIP codes with residential addresses. Three of these ZIP codes—41713, 29905, and 79010—are ZIP codes that are only used for post office boxes. ZIP code 78843 is for Laughlin Air Force Base in Texas (United States ZIP Codes 2016). There are 6,434 ZIP codes that are in our study but are not in the individual income tax dataset. These ZIP codes likely do not have enough returns for the IRS to report their data (Internal Revenue Service 2016).

This dataset contains several variables that represent count data, including the number of returns for each household status and the number of returns of each adjusted gross income level in a ZIP code (Internal Revenue Service 2016). It also contains the number of returns that contain different types of information or forms. We transform these 64 count variables into percentages by dividing by the estimated ZIP code populations from the CHSI dataset. The only count data that we use in our study is the number of individual income tax returns filed in each ZIP code. We transform all variables that contain total dollar, dependent, and exemption amounts into averages for that ZIP code by dividing by the number of returns that contain that information. After these transformations we have 117 different individual income tax observations per ZIP code.

C. ECONOMIC CENSUS AND COUNTY BUSINESS PATTERNS

The Census Bureau collects the economic census every ten years. This dataset includes the number of commercial establishments in eight different business sectors such as health care, retail, and education (United States Census Bureau 2011). We create additional variables to indicate the percentages of total establishments of each sector in a ZIP code.

The Census Bureau collects the annual county business patterns survey that includes the number of commercial establishments in a ZIP code, the number of employees, and the total annual and quarterly payrolls (United States Census Bureau 2011). This dataset contains 33,485 of the ZIP codes we are working with and an additional 4,948 ZIP codes. The majority of these ZIP codes are the same as those in the economic census.

We map all 5,479 commercial ZIP codes to the nearest residential ZIP code using the great circle distance between the ZIP code geographic centers. We use the geographic centers from Breen's (2015) *zipcode* package for *R*. We find that two of the ZIP codes do not have latitudes and longitudes in Breen's dataset. For these—and for all ZIP codes without latitudes and longitudes—we use the information from United States Zipcodes (2016). This commercial dataset contains geographic, demographic and other ZIP code data (United States Zipcodes 2016). Once we determine the nearest residential ZIP code to each commercial ZIP code, we inspect the distances.

We see 98% of the ZIP codes are within 20 miles of the nearest residential ZIP code. We need to accurately represent the presence of economic opportunity but also want to maximize the amount of data that we include. We inspect the commercial ZIP codes that are more than 20 miles from the nearest residential ZIP codes and see that the majority are in Arizona (United States Zipcodes 2016) and are geographically large with fewer than 10 business establishments (United States Census Bureau 2007). The only commercial ZIP code that is more than 20 miles from a residential ZIP code with more than 10 businesses is Avalon, California on Catalina Island (United States Zipcodes 2016). We decide not to include the commercial ZIP codes that are more than 20 miles from the nearest residential ZIP code.

Once we determine which commercial ZIP codes to include, we combine the data with any existing data for the nearest residential ZIP code. We sum the count data, to include the number of establishments and payroll data then adjust the percentage data based on the new totals. When we combine the economic census and county business pattern datasets we have economic data for 34,006 of our 34,007 ZIP codes.

D. MILITARY BASES

The military base data comes from two different sources. The location data comes from the United States Census Bureau (2012) and the military population data comes from the Defense Manpower Data Center (2009). We use these datasets because these are the most recent years with data available. The location data contains the latitude and longitude of 782 different military locations. The population data contains the count of how many of over 2 million uniformed people live in 772 different bases or cities in the United States. The population data does not contain any information for the Coast Guard, while the location data includes 89 different Coast Guard stations (United States Census Bureau 2012).

The location dataset includes many unneeded entries, including recreation areas ranges, and radar sites. In most cases, these locations are duplicates of the actual bases that they support, which are located nearby (United States Census Bureau 2012). We manually remove from this list any location that is a duplicate of a main base location. We also remove any installations that are no longer open based on entries in the Department of Defense's (2016) list of military installations.

We manually combine the datasets using the base names where they are the same. In some instances, they are listed differently. For example, in the location dataset (United States Census Bureau 2012), the Naval Postgraduate School is listed as "Naval Postgraduate School (Monterey)," but in the personnel dataset (Defense Manpower Data Center 2009) it is listed as "Monterey."

