ANALYSIS AND FORECASTING OF THE 360TH AIR FORCE RECRUITING GROUP GOAL DISTRIBUTION

THESIS

Tyler Spangler, Captain, USAF

AFIT-ENS-MS-20-M-172

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-20-M-172

ANALYSIS AND FORECASTING OF THE 360TH AIR FORCE RECRUITING GROUP GOAL DISTRIBUTION

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Tyler Spangler, B.S.

Captain, USAF

26 March 2020

AFIT-ENS-MS-20-M-172

ANALYSIS AND FORECASTING OF THE 360TH AIR FORCE RECRUITING GROUP GOAL DISTRIBUTION

THESIS

Tyler Spangler, B.S.
Captain, USAF

Committee Membership:

Dr. Raymond R. Hill, Ph.D
Chair

Dr. Edward D. White, Ph.D
Member

AFIT-ENS-MS-20-M-172

Abstract

The 360th Recruiting Group (RCG) is comprised of 8 squadrons spread across 21 states employing 542 recruiters to access new enlistment contracts (NECs) across their zone. This research utilizes monthly open-source data from 2012-2017 to determine economic or demographic factors that significantly contribute to increased goaling and production potential in areas of the 360 RCG's zone. Using regression analysis, a model of recruiting goals and recruiting production is built to identify squadrons within the 360 RCG's zone that are capable of producing more or fewer recruits and the factors that contribute to this increased or decreased capability. This research identifies that a zone's high school graduation rate, the number of recruiters in a flight, and the number of JROTC detachments in a zone are positively correlated with recruiting goals and that a zone's obesity rate and voting participation rate are inversely related to recruit goaling. Additionally, this research found that the monthly number of recruits goaled and the number of JROTC detachments in a zone are positively correlated with recruit production and the unemployment rate is inversely related with recruit production. Using these two linear regression models, recruiting goal distribution and recruiting production are projected into 48 months of new data to identify higher goaling and production potential squadrons within the 360th Recruiting Group.

Acknowledgements

I would like to first offer my appreciation to my advisor, Dr. Hill. Thank you for your guidance and mentorship throughout this research process. I appreciate your feedback and direction while working through this research process and my time at AFIT.

I would also like to thank the AFIT faculty who provided valuable instruction throughout my time here. Thank you for introducing me to operations research and analysis and showing the impact that operations research can have in future assignments.

Tyler Spangler

Table of Contents

Page

Abstract . . . iv

Acknowledgements . . . v

List of Figures . . . viii

List of Tables . . . x

I. Introduction . . . 1

1.1 Air Force Recruiting Service Background . . . 1
1.2 Research Objectives . . . 5
1.3 Organization of Thesis . . . 6

II. Literature Review . . . 7

2.1 Introduction . . . 7
2.2 Recruitment Supply . . . 7
2.3 Recruitment Demand . . . 12
2.4 Enlistment Choices . . . 14
2.5 Conclusion . . . 17

III. Methodology . . . 19

3.1 Introduction . . . 19
3.2 Data Description . . . 19
Data Gathering . . . 19
Data Imputation . . . 24
Geographic Organization . . . 26
Final Data Set Structure . . . 31
3.3 Dimension Reduction . . . 32
3.4 Regression Analysis . . . 34
Multiple Linear Regression . . . 34
Hypothesis Testing . . . 37
Variable Selection . . . 39
Model Fit and Performance . . . 40
Model Adequacy . . . 42
Model Validation . . . 47
3.5 Conclusion . . . 48

Page

IV. Results and Analysis 49

4.1 Introduction 49
4.2 Recruitment Goaling Model 49
Principal Components Analysis 49
Model Formulation 51
Model Adequacy 53
Model Fit and Validation 56
Model Results 57
4.3 Recruitment Production Model 59
Principal Components Analysis 59
Production Model Formulation 60
Model Adequacy 61
Model Fit and Validation 64
Model Results 65
4.4 Recruit Goaling and Production Projections 67
Recruiting Goals Forecast 68
Recruiting Production Forecast 69

V. Conclusion 72

5.1 Future Research 74

Appendix A. Recruit Production and Goals for Flights 76

Appendix B. Independent Variables Forecast 80

Appendix C. Full Model Results 104

Bibliography 106

List of Figures

Figure Page

1 AFRS Geography [1] . . . 2

2 360th Recruiting Group Goal vs. Produced NECs . . . 5

3 ZCTA to County Relationship . . . 28

4 Example Residual Plots . . . 44

5 Goaled Model PCA Results . . . 50

6 Goal PCA Loading Matrix . . . 51

7 Goaled Model ($\sqrt[4]{y_1}$) Residual Analysis . . . 54

8 Durbin-Watson Test . . . 55

9 Actual vs. Predicted Goals for RY2012-2017 by Squadron . . . 58

10 Produced Model PCA Results . . . 59

11 Produced PCA Loading Matrix . . . 60

12 Produced Model Residual Plots . . . 62

13 Produced Durbin-Watson Test . . . 63

14 Predicted Recruit Production using Predicted Goals (RY2012-2017 . . . 66

15 Predicted Goals and Predicted Production (RY2012-2017) . . . 67

16 Forecasted Recruiting Goals . . . 69

17 Forecasted Recruiting Production . . . 70

18 Predicted Produced and Predicted Goals RY2012-2017 . . . 79

19 HS Graduation Rate Forecasts . . . 83

20 JROTC Detachment Forecasts . . . 87

21 Number of Recruiters Forecasts . . . 91

Figure Page

22 Obesity Rate Forecasts 95

23 Unemployment Rate Forecasts 99

24 Voting Participation Rate Forecasts 103

List of Tables

Table Page

1 Consolidated List of Research Variables . . . 23

2 Imputation Required for Each Research Variable . . . 26

3 Geographic Organization Required . . . 30

4 Consolidated Variables with Variable Notation . . . 31

5 Analysis of Variance Table . . . 38

6 Assumptions of Linear Regression . . . 42

7 Goal Model Parameters ($\sqrt[4]{y_1}$) . . . 52

8 Goal Model Summary Statistics . . . 56

9 Produced Model Parameters . . . 61

10 Produced Model Summary Statistics . . . 64

11 Goal Model Results . . . 104

12 Produced Model Results . . . 105

ANALYSIS AND FORECASTING OF THE 360TH AIR FORCE RECRUITING GROUP GOAL DISTRIBUTION

I. Introduction

1.1 Air Force Recruiting Service Background

The Selective Service Draft was eliminated in 1973 and the All-Volunteer Force was established [2]. With the establishment of the All-Volunteer Force, each service of the United States Military relied solely on volunteer enlistments and was responsible for recruiting volunteers into their service. The Air Force Recruiting Service (AFRS) was established in 1954 with the mission to "inspire, engage, and recruit the next generation of Airmen" [3]. The AFRS is headquartered at Joint-Base San Antonio, Texas and is comprised of approximately 2,862 members who are responsible for recruiting and accessing 100% of the enlisted force, 90% of the health professions officers, approximately 16% of the Line Officers, and 100% of the Air Force Chaplain Corps [3].

The AFRS is divided into three recruiting groups each responsible for a different region of the United States [3]. The three groups are the 360th Recruiting Group headquartered in Pennsylvania, the 369th Recruiting Group headquartered in Texas, and the 372nd Recruiting Group headquarted in Utah [3]. Additionally, there is also an AFRS presence overseas in England, Germany, Italy, Japan, Puerto Rico, and Guam [3]. Each group is divided into recruiting squadrons of which there are 28 total squadrons in the AFRS. Figure 1 shows the headquarters, geographic area of responsibility that each group is responsible for, and the locations of each of the 28

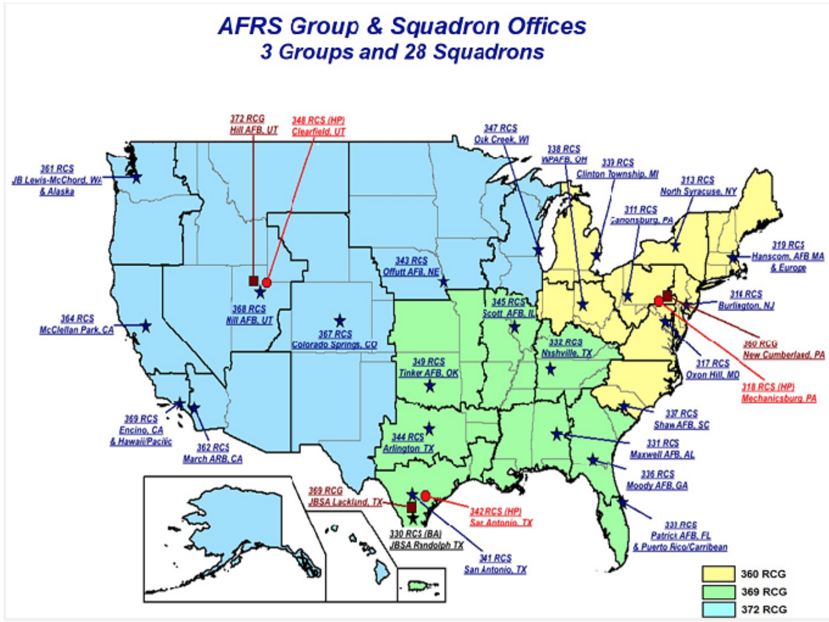


Figure 1. AFRS Geography [1]

Recruiting Squadrons. The 28 recruiting squadrons are further divided into flights that are comprised of recruiters. AFRS employs approximately 1,294 recruiters located in 1,040 recruiting offices throughout the United States and countries mentioned above. The 1,294 recruiters are responsible for recruiting and accessing new recruits to meet established recruiting goals.

Recruiters utilize a variety of methods to meet their monthly goals and to fill New Enlistment Contracts (NECs). The basic eligibility requirements to join the Air Force include being 17-39 years of age, having either a high-school diploma, a general education development (GED) with at least 15 college-credits, or a GED, and being a United States citizen or legal resident [4]. Recruiters aim to identify eligible and interest candidates that fulfill these criteria through a process called prospecting. Prospecting is accomplished through a combination of school visitation, participation in community events, or interested candidates contacting a recruiter. As part of the prospecting process, recruiters generate leads which are refined and prioritized into three categories. Priority 1 leads are individuals who have passed the high school

Armed Services Vocational Aptitude Battery (ASVAB) test and who are interested in joining the military [5]. Priority 2 leads are those that have passed the high-school ASVAB test, but are not interested in military service [5]. Priority 3 leads are all other leads outside of priority 1 and priority 2 leads [5].

When individuals decide to enlist, they must take the ASVAB test and their scores categorize them into 5 overall categores. The categories provide "an overall quality indicator" based on the percentile that the score falls in [5]. Category 1 is those who score within the 93-99 percentile, Category 2 is those individuals who score within 65-92 percentile, Category 3A is those who score in the 50-64 percentile, Category 3B is those who score within 31-49 percentile, Category 4A is individuals who score in the 21-30 percentile, Category 4B is percentiles 16-20, Category 4C is for percentiles between 10 and 15, and Catebory 5 is for those who score in percentiles 0-9 [5]. AFRS has a target that at least 70% of enlistees will be in category 3A or higher [6]. If an interested individual is a high school graduate and is in Category 4A or higher, or if they are a GED holder in Category 3A or higher, or if they are non-graduate or non-credential holder in Category 2 or higher, they are ineligible to enlist [5]. When enlisting potential recruits, it is the recruiters responsibility to ensure that recruits do not contain any disqualifying factors. The criteria that can render a recruit ineligible are wide ranging, but some examples include having been previously discharged from the United States Military with an other than honorable characterization, has a moral, drug, or dependency disqualification, or not meeting the minimum ASVAB score requirements [5].

The 360th Recruiting Group headquartered in Pennsylvania is responsible for and oversees the operations of eight enlisted recruiting squadrons in 21 states ranging from the Canadian border to South Carolina and westward to Michigan [7]. The 360th Recruiting Group also oversees recruiting efforts in Europe and the District of Columbia [7]. To recruit effectively in this vast area of responsibility, the 360th Recruiting Group utilizes 542 recruiters located in 60 recruiting flights throughout the area for which they are responsible. Each recruiter is responsible for accessing potential recruits in a group of zip codes. Each recruiter is assigned to a school or multiple schools and the zip codes that fall within that school's district is assigned to a recruiter.

Every fiscal year, the AFRS releases the Annual Market Mission Objective (AMMO) to each of the organizations under its command detailing the goal for each group and squadron for the fiscal year. The purpose of the AMMO is to meet the Air Force's accession requirements, provide indicators to allow organizations to identify and address production shortfalls, and provide a basis for rewarding organizations in meeting recruiting objectives [6]. The AMMO assigns each group and squadron a goal for each category of accession to meet the Air Force's accession requirement for the year. The different categories for enlisted accessions include Prior-Service (PS), Non-Prior Service (NPS), and Extended Active Duty (EAD). When recruiters enlist an individual this is counted a NEC and each recruiting organization is responsible to fill a certain amount of NECs each month.

The number of NECs each organization is responsible for is referred to as a goal. The goal for each group and squadron is provided at the beginning of each fiscal year and includes the goal for each month. The method of assigning the goal to each group and squadron for the New Enlistment Contract (NEC) category is using the non-prior service (NPS) production from fiscal years 2014-2018 with each year weighted equally [6]. After the AFRS releases the AMMO, each group distributes monthly goals to each of their squadrons. Prior to April 2019, the 360th Recruiting Group passed the AFRS goal directly to each of their organizations as their monthly goal. Beginning in May 2019, the 360th Recruiting Group developed a new goaling formula that consists of taking the squadron's capability multiplied by a manning factor. A squadron's capability is defined as the squadron's number of enlistments over the last five years compared to the group's total over the same period. The manning factor is the number of recruiters they have available during the goaled period.

Figure 2. 360th Recruiting Group Goal vs. Produced NECs

1.2 Research Objectives

Between the years 2011 through 2018 the 360th Recruiting Group was responsible for recruiting an average of 8,097 NECs annually. Figure 2 shows the number of NECs produced and the number of goals for recruiting years 2011 through 2018. The 360th Recruiting Group is interested in studying how to effectively goal their squadrons and recruiters. This research will focus on the goaling procedure for enlisted contracts within the United States. Although the 360th Recruiting Group is responsible for recruiting in Europe, this is outside the scope of this study. This research will utilize economic and demographic data to identify if there are any factors that contribute significantly to producing more or fewer recruits in a given area and how the 360th

Recruiting Group can effectively goal their squadrons and recruiters to meet the AFRS goal.

1.3 Organization of Thesis

This first chapter provides an overview of the AFRS, some recruiting practices, current goaling procedures, and the purpose of this research. The second chapter will review relevant studies that have been conducted to analyze military recruiting. The third chapter will discuss the methodology of this research including data collecting and preparation and model formulation. The fourth chapter will present the results and performance of the model. The final chapter summarizes the contributions of this research and provides areas for further research.

II. Literature Review

2.1 Introduction

Military recruiting has been researched for many years. The various research projects explore factors that contribute to the number of eligible recruits that could enter service, the factors that affect the recruiting effort and desire for recruiters to fill contract goals, and the conversion of possible recruits to enlistees by signing an enlistment contract. The first area that has been extensively studied is the area of recruiting supply which focuses on the econometric data that can help determine the number of eligible and interested recruits nationwide. The second area is that of the demand for recruits which focuses on "the behaviors of recruiters and job counselors, as well as military advertising" that influence enlistment decisions of potential recruits [8]. The final area of research pertains to the choice of eligible recruits to enlist in the military, which is also referred to as the conversion of recruits into enlistees [9]. These three areas of research interact to produce a projection of a market area's potential to produce an amount of enlistees in a given period. This literature review explores various techniques used to study these three areas and produce varying models which will aid in developing the methodology for this research project.

2.2 Recruitment Supply

A variety of techniques have been used to study enlistment supply. The use of econometric data has aided many researchers in developing models of enlistment supply.

The first widely used technique to model enlistment supply is various forms of regression models. Murray and McDonald [8] apply a linear regression model using the number of monthly contracts as the response variable to a time-series set of econometric data spanning fiscal year 1983-1993. The authors choose a linear model because for some months there were zero contracts signed making a logarithmic linear regression model impossible [8]. To compare the linear regression model to a logarithmic model where a month with zero contracts was eliminated, they used the model created for the Army; the Army had the largest number of observations [8]. The results of this comparison found that the responsiveness of the variables to changes was similar between the linear and logarithmic model with the deleted observations [8]. In [10] and [11], Dertouzos and then Simon and Warner, respectively, utilized a logarithmic linear regression model applied to a time-series dataset to model the recruiting supply. Dertouzos [10] utilized a dataset comprised of econometric data for 1980-1981 gathered from 33 Military Entrance Processing Stations. Simon and Warner [11] applied a log-linear regression model using the number of high-quality enlistments per youth in a state. Intrater et al. [12] utilized two regression techniques to model Naval enlistment supply at both the recruiting station level and at the zip code level for the east region, the west region, and the nation as a whole. To model enlistment supply at the station level, they utilized backward stepwise multiple linear regression using the number of annual Navy accessions as the response variable and 71 explanatory variables [12]. For the national and east region models, they used a square root transformation of the response and for the west region he utilized a cube root transformation of the response variable to meet the assumptions of a linear model [12]. Intrater et al. [12] used a zero-inflated negative binomial regression model to study enlistment supply at the zip code level which allowed them to first implement a logistic model to determine zip-codes that will not produce any accessions or structural zeros. They then utilize a negative binomial model to predict the non-structural zeros and the count data, or number of accessions in a zip code [12].

A second widely used array of techniques to study enlistment supply are multivariate statistical methods including principal components analysis, discriminant analysis, cluster analysis, and neural networks. McDonald et al. [13] utilized a multivariate approach using principal component analysis and mixed stepwise regression to formulate a model of enlistment supply in the Army. Using principal components analysis allowed them to "reduce model dimensionality and improve the level of statistical rigor" [13]. Williams [14] also used a multivariate approach in where he applied discriminant analysis to analyze econometric data to classify cities as either high-producing or low-producing cities. Using this technique, Williams [14] was able to identify econometric variables that were significant in determining whether a city was a high-producing or low producing city. Employing cluster analysis, Fulton [15] used 347 variables divided into five categories including economic, military, demographic, health, and education to cluster similar zip-codes. He utilized between two and eighteen clusters for each category of variable to group similar zip codes [15]. After clustering similar zip codes, he used a Poission regression model to predict the number of Army leads using the number of Army leads as the response variable and the cluster assignment as the predictor variables [15]. Monaghan [16] utilized a similar technique to Fulton [15], but she included estimates of the zip code's population as a predictor variable with the cluster assignments from [15] to predict the number of Navy leads produced at the zip code level. Marmion, in [17], built a linear regression model to study the potential of Army recruiting centers and compared the results of the linear regression to an artificial neural network (ANN). He created 15 different ANN models varying the number of hidden layers, hidden layer nodes, learning rates, and the number of boosts which corresponds to creating different models for the given inputs by varying the learning rate [17]. He found that the multiple linear regression models were better predictors than the ANN and that the ANN performed better on the training set of data, but worse than the regression models on the test dataset [17].