In some cases, the population data is divided into two different locations. For example, there are Army populations in both "Fort Irwin," and "Barstow." We know that Barstow is the city directly outside Fort Irwin and that there are no other military bases nearby (Department of Defense 2016). In these cases, we sum the numbers of uniformed personnel. This results in 450 military installations, of which 191 include the data for military population that lives there.

To transform these datasets into variables that indicate military influence in ZIP codes, we first remove all duplicates and assign populations to military bases. We assign

each military location a factor indicating which military branch it belongs to. In addition to the four basic branches, we create categories for depots, Coast Guard stations, Army National Guard, and Air Force Reserve bases. We assign these categories based on the locations name or the largest population assigned to the location (United States Census Bureau 2012).

Once we calculate the distance from each ZIP code to each military base, we create variables for the type, distance to and military population of the nearest base. We also create variables for the types and populations of the largest base within 10 and 50 miles of each ZIP code. This results in seven variables for each of our 34,007 ZIP codes.

E. VETERAN POPULATIONS

We use veteran population data from the United States Census Bureau's (2014) American Community Survey. This dataset contains 246 variables for each ZIP code. Many of these variables indicate demographic information or margins of error of the estimates. We only use the variables that include counts of veterans of different genders and ages. The Census Bureau collected this data at the county level. We transform the data to ZIP code level using the same methods we use for the CHSI dataset, as we discuss in Section A of Appendix B.

We transform these data points from count data to percentages of populations using the total estimated ZIP code populations from this dataset. By using percent data, we better represent the military influence in a ZIP code. After mapping the data, selecting the variables, and transforming them, we have 41 variables for our 34,007 ZIP codes.

F. SECONDARY EDUCATION INSTITUTIONS

We use the secondary education institution data from the National Center for Education Statistics (2016). This dataset contains the locations, enrollment, sizes, school types, and other information for and sizes for 7,687 secondary education institutions in the United States. This dataset includes major universities and small trade schools. We sort the institutions by the five different size categories that indicate the enrolled

populations. The small category is under 1,000 students and the largest is 20,000 students and above (National Center for Education Statistics 2016).

To use this dataset, we transform the variables into count data indicating the number of each size of institution near each of our ZIP codes. We use two counts: institutions within 10 miles, and institutions within 50 miles. We calculate the distances between the ZIP codes and the education institutions using the same method we use for the commercial ZIP codes in the economic census. This shows that we have between zero and 306 of the smallest institutions within 50 miles of our ZIP codes. We have between zero and 22 of the largest institutions within 50 miles of each ZIP code. With the two distances and five sizes, we have ten variables for each of our 34,007 ZIP codes.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Breen J (2015) U.S. ZIP Code database for geocoding. R package version 1.0. Retrieved May 14, 2016, http://CRAN.R-project.org/package=zipcode.

Brown U, Rana D (2005) Generalized exchange and propensity for military service: The moderating effect of prior military exposure. *Journal of Applied Statistics.* 32(3): 259–270.

Buttrey S (2015) treeClust: Cluster distances through trees. R package version 1.1-1. Retrieved May 14, 2016, https://cran.r-project.org/web/packages/treeClust/index.html.

Buttrey S, Whitaker L (2015) treeClust: An R package for tree-based clustering dissimilarities. *The R Journal* 7/2 (December). Retrieved April 8, 2016, https://journal.r-project.org/archive/2015-2/buttrey-whitaker.pdf.

Centers for Disease Control and Prevention (2010) Community health status indicators to combat obesity, heart disease and cancer. Retrieved May 8, 2016, http://www.healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer.

Clingan M (2007) U.S. Army custom segmentation system. Presentation, 75th Military Operations Research Society Symposium, June 12–14, Annapolis, MD.

Defense Manpower Data Center (2009) Distribution of personnel by state and selected locations. Retrieved May 14, 2016, https://www.dmdc.osd.mil/appj/dwp/rest/download?fileName=M02.zip&groupName=pubSelectedLocations.

Department of Defense (2016) Military installations. Retrieved May 8, 2016, http://www.militaryinstallations.dod.mil/.