The above modeling techniques were applied to econometric data using variables to relate demographic and socioeconomic factors to military enlistment.

The various models discussed were applied to varying sets of econometric data to develop a model of recruitment supply. The authors of each study gathered data on variables that were believed to reflect factors that would contribute to and predict enlistment behavior. In [8], the authors utilized variables that represented civilian opportunities, military opportunities, recruiting effort, and the number of youths and contracts. To describe civilian opportunities, the authors used the unemployment and civilian earnings provided by the March Current Population Survey [8]. To measure military opportunities, the authors created a civilian to military pay ratio using the earnings described above and fringe benefits including enlistment bonuses and educational benefits [8]. The variables for recruiter effort are described below when detailing recruitment demand. Dertouzos [10] used similar variables to develop his supply model including the unemployment rate, wages for manufacturing production, and population of males ages 15-19. McDonald [18] used 12 variables to describe the level of enlistment supply. Some variables that McDonald used are similar to those in [8] and [10], but the variables that differ include voter participation rate, sponsor share or the number of Army service members in the area, the number of violent crimes and illicit drug use rate, adult obesity rates, and the population of youth ages 17-24 [18]. Williams [14] used data that summarized the composition of the city including median income, average income, population density and ages of the population, and the composition of the workforce broken out by different fields.

Intrater et al. [12] focused on using socioeconomic variables that included economic, educational, and veteran population variables. They used five economic variables including the zip code's unemployment rate, unemployment compensation, an index relating the number of retirees in a zip code not dependent on the job market, and the percentage of a zip code within each of the six categories of adjusted gross income [12]. To describe educational opportunities, they calculated the distance to the nearest divison I university and counted the number of division I universities within 50 miles of each zip code centroid, the total number of universities in each zip code, and used the maximum enrollment data for the largest university in the zip code to divide schools into five population categories [12]. To measure veteran population they used veteran status divided into five categories based on age and gender gathered from the United States Census Bureau [12]. To cluster zip codes, Fulton [15] and Monaghan [16] utilized 347 variables divided into five categories including economic, demographic, health, military, and education. The five categories of variables include variables used in studies previously mentioned including population, age, race, income, employment statistics, unemployment, high-school graduation rates, school sizes, obesity rates, and veteran populations divided by gender and age [15]. Marmion [17] used the four-year weighted average number of enlistments to measure previous performance, unemployment rate, size of recruiting boundary, and distance to recruiting center. He also includes index scores that measure the representation of high performing segments, high performing social groups, and high performing lifestyle groups within a center developed by the Army Potential Rating Index for Zip Markets, New Evolution (PRIZM NE) [17]. The common variables used throughout the research pertain to economic and demographic factors and the results of the models show how they help to relate these factors to enlistment supply.

The results of the models described above can help determine which of the included econometric or demographic variables are useful in describing enlistment supply. Murray and McDonald [8] applied their enlistment supply model to each of the four services and the results for the Air Force model show that the variables representing the population of youth and the civilian to military pay ratio were significant in determining the supply of Air Force recruits. Similarly, Dertouzos [10] found that the unemployment rate and wages were significant in describing enlistment supply. Using principal components analysis, McDonald [18] redefined variables that were loaded on the same principal component. The supply variables that he redefined were the propensity, obesity rate, and high-school graduation rate and the unemployment rate was included in the model as its own independent variable [18]. Using discriminant analysis, Williams [14] determined that variables pertaining to income levels, education levels, and the composition of the cities workforce were significant in classifying areas as high or low producing cities for recruiting. Intrater et al. [12] found that the variables significant in predicting supply at the national level include income, the unemployment rate, and unemployment compensation. Using the logistic model to predict structural zeros, they found that the supply variables of violent crime reports, unemployment compensation, and income were significant in predicting zip codes that would produce zero recruits [12]. Fulton [15] found that the utilizing the economic data to determine clusters provided more predictive power than using the other categories or a combination of all the categories. Marmion [17] found that the significant factors that contributed to enlistment supply using the linear regression model are the four-year weighted average of contract performance, the index score of high-performing segments, and the index score of high-performing social groups. The models and econometric models described above can used with similar models of recruiting demand to develop a comprehensive econometric model of recruitment behavior.

2.3 Recruitment Demand

Dertouzos [10] notes that it is important to not only consider enlistment supply because “recruiters do not passively process enlistments.” Recruiting demand is defined to include the factors that contribute to recruiters using resources and effort to fulfill contracts. While econometric data have been used to study enlistment supply, many factors of enlistment demand are specific to and controlled by the military. Many of the studies previously discussed are used in this section to discuss enlistment demand. The methods employed by the studies are the same methods, but the variables used to model the enlistment demand are detailed here.

Similar to enlistment supply, there are many variables that that appear to have an effect on enlistment demand. These measures indicate a recruiter's level of effort or use of resources to fill enlistment contracts. Murray and McDonald [8] captured recruiter effort through the use of recruiter goals and the number of recruiters in an area. Recruiter goals are the number of contracts that recruiters are assigned to fill during a given period (usually one month). The research in [8], [10], and [19] focus on Army recruiting and therefore differentiate between two categories of recruits: high-quality recruits and low-quality recruits. Dertouzos [10] found that there is a trade-off between filling contracts for these two categories. The effort of recruiters to fill these two categories captures enlistment demand and can be measured through the use of the goals or quotas. Both [8] and [10] found that the number of recruiters and the effort of recruiters was significant in determining the enlistment demand. Effectively setting the quotas or goals for recruiters is a significant indicator of future performance; if recruiters perceive that the goal is too easy they will under-produce and if the goal is perceived to be too difficult they will become overwhelmed and not produce enough enlistments [19]. Dertouzos [19] also found that "stations with more than one regular Army recruiter tend to be less productive than those in one-recruiter stations." Similarly, when studying Navy enlistment, Intrater et al. [12] determined that the average number of recruiters per year was negatively related to the number of accessions each year, which indicates that as the number of recruiters increase, fewer recruits are accessed. Incorporating enlistment demand into his research, McDonald [18] measured enlistment demand through the use of the number of goals, the recruiter share, the number of appointments made, the number of appointments conducted, and the number of days to process a recruit. The effective use of goals or quotas is significant in describing enlistment demand. The models incorporate both enlistment supply and enlistment demand to fully capture the number of potential recruits in a market area. The third area of previous research explored is the area of differing choices for potential recruits.

2.4 Enlistment Choices

In many studies of the enlistment supply, the population of individuals between ages 15-24 were used as the target demographic for recruiting efforts [8] [10] [18]. This age range, high-school youth and after graduating from high-school, typically face the choice of joining the military, pursuing further education, or joining the workforce. The choice faced by these potential recruits is an important consideration in determining the overall enlistment potential of areas.

There has been considerable research on what factors drive individuals to enlist in the armed forces. A study conducted in 1976 analyzed the survey responses of individuals who were currently in Basic Military Training at Lackland AFB, Texas in 1970-1971 [20]. This research was conducted prior to the end of the draft and was administered to volunteers to see what incentivizes volunteering. The survey consisted of questions about where the volunteer was from and what aspects of the military were appealing to their volunteering [20]. The survey results found that volunteers were more likely to join the military if they were from the Far West and Great Lakes regions [20]. The survey also found that volunteers were likely to enlist because of job related factors of the military including technical training, competent supervision, security, interesting and challenging work, and equitable pay [20].

Another factor that affects youths' choice to enlist is the role of influencer attitudes and recruiters' access to potential recruits [9]. The research in [9] explores the reasons behind the reported increase in difficulty in filling enlistment contracts. Using results from the Youth Attitude Tracking Study (YATS), the researchers explored propensity of high-school students to enlist and influences that would entice them to enlist [9]. This study analyzed the YATS for trends in the "proportion of mothers, fathers, or friends advising" potential recruits against joining and did not find a significant change in the perception of joining the military [9].

Previous studies mentioned in this literature review have included variables that relate the number of potential influencers in a geographic area including the sponsor share in [18] and the veteran population in different age ranges in [19]. Additionally, Intrater et al.'s [12] national level regression model found that the number of male veterans aged 35-54 were significant and had a positive effect on the number of recruits access. Their logistic regression model to determine the probability of a zip-code producing zero recruits found that the population of veterans in an area is significant in reducing the probability of a zip-code being a structural zero [12]. These factors are important in developing youth perception of the military and can affect the choice of potential recruits. A second reason that Orvis and Asch [9] explored for the difficulty in filling enlistment contracts is the access of recruiters to potential recruits. The research used Recruiter Surveys from 1991-1996 and ASVAB testing rates. The research in [9] shows that there was a decrease in recruiter contact with high-school students which decreases the ASVAB testing rate. The reduced contact with high school students relates to increased difficulty for enlisting youth into the Delayed Entry Program [9].

Potential recruits also face the choice of pursuing further education through either two or four-year colleges or technical schools. Including educational variables into their study, Intrater et al. [12] looked to measure the effect of further education on Navy enlistments. In the logistic model used to measure the probability of producing a structural zero, they found that the number of division I universities within 50 miles of the zip code and the total number of universities in the zip code were significant and increased the probability that a zip code would produce zero recruits [12]. Williams [14] found that the percentage of college graduates decreases the recruiting potential in an area. He found that as the number of college graduates increases in an area, the city is more likely to be a low-producing area for recruiting [14].

To further analyze the trends of youth who plan to pursue further education, Kilburn and Asch [21] utilized the data from the Monitoring the Future (MtF) survey and analyzed different factors that affected the decisions to pursue further education. The MtF survey consists of questions asking respondents about their intentions after high school to include attending a technical school, serve in the armed forces, graduate from a two-year or four-year college, or attend graduate or professional school after college [21]. The possible responses for each of these questions include 'Definitely Won't', 'Probably Won't', 'Probably Will', or 'Definitely Will' [21]. Kilburn and Asch [21] begin by defining the term youth in the college market as "high school youth who plan to go to college soon after completing high school, youth who are already in college, or youth who might have recently left college". They define this term because youth in the college market are a primary source of potential recruits and understanding the trends in the decisions to pursue military enlistment versus pursuing further education is important in recruiting these potential recruits.

To analyze the MtF data, the Kilburn and Asch [21] utilize multinomial logistic regression because it allows the researcher to retain all the information of the survey instead of reclassifying the responses into either binary variables for logistic regression or continuous variables. The authors developed three separate models for the responses of the MtF; a model of the predictors for two-year college intentions, a model of the predictors for four-year college intentions, and a model of the predictors for military service [21]. Each model contains a set of independent variables expected to influence intentions of each model and the results of the multinomial logistic regressions indicate estiamtes of the "amount by which the predicted odds of a given outcome are multiplied for each one unit change in the independent variable" [21]. The independent variables used in the models include information on the respondent's family information, geographic information, and high-school educational performance. The results of the model in [21] model focus on how intentions of pursuing two-year or four-year college are affected by intentions of joining the military. The results of the model for two-year college intentions compared to intentions of joining the military show that for respondents with positive military intentions, pursuing two-year college more incompatible with military service indicating that competition between two-year college and military service is more pronounced than the competition between four-year college and military service [21]. The model studying four-year college intentions show that students who indicate that they 'Probably Won't' serve in the military have positive inclination to pursue further education in a four-year college and students who reflect positive intention of serving in the military are less likely to have positive intention of pursuing a four-year college [21]. The results of this multinomial logistic regression highlight the complex relationship between the choices of high-school youth after graduating.

2.5 Conclusion

This literature review summarized previous research in regards to military recruiting in the three general areas of enlistment supply, enlistment demand, and factors that contribute to the choices of potential recruits to choose military service over pursuing further education or entering the civilian workforce. Much of the previous research summarized here was research applied to Army and Navy recruiting. The remainder of this thesis will build on previous research to develop a model of Air Force recruiting to aid the 360th Recruiting Group to effectively establish monthly goals for their squadrons while considering the effect of enlistment supply, enlistment demand, and enlistment choices in their geographic area.

III. Methodology

3.1 Introduction

To address the research objectives of determining whether there are economic or demographic factors that would improve the recruitment goaling process for 360th Recruiting Group's area of responsibility, a variety of methods were used to gather, clean, and organize open source data, to build a statistical model of recruit goaling and production , and to finally assess the validity of the model and its ability to predict future recruit production. This research utilized open source data to characterize a recruiting flight's area of responsibility or zone. Although goals are distributed to individual recruiters, data relating individual recruiters to specific zones was unavailable. For this reason, this research focuses on the production and goaling of the flights within the 360th Recruiting Group. The first section of this chapter details the process of identifying relevant research variables, gathering data to represent these variables, and the data cleaning and organization to form a usable data set. The remaining sections detail a variety of statistical techniques used to build a model of recruitment goaling and assess the validity and performance of the model in predicting recruiting goals.

3.2 Data Description

Data Gathering.

The data used to build a model of recruitment goaling procedures was gathered using open source data. For the purpose of this research, open source data is defined as any source of data that is publicly available through the internet or other means that can be accessed by government or non-government organizations. The first step was to identify which variables could be used to characterize a recruiting zone and build a model of recruitment goaling. Utilizing the framework outlined in the literature review and variables used in previous studies, a total of 20 variables were identified to characterize recruiting supply, recruiting demand, and the choice of potential recruits to enlist in the USAF.

Recruiting supply is the economic or demographic characteristics of an area that can help determine the number of eligible and interested recruits in an area. This research identified 11 variables from various data sources to describe recruiting supply throughout the 360th Recruiting Group's zone. The first data source used was the United States Census Bureau American Community Survey (ACS) 5-year estimates [22]. This data provided the total population, the population of 15-19 year olds, and the population of 20-24 year olds at the county level for years 2011-2017. From these variables, an additional variable was created to represent the proportion of the population of 15-24 year olds by adding the two gathered age variables together. The population of 15-24 year olds was divided by the total population to find the proportion of the population in this age range.

The second data source used to gather variables relating to recruiting supply was the Bureau of Labor Statistics (BLS) Local Area Unemployment Statistics (LAUS) [23]. The BLS LAUS data provided the number of unemployed and the size of the labor force at the county level at monthly intervals for the years 2011-2017. From this data, the unemployment rate was calculated by dividing the number of people unemployed by the size of the labor force and the labor force participation rate was calculated by dividing the size of the labor force by the total population from the ACS population data.

The third data source was the County Health Rankings and Roadmap data [24]. This data source provided multiple variables including the high school graduation rate, the obesity rate, and the number of violent crimes. The high school graduation rate measures the percentage of the ninth grade cohort that graduates within four years. The obesity rate measures the percentage of the adult population (aged 20 and older) that is considered obese. The violent crime variable measures the number of reported violent crimes per 100,000 people. This data was organized at the county level and was available for 2011-2017.

The next variable used to characterize recruit supply was the voter participation rate. The number of total votes cast for the 2012 and 2016 U.S. Presidential Election in each county of the U.S. was found on [25]. Using the total number of votes cast, the voter participation rate was found by dividing the total number of votes by the population of individuals over the age of 18 [22]. The voting participation rate for the 2012 presidential election was used for years 2012-2015 and the 2016 presidential voting rate was used for years 2016-2017.

The final variable describing recruiting supply was Qualified Military Available (QMA) provided by the AFRS. The QMA is a Department of Defense estimate of the number of youth who are eligible and available for military service without a waiver [26]. QMA is measured as the population aged 17-24 who meet eligibility requirements included in the following categories: medical/physical, overweight, mental health, drugs, conduct, dependents, and aptitude [26]. This data was provided for each zip code in AFRS and was available for the years 2011-2017.

The second set of variables used to develop a model of recruit production was that of recruiting demand. Recruiting demand relates to recruiting specific behaviors that influence enlistment decisions of potential recruits. Recruiting demand variables were gathered from the AFRS and the Air Force Recruiting Information Support System - Total Force (AFRISS-TF) which is the information system that recruiters use to input their goals, production, and marketing efforts [27]. The first variable provided by the AFRS is the zone area in miles which measures the total area for which a recruiter is responsible. This was provided for each zip code in the 360th Recruiting Group and the flight's total zone area is the aggregation of each recruiting office's zone. This data was provided for only the current zone delineations, but was applied to the years 2011-2017 under the assumption that zone areas for recruiting flights do not change frequently. Because data was only available for the current recruiting year, to capture the impact of the zone area over time, the total zone for a flight was divided by the number of recruiters in the flight each month. This produced a variable to represent the zone per recruiter in a flight. Both variables were included in the study.

The data gathered from AFRISS-TF include the number of recruits produced, the number of recruits goaled, the previous number of recruits produced, and the number of recruiters in each flight. The number of recruits produced is the number of NECs that a flight accesses during each month and was available monthly for the years 2012-2017. The number of recruits goaled is the number of recruits that each flight is assigned to access during each month and monthly data for years 2012-2017 was available. The previous number of recruits produced is the number of recruits accessed in the previous month for each flight.

The final area of recruiting goaling identified was enlistment choice which identifies factors that influence potential recruits to enlist in the USAF. Four variables were identified to represent the choice of recruits to enlist. The first variable used to characterize this choice is the rate of the adult population (aged 25-44) who possess some post-secondary education from the County Health Rankings [24]. This data was organized at the county level and available annually for years 2011-2017.

The second variable identified was the proportion of veterans in each county. This data was from the U.S. Census Bureau ACS five-year estimates [22] and was organized at the county level and annual data was available for years 2011-2017. The proportion of veterans in each county was calculated as the number of veterans in a county divided by the total population of the county.

The third variable to characterize enlistment choice is the number of active duty members in an area. This data was provided by the Annual Demographic Profiles provided by Military One Source [28]. This data was organized by zip codes and available for years 2011-2017.