Dorminey D (2007) Are you a "mover & shaker"...or an American classic. *Recruiter Journal* (November). Retrieved April 4, 2016, http://www.usarec.army.mil/hq/apa/download/RJ/nov07.pdf.

Faraway J (2006) *Extending the Linear Model with R*. (Taylor and Francis Group, Boca Raton, FL).

Fleischmann M, Nelson M (2014) Recruiting mission allocation using a recruiting market index. Presentation, Army Operations Research Symposium, November 4, Aberdeen Proving Ground, Maryland.

Gibson J, Hermida R, Luchman J, Griepentrog B, Marsh S (2011) ZIP code valuation study technical report. Joint Advertising, Market Research & Studies (JAMRS), Defense Human Resources Activity, Arlington, Virginia.

Gower J (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Housing and Urban Development (2016) County—ZIP Crosswalk, 4th Quarter 2015. Retrieved April 8, 2016, https://www.huduser.gov/portal/datasets/usps_crosswalk.html

Hojnowski R (2005) Analyzing the assignment of enlisted recruiting goal shares via the Navy's enlisted goaling and forecasting model. Master's thesis, Naval Postgraduate School. Retrieved May 14, 2016, http://calhoun.nps.edu/bitstream/handle/10945/2284/05Mar_Hojnowski.pdf

Internal Revenue Service. (2016) Internal Revenue Service SOI. Retrieved April 20, 2016, https://www.irs.gov/uac/SOI-Tax-Stats-Individual-Income-Tax-Statistics-2013-ZIP-Code-Data-(SOI).

Intrater B (2015) Understanding the impact of socio-economic factors on Navy accessions. Master's thesis, Naval Postgraduate School. Retrieved May 14, 2016, http://calhoun.nps.edu/bitstream/handle/10945/47279/15Sep_Intrater_Bradley.pdf .

Joint Advertising, Market Research & Studies Group (2005) A national segmentation analysis of the joint services (FY00-FY05). Unpublished presentation.

Kahle D, Wickham H W (2016) Spatial visualization with ggplot2. R package version 2.6.1. Retrieved May 14, 2016, http://CRAN.R-project.org/package=ggmap.

Kaufman L, Rousseeuw P (1990) *Finding Groups in Data: An Introduction to Cluster Analysis.* (John Wiley and Sons, New York).

Krijthe J (2015) Package "Rtsne." Retrieved from https://cran.r-project.org/web/packages/Rtsne/Rtsne.pdf.

Lopez C (2014) Recruiting force remains unchanged, despite shrinking goals. *Army News Service* (January 16) Retrieved May 16, 2016, http://www.army.mil/article/118369/.

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (1: Statistics):281-297. Retrieved May 16, 2016, http://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992.

Market Analysis Division (2015) Segmentation analysis and market assessment. USAREC G2, Unpublished Presentation.

Marmion W (2015) Evaluating and improving the SAMA (segmentation analysis and market assessment) recruiting model. Unpublished master's thesis, Naval Postgraduate School. Retrieved May 14, 2016, http://calhoun.nps.edu/handle/10945/45894.

McDonald J (2016) Analysis and Modeling of U.S. Army Recruiting Markets. Master's thesis, Air Force Institute of Technology.

National Center for Education Statistics (2016) Integrated Postsecondary Education Data System. Retrieved April 10, 2016, http://nces.ed.gov/ipeds/.

Nielsen Company (2013) Nielsen Company's Potential Rating Index for ZIP Markets, New Evolution (PRIZM NE) Market Segmentation. http://www.srds.com/frontMatter/ips/lifestyle/reports/prizm.html.

Nielsen Company (2016) My best segments. Retrieved April 20, 2016, https://segmentationsolutions.nielsen.com/mybestsegments/

Oh, Y (1998) An analysis of factors that influence enlistment decisions in the U.S. Army. Master's thesis, Naval Postgraduate School. Retrieved May 14, 2016, http://calhoun.nps.edu/handle/10945/32726.

Parker N (2015) Improved Army Reserve unit stationing using market demographics. Master's thesis, Naval Postgraduate School. Retrieved May 14, 2016, http://calhoun.nps.edu/handle/10945/45921.