Table 1. Consolidated List of Research Variables

Variable Type	Variable Name	Time Unit/Geographic Level
SupplyVariables	Total Population	Annual/County
	15-19 Population	Annual/County
	20-24 Population	Annual/County
	15-24 Population	Annual/Count
	Unemployment Rate	Monthly/County
	Labor Force Participation Rate	Monthly/County
	Obesity Rate	Annual/County
	HS Graduation Rate	Annual/County
	Violent Crime Rate	Annual/County
	QMA	Annual/Zip Code
	Voter Participation Rate	Annual/County
DemandVariables	Goal	Monthly/Recruiting Flight
	Recruit Production	Monthly/Recruiting Flight
	Previous Recruit Production	Monthly/Recruiting Flight
	Zone Area	2019 Data/Recruiting Flight
	Number of Recruiters	Monthly/Recruiting Flight
	Zone per Recruiter	Monthly/Recruiting Flight
Enlistment ChoiceVariables	Some College Rates	Annual/County
	Veteran Population	Annual/County
	Number of Active Duty Personnel	Annual/Zip Code
	Number of JROTC Units	Annual/Zip Code

The final variable used to model enlistment choice is the number of Junior Reserve Officer Training Corps (JROTC) detachments in a recruiting flight's zone. The number of JROTC detachments is publicly available on the Holm Center website unit locator [29]. This unit locator shows all the active JROTC detachments around the world and the zip codes associated with the detachment. A listing was provided by the Holm Center that showed the deactivated units and the date they were deactivated. From the unit locator website and this listing, a complete list of all detachments for years 2012-2017 was generated. Table 1 shows the final list of research variables included in this study and the time unit and geographic organization for which the data was available. After gathering all the relevant research variables, this research used multiple different imputation techniques to both fill missing values and to interpolate monthly values from the annual data sets.

Data Imputation.

Data imputation is the process of identifying and estimating missing values within the data [30]. This process involves identifying missing values within the data and estimating the value with a "likely" value based on information contained within the data set or outside information to improve the most likely value [30]. Data imputation was first used in this research to correct missing values in the annual data. To correct the missing data in the annual variables, a form of hot deck imputation and mean value imputation was used. Hot deck imputation is the process of replacing missing values in the data with values that already occur in the data set [30]. When using hot deck imputation, the missing value is estimated by using values from similar observations [30]. These values are not randomly chosen to estimate the missing value, but are chosen based on the complete observation's similarity to the observation with a missing value. Mean value imputation replaces the missing value with the sample average of the non-missing observations [30].

Much of the data collected for this research was organized at the county level and a method of identifying similar observations was needed as there were not enough complete observations to use to compare them to each other. Using the counties as a measure of similarity between observations under the assumption that there will be homogeneity among neighboring counties, the closest county with a non-missing value was used as the value to replace the missing value. The National Bureau of Economic

Research publishes a County Distance Database that contains the distances between counties collected from the 2010 Census [31]. This data calculates the distance to all neighboring counties within 25 miles for each county in the U.S. If the value for the nearest county was also missing, the value for the second closest county was used to replace the missing value. If there were no counties within 25 miles containing data, mean value imputation was used calculating the mean of the county for the years that data was available. Many of the variables gathered were rates, with the numerator and denominator for the rates. In most cases, the data for the rates was more complete than the numerator or denominator, so the rate was first imputed followed by the numerator. The denominator was then derived using the rate and the numerator.

The second source of imputation required for this research was to interpolate monthly values of the annual data. Because recruiting production and goals are tracked monthly, imputation of monthly values for each of the variables was necessary. To interpolate the monthly values for each annual variable, stochastic mean value imputation was performed. This method of imputation is a variation of the mean value imputation method that adds in a random component to the mean value to capture variability in the data [30]. The stochastic mean value imputation used in McDonald [18] was used for this research. In this method, McDonald [18] derived the means for the monthly value by first finding the gradient, δ, between the two annual values. For example, to impute the monthly values for a variable x, the first annual value is denoted as x_t and the annual variable for the subsequent year is denoted as x_{t+11}. The gradient is then the difference between these two values divided by 12. The mean for the monthly values is then denoted as $\hat{\mu_t} = x_t + \delta t$ for $t \in \{1, 2, ..., 11\}$. He then found the standard deviation for the monthly values as $\sigma = \frac{12\delta}{4} = 3\delta$. Using the mean and standard deviation to characterize a normal distribution, he applied the inverse-transform technique to generate a random standard error, ϵ, value for each monthly value using the norm.s.inv function in Excel. The standard error

Table 2. Imputation Required for Each Research Variable

Variable Type	Variable Name	Time Unit/Geographic Level	Imputation Required
SupplyVariables	Total Population	Annual/County	Stochastic Mean Value Imputation
	15-19 Population	Annual/County	Stochastic Mean Value Imputation
	20-24 Population	Annual/County	Stochastic Mean Value Imputation
	15-24 Population	Annual/County	Stochastic Mean Value Imputation
	Unemployment Rate	Monthly/County	No Imputation Required
	Labor Force Participation Rate	Monthly/County	No Imputation Required
	Obesity Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation
	HS Graduation Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation
	Violent Crime Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation
	QMA	Annual/Zip Code	Stocastic Mean Value Imputation
	Voter Participation Rate	Annual/County	No Imputation Required
DemandVariables	Goal	Monthly/Recruiting Flight	No Imputation Required
	Recruit Production	Monthly/Recruiting Flight	No Imputation Required
	Previous Recruit Production	Monthly/Recruiting Flight	No Imputation Required
	Zone Area	2019 Data/Recruiting Flight	No Imputation Required
	Number of Recruiters	Monthly/Recruiting Flight	No Imputation Required
	Zone per Recruiter	Monthly/Recruiting Flight	No Imputation Required
Enlistment ChoiceVariables	Some College Rates	Annual/County	Stochastic Mean Value Imputation
	Veteran Population	Annual/County	Stochastic Mean Value Imputation
	Number of Active Duty Personnel	Annual/Zip Code	Stochastic Mean Value Imputation
	Number of JROTC Units	Annual/Zip Code	No Imputation Required

calculation is $\epsilon = norm.s.inv(Rnd()) * \frac{\sigma}{\sqrt{12}}$. The final formula to impute the monthly values is $x_t = x_t + t\delta + \epsilon$ for $t \in \{1, 2, ..., 11\}$. Utilizing this technique for each annual variable in the data set generated monthly values for each year of data gathered. The final result of this imputation yielded 72 months of data across all variables. Table 2 shows each variable and the imputation technique required.

Geographic Organization.

The 360th Recruiting Group is organized into four different levels of recruiting organizations. The lowest level of the recruiting organization is the individual recruiter. Each recruiter in the 360th Recruiting Group is assigned a set of zip codes and they are responsible for accessing recruits within this set of zip codes. This set of zip codes is called their zone. The next level of the recruiting organization is the recruiting office. A recruiting office is an aggregation of individual recruiters. The zone for a recruiting office is the set of zip codes of all the recruiters within the recruiting office. The recruiting flight is the next level of organization and is an aggregation of multiple recruiting offices and the recruiting zone for each flight is the zone of each of the offices within the flight. This research focuses on the flight level of the recruiting organization. The final level of recruiting organization is the squadron. The squadron's zone is the largest and is the aggregation of the zip codes within each flight in the squadron.

The zone for the entire 360th Recruiting Group contains 15,023 zip codes within 21 states. These 15,023 zip codes are dispersed throughout its flights. As discussed previously, much of the data gathered to represent each recruiting zone was available at either the county level or zip code level. The county level and zip code level data was pulled for each county or zip code within all the states for which the 360th Recruiting Group is responsible. With data organized at two different geographic levels, a method to map county data to zip codes and organize the data for each flight was necessary.

A Zip Code Tabulation Area (ZCTA) is a generalized areal representation of zip codes and is a trademark of the U.S. Census Bureau [32]. The U.S. Census Bureau creates ZCTAs by first examining addresses within each census block and assigns a ZCTA as the most frequently occurring zip code in that block. If the block did not have a single most occurring zip code, the census block was assigned the ZCTA of the longest boundary of that area. In many cases, the ZCTA is the same as the zip code for the area. A ZCTA can span multiple counties or zip codes. The U.S. Census Bureau publishes a relationship file that shows the relationship between a ZCTA and county. Figure 3 shows an example of this relationship [33]. This figure shows the example ZCTA, 85602, highlighted in yellow, and encompasses parts of county 04003 and county 04019. The U.S. Census Bureau publishes a relationship file that shows the proportion of the population that reside in the counties encompassed by each ZCTA [34]. This file can be used to map each county to a ZCTA.