Pinelis Y, Schmitz E, Miller Z & Rebhan E (2011) An analysis of Navy recruiting goal allocation models. (Center for Naval Analysis, Arlington, VA)

United States Army Recruiting Command (n.d.). Brigades & Battalions Map, http://www.usarec.army.mil/bdemap.html.

United States Army Recruiting Command (2009) USAREC Manual 3–0: Recruiting Operations. (USAREC, Fort Knox, KY). Retrieved May 14, 2016 http://www.usarec.army.mil/im/formpub/rec_pubs/man3_0.pdf.

United States Army Recruiting Command (2013) *Recruiting Overview*. Retrieved April 8, 2016, http://www.usarec.army.mil/downloads/hq/Recruiting_Overview.ppt

United States Army Recruiting Command (2015a) National Leads Data. Unpublished dataset.

United States Army Recruiting Command (2015b) Recruiting station identification to ZIP code crosswalk. Unpublished dataset.

United States Army Recruiting Command (2015c) PIRZM NE market segments by ZIP code for 2011–2015. Unpublished dataset.

United States Army Recruiting Command (2016) Frequently asked questions about recruiting. Retrieved May 14, 2016, http://www.usarec.army.mil/support/faqs.htm.

United States Census Bureau (2007) United States Census Bureau economic census. Retrieved April 20, 2016, http://www.census.gov/econ/census/about/.

United States Census Bureau (2011) County business patterns. Retrieved May 8, 2016 http://www.census.gov/data/datasets/2011/econ/cbp/2011-cbp.html.

United States Census Bureau (2012) TIGR maps of U.S. military installations. Retrieved May 8, 2016, https://www2.census.gov/geo/tiger/TIGER2012/MIL/.

United States Census Bureau (2013a) Current lists of metropolitan and micropolitan statistical areas and delineations. Retrieved April 20, 2016, http://www.census.gov/population/metro/data/metrodef.html.

United States Census Bureau (2013b) Map of Metropolitan and Micropolitan Statistical Areas of the United States and Puerto. http://www.census.gov/population/metro/files/metro_micro_Feb2013.pdf .

United States Census Bureau. (2014) Veteran status 2010–2014 American community survey 5-year estimates. Retrieved April 20 2016, http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_S2101&prodType=table

United States Department of Defense (2015) DOD Announces Recruiting and Retention Numbers for Fiscal 2015, Through May 2015. *News Release*(NR-312-15, July 31) Retrieved May 7, 2016, http://www.defense.gov/News/News-Releases/News-Release-View/Article/612817.

United States Department of Defense (2016) Welcome to the joint advertising, market research & studies. Retrieved 8 May, 2016, http://jamrs.defense.gov/Home.aspx.

United States Office of Management and Budget (2015) Revised delineations of metropolitan statistical areas, micropolitan statistical areas, and combined statistical areas, and guidance on uses of the delineations of these areas. *OMB Bulletin*(15–01). Retrieved May 8, 2016, https://www.whitehouse.gov/sites/default/files/omb/bulletins/2015/15-01.pdf.

United States Postal Service (2016) History of the United States Postal Service. Retrieved April 20, 2016, https://about.usps.com/publications/pub100/pub100_001.htm

United States Postal Service Office of Inspector General (2013) The untold story of the ZIP code. Retrieved April 20, 2016, http://postalmuseum.si.edu/research/pdfs/ZIP_Code_rarc-wp-13-006.pdf.

United States Zip Codes (2016) United States ZIP Codes. Retrieved June 6, 2016, http://www.unitedstateszipcodes.org/.

Van der Maaten L., Hinton G (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research*, 1, 1–48.

Vergun D (2015) Army sees challenges ahead to recruiting future soldiers (August 21) Retrieved May 16, 2016, https://www.army.mil/article/154299/.

Wickham H, Chang W (2016) An implementation of the grammar of graphics. R package version 2.1.0. Retrieved May 14, 2016, http://CRAN.R-project.org/package=ggplot2.

Williams T (2014) Understanding factors influencing Navy recruiting production. Master's thesis, Naval Postgraduate School. Retrieved May 14, 2016, http://calhoun.nps.edu/bitstream/handle/10945/48125/14Dec_Williams_Taylor.pdf.

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