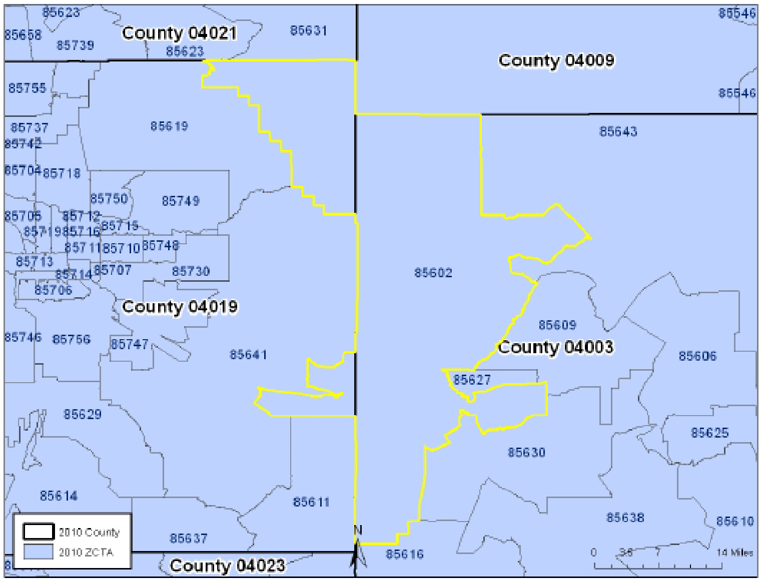


Figure 3. ZCTA to County Relationship

In many cases the ZCTA will be the same as the zip code, but this is not always the case. To align all the gathered data with each flight, all zip codes within the 360th Recruiting Group's zone were mapped to a ZCTA. The Uniform Data System (UDS) Mapper publishes a file that shows the 2018 zip code to ZCTA crosswalk [35]. The recruiting zone zip codes are assigned for 2019, but the most recent file published by the UDS Mapper is 2018, so this crosswalk was used to map the recruiting zip codes to ZCTAs. This mapping process resulted in 32 zip codes within the 360th Recruiting Group's zone not being mapped to ZCTAs. The U.S. Census Bureau acknowledges that some zip codes will contain very few addresses and will not be assigned ZCTAs [32]. After the zip codes in the 360th Recruiting Group zone was assigned ZCTAs the county data had to be mapped to ZCTAs.

There are 18,884 ZCTAs contained within the 360th Recruiting Group's 21 state zone. To map the county to a ZCTA, a weighting procedure, adapted from [18], was used where the value of a variable was multiplied by the percentage of the county population that resides in the ZCTA. Using the ZCTAs and weighted values, each weighted value of the variable was summed over each ZCTA in each recruiting flight. Showing the mathematical interpretation of this, let Z be the set of (m= 1,2,3...,18,884) ZCTAs within the 360th Recruiting Group's Zone and the set C be the set of (n = 1,2,3...,1084) counties in the 21 states within the 360th Recruiting Group's Zone. The weighted values of each variable x_i was found using equation 1.

$$x_i' = \upsilon_{m(n)} x_n \tag{1}$$

where

$x_i' \equiv$ the ZCTA weighted value of variable x_i

$\upsilon_{m(n)} \equiv$ the proportion of county n contained within ZCTA m

$x_n \equiv$ the unweighted value of variable x for county n.

After the ZCTA weighted values were found, the ZCTAs were assigned to a recruiting flight using the previously described zip code to ZCTA crosswalk and the values of all the ZCTAs within the flight was summed to produced a flight level value for each variable. This final flight level value is shown in equation 2 .

$$x_{i,j} = \sum_{Z_j} x_i' \tag{2}$$

where

$x_{i,j} \equiv$ the value of variable i for flight j, where j is the set of all recruiting flights

$x'_i \equiv$ the ZCTA weighted value of variable i

$Z_j \equiv$ All ZCTAs within recruiting flight j.

Table 3 shows the geographic organization technique required for each research variable. With the rate variables gathered, this mapping procedure was used on the

Table 3. Geographic Organization Required

Variable Type	Variable Name	Time Unit/Geographic Level	Imputation Required	Geographic Organization Required
SupplyVariables	Total Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	15-19 Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	20-24 Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	15-24 Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	Unemployment Rate	Monthly/County	No Imputation Required	ZCTA Weighting
	Labor Force Participation Rate	Monthly/County	No Imputation Required	ZCTA Weighting
	Obesity Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation	ZCTA Weighting
	HS Graduation Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation	ZCTA Weighting
	Violent Crime Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation	ZCTA Weighting
	QMA	Annual/Zip Code	Stochastic Mean Value Imputation	Zip-to-ZCTA Crosswalk
	Voter Participation Rate	Annual/County	No Imputation Required	ZCTA Weighting
DemandVariables	Goal	Monthly/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	Recruit Production	Monthly/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	Previous Recruit Production	Monthly/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	Zone Area	2019 Data/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	Number of Recruiters	Monthly/Recruiting Flight	No Imputation Required	N/A
	Zone per Recruiter	Monthly/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
Enlistment ChoiceVariables	Some College Rates	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	Veteran Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	Number of Active Duty Personnel	Annual/Zip Code	Stochastic Mean Value Imputation	Zip-to-ZCTA Crosswalk
	Number of JROTC Units	Annual/Zip Code	No Imputation Required	Zip-to-ZCTA Crosswalk

numerator and denominators of each rate variable. The rate was then found using the weighted values for the numerator and denominators. In very few instances, about 0.7% of observations, the rate was greater than 1. These values were truncated to one as it was determined to be the result of the ZCTA weighting procedure. This mapping procedure produced flight level ZCTA weighted values for each of the 20 variables used in the study.

Final Data Set Structure.

After imputing missing values for the annual variables, interpolating monthly values for the annual variables, and performing the ZCTA weighting procedure for all the variables, the final data set was created. The final data set consisted of a total of 72 months of data over each recruiting flight in the 360th Recruiting Group's zone. The years included in this are based on recruiting years (RY) 2012 through 2017. After examining the final data set, there were three months where the AFRS instructed each group to set the number of recruiting goals at zero. These months were dropped from the data as it is not indicative of normal recruiting procedures. It is important to explain the notation used to represent both the flights, variables, and how time was captured in the data set. This notation is used in the remainder of this research.

Each flight was denoted using the last two numbers of their squadron followed by the letter of the flight to which they belong. For example, flight 11A is the A Flight of the 311th Recruiting Squadron. To capture monthly values for each flight, the month number was then added to the end of the flight notation. The first month of the study for flight 11A was denoted as 11A01. Each flight and month was an observation in the data and had values for each of the 20 variables discussed previously. The notation for each of the variables is $x_{i,j}$ where i is the set of all

Table 4. Consolidated Variables with Variable Notation

Variable Type	Variable Number	Variable Name	Time Unit/Geographic Level	Imputation Required	Geographic Organization Required
Supply Variables	x_1	Total Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	x_2	15-19 Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	x_3	20-24 Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	x_4	Unemployment Rate	Monthly/County	No Imputation Required	ZCTA Weighting
	x_5	Labor Force Participation Rate	Monthly/County	No Imputation Required	ZCTA Weighting
	x_6	Obesity Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation	ZCTA Weighting
	x_7	HS Graduation Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation	ZCTA Weighting
	x_8	Violent Crime Rate	Annual/County	Hot Deck & Stochastic Mean Value Imputation	ZCTA Weighting
	x_9	QMA	Annual/Zip Code	Stochastic Mean Value Imputation	Zip-to-ZCTA Crosswalk
	x_{10}	Voter Participation Rate	Annual/County	No Imputation Required	ZCTA Weighting
Demand Variables	y_1	Goal	Monthly/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	y_2	Recruit Production	Monthly/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	x_{11}	Previous Recruit Production	Monthly/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	x_{12}	Zone Area	2019 Data/Recruiting Flight	No Imputation Required	Zip-to-ZCTA Crosswalk
	x_{13}	Number of Recruiters	Monthly/Recruiting Flight	No Imputation Required	N/A
	x_{14}	Zone Per Recruiter	Monthly/Recruiting	No Imputation Required	Zip-to-ZCTA Crosswalk
Enlistment Choice Variables	x_{15}	Some College Rates	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	x_{16}	Veteran Population	Annual/County	Stochastic Mean Value Imputation	ZCTA Weighting
	x_{17}	Number of Active Duty Personnel	Annual/Zip Code	Stochastic Mean Value Imputation	Zip-to-ZCTA Weighting
	x_{18}	Number of JROTC Units	2019 Data/Zip Code	No Imputation Required	Zip-to-ZCTA Crosswalk

variables included in this research ($i = \{1, 2, 3, ..., 21\}$) and j is the set of all the recruiting flights (j = {11A, 11B,...,39H}). Table 4 shows a consolidated table with all research variables and denotes the variable numbers according to the notation outlined above. The subscripts j and t are omitted because the variables are for each flight and time period.

3.3 Dimension Reduction

In cases where there are a large number of independent or predictor variables there is often a need for dimension reduction because multicollinearity could exist among the data. Multicollinearity occurs when the independent variables are highly correlated with each other and can lead to unstable solutions or overemphasize variables in a solution [36]. In cases where there is multicollinearity, dimension reduction techniques can be used to reduce the number of independent variables and ensure the independence of variables included in the study [36]. Principal Components Analysis (PCA) is a commonly used dimension reduction technique and creates a subset of independent linear combinations of variables to explain as much of the total variance in the data [37].

When extracted from the data, the first principal component will explain the largest proportion of the variance and each subsequent principal component will account for a decreasing amount of variance until the total variance of the data set is explained [37]. For a data set with p independent variables, p independent principal components will be formed. The mth principal components will be in the form:

$$PC_{(m)} = w_{(m)1}X_1 + w_{(m)2}X_2 + \ldots + w_{(m)p}X_p \tag{3}$$

where $w_{(m)1}, w_{(m)2}, \ldots, w_{(m)p}$ are weights that have been chosen to maximize the variance of all linear combinations that are uncorrelated with all previously constructed principal components subject to the constraint $\sum_{j=1}^{p} w_{(m)j}^2 = 1$ [37].

One common use of PCA is feature extraction where the most the important features are retained through PCA and the least important features are dropped [38]. The features that are extracted from PCA are the principal components which are independent linear combinations of the original variables. This makes PCA less interpretable as the principal components act as new variables made up of the loadings of the original variables [38]. Analyzing the component loadings can help interpretation of the principal components and allow variables to be extracted for analysis.

Principal components is calculated using the correlation matrix of the original data, $\mathbf{R}$. The eigenvalues and eigenvectors of $\mathbf{R}$ are found with the largest eigenvalue corresponding to the first principal component and each eigenvalue has a corresponding eigenvector. The loading for the ith variable on the jth principal component is given by $a_{i(j)}\sqrt{l_{(j)}}$, where $l_{(j)}$ is the eigenvalue corresponding to principal component j and $a_{i(j)}$ is the value of variable i in the eigenvector corresponding to the eigenvalue $l_{(j)}$ [37]. The proportion of total variance explained by principal component j is given by $l_{(j)}/p$, where $l_{(j)}$ is the eigenvalue corresponding to the jth principal component and p is the number of variables in the original data set [37].

Although it is possible to extract p principal components from a data set with p variables, it is desired to account for most of the total variance with as few principal components as possible [37]. To determine how many principal components to retain, this research uses the methods of the amount of cumulative variance explained and the scree plot [37]. The first technique identifies the principal component where a pre-identified percent of variance is explained. This research sets the amount of variance explained to 80%, so all principal components up to the principal component where the amount of cumulative variance explained is 80% are retained. The second measure is the scree plot. A scree plot is graphical representation of the amount of variance explained in each component [36]. The scree plot will plot the principal component number with the eigenvalue. A good representation of how many components to extract is where an elbow or bend occurs on the plot. This represents that a large decrease in the amount of variance explained is contained in the principal component where this elbow occurs. At this elbow, the remaining principal components should not be extracted from the data as the proportion of variance explained by the remaining components is much less than the components before the elbow.

After the number of principal components to retain is determined, the loadings of each variable on these principal components can be used to extract important variables from the data set. This research extracts the variables with the largest positive and negative loadings from each principal component retained. Because the principal components are linearly independent, this method of identifying variables to retain in the study will aid in minimizing any multicollinearity in the research.

3.4 Regression Analysis

Multiple Linear Regression.

Multiple linear regression is a statistical modeling technique used to investigate the relationship between a response variable and a set of predictor variables or regressors [30]. The form that a multiple linear regression model takes is shown in equation 4

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{4}$$

where y is the response variable, the β values refer to the regression coefficients, x values are the regressor variables, the subscript k is the number of regressor variables included in the model, and ϵ is a random error component [39]. The regression parameters, β_j, represent the expected change in the response variable, y, for a one unit change in the regressor variable, x_j, when all other regressor variables are held constant [39]. This is called a linear regression because the model is linear in the regression parameters and not necessarily because the relationship between the response and predictor variables is linear [39]. A multiple linear regression model can also include more complex features including interactions between regressor variables and regressor variables raised to a power. For example, equation 5 is a second-order linear regression model with interactions [39].

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \tag{5}$$

This is still considered a multiple linear regression model, but can capture more complex relationships between the response variable and regressor variables.

When an independent variable is a qualitative variable, it can be incorporated into the regression model using an indicator variable [39]. Indicator variables can be represented as binary variables, 0 or 1, indicating whether the variable is at a certain level or has a certain attribute. When using binary indicator variables, a 0 indicates that the variable does not have that attribute and 1 indicates that it does have that attribute. For qualitative variables with a distinct levels, $a-1$ indicator variables are used to represent all levels of the qualitative variable [39]. In this research, indicator variables are used to represent the recruiting flight. Because there are 60 recruiting flights in this research, 59 indicator variables are used to represent the flights.

Indicator variables can be incorporated into the regression model in two manners. The first is to incorporate the qualitative variable as its own variable in the model. For example, consider the regression equation in 6

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{6}$$

where β_o is the intercept of the regression equation, x_1 is a quantitative variable with the coefficient β_1, and x_2 is a binary indicator variable with coefficient β_2. When $x_2 = 1$, the intercept of the regression model is shifted by β_2 creating two parallel regression lines with a slope of β_1. The new regression model is shown in equation 7.

$$y = (\beta_o + \beta_2) + \beta_1 x_2 + \epsilon \tag{7}$$

Where there are multiple indicator variables present in the model, the difference between the two levels of indicator variables can be measured by subtracting the coefficients of the indicator variables [39].

The second method to incorporate indicator variables into a regression model is for the indicator variable to interact with a quantitative variable. The interaction between the qualitative and quantitative variables indicate that the level of the qualitative variable will change the slope of the quanitative variable. For example, consider the following regression equation in 8

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \tag{8}$$

where β_o is the intercept of the regression equation, x_1 is a quantitative variable with the coefficient β_1, and x_2 is a binary indicator variable with coefficient β_2. Where there is an interaction between the indicator variable and quantitative variable, both the slope and intercept of the model are shifted. This is illustrated in equation 9.

$$y = (\beta_o + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon \tag{9}$$

This situation can be extended to instances where there are multiple qualitative levels and quantitative variables in the model.

The regression coefficients are estimated using the method of least squares estimation. This method is used to estimate the regression coefficients to ensure that the sum of squares from the differences between the observations, y_i, and the straight line fit to model the data is a minimum [39]. The calculation for the β coefficients can be represented in matrix notation by equation 10

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{10}$$

where $\boldsymbol{y}$ is a n x 1 vector of the observations, $\boldsymbol{X}$ is a n x $(k+1)$ matrix of the regressor variables, $\boldsymbol{\beta}$ is a $(k+1)$ x 1 vector of regression coefficients, and $\boldsymbol{\epsilon}$ is a n x 1 vector of random errors [39]. The least squares normal equations which are used to minimize the distance between the observation and line is represented in matrix notation by equation 11 [39].

$$\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'y} \tag{11}$$

Solving equation 11 for $\boldsymbol{\hat{\beta}}$ yields the estimation for the regression coefficients[39].

$$\boldsymbol{\hat{\beta}} = \boldsymbol{(X'X)^{-1}X'y} \tag{12}$$

Using least squares estimation, the estimated coefficients found in equation 12 are unbiased estimators for the actual value of the parameter [39].

Hypothesis Testing.

After estimating the parameters for the multiple linear regression model, hypothesis testing is used to determine the overall significance of the regression and the significance of the individual parameters in the model. Testing for significance of the regression is used to determine if a linear relationship exists between the response y and the regressor variables x that are included in the model [39]. The hypothesis test used for this is:

$$H_o : \beta_o = \beta_1 = \beta_2 = \cdots = \beta_k$$
$$H_A : \beta_j \neq 0 \text{ for at least one j}$$

The null hypothesis is testing if all regression coefficients are equal to zero and the alternative hypothesis is testing that at least one of the regression coefficients does not equal zero. If the null hypothesis is rejected, this indicates that a linear relationship exists. To determine if the null hypothesis is rejected, an analysis of variance (ANOVA) table is used to determine an F-statistic for the model. In the ANOVA, the total sum of squares is partitioned into a sum of squares due to regression and sum of squares due to residuals [39]. Using an ANOVA table, shown in Table 5, the F-statistic, F_o is found [39]. F_o follows the $F_{k,n-k-1}$ distribution and the null

Table 5. Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_o
Regression	SS_r	k	MS_r	$\frac{MS_r}{MS_{res}}$
Residual	SS_{res}	$n-k-1$	MS_{res}	
Total	SS_T	$n-1$		

hypothesis is rejected if $F_o > F_{\alpha,k,n-k-1}$ for a specified value of α [39]. If the null hypothesis is rejected for this hypothesis test, it is then important to test the individual regression coefficients to determine which coefficient is significant in the model.

The hypothesis test for individual coefficients tests to determine if each individual regression coefficient is equal to zero. The hypothesis test is formulated as:

$$H_o : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

If the null hypothesis is rejected in this test, the regression coefficient is signficant in the model and should remain as part of the model. If the null hypothesis is not rejected, then the regressor can be removed from the model [39]. The test statistic for this hypothesis is found using equation 13.

$$t_o = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}} \tag{13}$$

To determine if the null hypothesis should be rejected or not, t_o is compared to $t_{\alpha/2,n-k-1}$ for a specified value of α. If $|t_o| > t_{\alpha/2,n-k-1}$, the null hypothesis is rejected and the regression coefficient is significant in the model [39]. A primary challenge in developing regression models is how to select the variables that should be included in the model. The next section discusses the various methods of variable selection.

Variable Selection.

The variable selection problem is the process of selecting the appropriate subset of all the candidate regressors available to include in the model [39]. The model building process aims to achieve the two goals of having as many regressors as possible so that the information content in the model influences the predicted response variable and to have as few regressors as possible to limit the variance of the prediction [39]. Stepwise regression is set of procedures that aid in the variable selection problem.

Backwards elimination starts with the assumption that all candidate regressors are included in the model [39]. A partial F-statistic is calculated for each candidate regressor under the assumption that it was the last regressor to enter the model. This partial F-statistic is then compared to a predetermined F-value and if the partial F-statistic is less than the F-value, the candidate regressor is removed from the model. Backwards elimination is completed when there are no candidate regressors that have a partial F-statistic that is less than the predetermined F-value.

Mixed stepwise regression is a variable selection technique where the procedure begins with the assumption that there are no regressors in the model. Mixed stepwise regression adds a regressor into the model if the F-statistic for the model with the added regressor is greater than a pre-determined F-statistic or significance level [39]. The first variable added to the model will be the regressor with the highest correlation with the response variable as this will produce the highest F-statistic for the model. As each variable is added to the model, mixed stepwise regression will also reevaluate all regressors currently in the model to ensure they are still significant, similar to backwards elimination [39]. If a variable is no longer significant after a regressor is added, this variable is dropped from the model. This procedure continues until all candidate regressors are assessed.

These techniques are useful in identifying which candidate regressors to add to the model, but can result in different models, when used separately [39]. In some instances, it is beneficial to use a combination of the stepwise techniques. The authors of [39] recommend using mixed stepwise regression followed by backwards elimination. After the model is built using these variable selection procedures, model fit and performance is assessed to determine the strength of the model.

Model Fit and Performance.

After a model is built and hypothesis testing determines that the model is significant and the regressors in the model are significant, the overall fit or performance of the model must be assessed. This is accomplished using multiple model summary statistics. The first set of model statistics used is the R^2 or coefficient of determination values. This value is measure of the proportion of variance explained by the model [39]. R^2 is the ratio of the sum of squares of the regression to the total sum of squares as shown in equation 14 [39].

$$R^2 = \frac{SS_r}{SS_t} = 1 - \frac{SS_{res}}{SS_t} \tag{14}$$

Because regressors are added to the model R^2 will never decrease, adjusted R^2, R^2_{adj}, is often used to prevent overfitting the model or adding terms that are not helpful [39]. R^2_{adj} accounts for the additional variables added and will only increase if the added regressor decreases the residual mean square. Equation 15 gives the from of the R^2_{adj} [39].

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-p)}{SS_t/(n-1)} \tag{15}$$

The final R^2 value used in assessing the performance of the model is called the R^2 for prediction, R^2_{pred}. An important use of a regression model is its ability to predict the response value for new observations. R^2_{pred} gives an indication of the model's predictive capability using the PRESS statistic which is a measure of how well the model will predict new data [39]. The PRESS statistic is given by

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2 \tag{16}$$

and is used in the R^2_{pred} given by equation 17 [39].

$$R^2_{pred} = 1 - \frac{PRESS}{SS_t} \tag{17}$$

Small values of the PRESS statistic are desired which will increase the R^2_{pred} indicating that model predicts new observations well.

The second method of assessing model fit is to identify if there is multicollinearity present in the model. Multicollinearity indicates that the regressors in the model are correlated and may degrade model performance. When multicollinearity exists in the model, the estimates of the coefficients are too large, the estimates could change when using a different subset of the data, and the model will be a poor predictor of the response [39]. Variance inflation factors (VIFs) indicate if strong multicollinearity is present in the model. VIFs are given by equation 18

$$VIF_j = \frac{1}{1 - R_j^2} \tag{18}$$

where R_j^2 is the R^2 value when the variable j is regressed on the other $(p-1)$ variables [39]. In instances where VIFs are larger than ten, there is evidence of strong multicollinearity in the model [39]. When multicollinearity is present, dimension reduction can be used prior to building the model to aid in selecting independent variables or the variables with VIFs larger than ten are dropped from the model.

Model Adequacy.

Regression analysis utilizes five assumptions when a regression model is built. A regression model must meet these five assumptions or the model may be inadequate or unstable [39]. The five major assumptions are shown in Table 6. Model adequacy is

Table 6. Assumptions of Linear Regression

1.	The relationship between the response y and the regressors is at least approximately linear
2.	The error term, ϵ, has zero mean
3.	The error term, ϵ, has constant variance
4.	The errors are uncorrelated
5.	The errors are normally distributed

the process of checking that the built model satisfies these assumptions. The methods of assessing model adequacy include residual analysis, transformations of the model to satisfy the assumptions, and detecting and treating leverage and influence points.

Residual Analysis.

The residuals of a linear regression model are defined as the difference between the true value of the response and the predicted value of the response generated by the model [39]. The residual is given by the form:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, ...n. \tag{19}$$

The residual is a measure of the variability of the response variable not explained by the regression model [39]. The residuals defined in equation 19 are referred to as the raw residuals. Scaling residuals is useful in regression analysis to identify outliers or extreme values [39]. There are many different methods of scaling residuals and this research utilizes the studentized residuals given by equation 20. Studentized residuals have constant variance equal to one regardless of the location of $\boldsymbol{x_i}$ [39].

$$r_i = \frac{e_i}{\sqrt{MS_{res}(1 - h_{ii})}} \tag{20}$$

The studentized residual uses the hat values, h_{ii}, which is a measure of where the point lies in the x-space. As the point gets further away from the center of the data, the raw residual will get smaller because the point will pull the regression line towards the point. The studentized residual is useful because when a point has a large residual and large hat value, it may be an outlier and should be examined further.

Examining the studentized residuals using residual plots is necessary in determining model adequacy. A normal probability plot is used to determine if the studentized residuals are normally distribution, satisfying assumption 5. A normal probability plot is designed so that the cumulative normal distribution plots is a straight line [39]. When the plot is not a straight line, this indicates non-normality of the residuals [39]. Although slight deviations from this straight line do not drastically affect the model, large deviations could impact the hypothesis testing used to evaluate the model [39].

The second type of residual plot used to ensure model adequacy is a plot of the studentized residuals against the predicted response [39]. The studentized residuals against predicted response plot assesses how well the model meets assumptions 2 and 3. If the model satisfies these assumptions the studentized residuals will be spread between two horizontal bands, shown in Figure 4(a), indicating that the variance is constant [39]. If the studentized residuals are in an outward funneling pattern or in a bowing pattern, shown in Figure 4(b) and Figure 4(c), this is evidence of nonconstant variance and the assumption is not met [39]. If the studentized residuals follow a curved pattern, shown in Figure 4(d), this is evidence of a nonlinearity and a higher order term may be required in the model [39]. When there is evidence

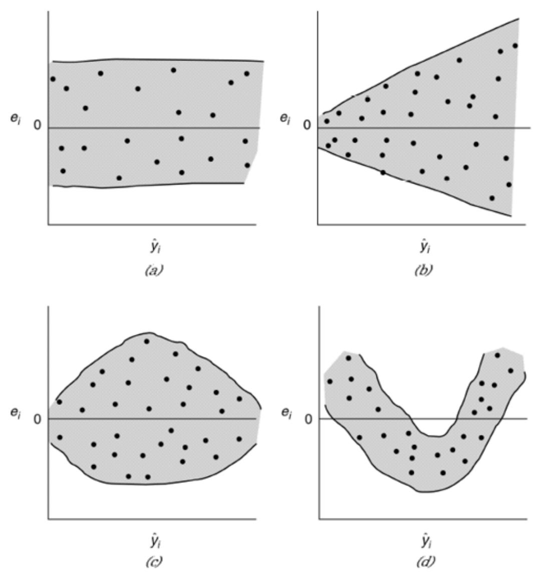


Figure 4. Example Residual Plots

of nonconstant variance in the studentized residuals, oftentimes there is need for a transformation of the response.

The final model adequacy check is to determine if there is autocorrelation in the residuals. When the data used in building a regression is time series data, autocorrelation is the presence of correlated error terms [39]. When autocorrelation is present in the model, a lagged response variable is included to ensure the error terms are uncorrelated. This lagged response variable is denoted as $y_{t-\tau}$ with a coefficient of ϕ_τ [40]. After including a variable for a lag one response, y_{t-1}, there may be a requirement to add successive lag variables to ensure the residuals are uncorrelated.

The Durbin-Watson test is a statistical test for the presence of positive autocorrelation. In this test, the null hypothesis, H_o, is that there is no autocorrelation present in the data [39]. The Durbin-Watson test statistic is calculated using equation 21,

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \tag{21}$$

where e_t, $t = 1, 2, ..., T$ are the residuals from a regression.The test statistic d is used to determine if there is positive autocorrelation in the data, while testing for negative autocorrelation uses the test statistic $4 - d$. When the errors are uncorrelated, the value of d will be approximately two, but statistical testing is necessary to determine how far the value of d is from two to fail to reject the null hypothesis [39]. Durbin and Watson determined limits for the test statistic, d, where if the test statistic is less than the lower limit, d_L, the null hypothesis is rejected and if the test statistic is greater than the upper limit, d_U, the null hypothesis is not rejected [39]. Where the test statistic falls between the upper and lower limits, the test is inconclusive [39].

Transformations of the Response Variable.

Where there is non-constant variance of the studentized residuals, a transformation on the response variable is often necessary. Correcting non-constant variance in the model gives more precise estimates of the parameters in the model and increases the sensitivity for the statistical tests [39]. This research utilizes the Box-Cox method to determine the appropriate transformation of the data. The Box-Cox method utilizes the power transformation, y^λ and determines the best value of λ where the residual sum of squares is minimized [39]. To determine this value, the Box-Cox method fits models with various values of λ and plots the residual sum of squares for each model. The value of λ corresponding to the minimum residual sum of squares is selected as the value for the transformation [39]. The response of the transformed model is then y^λ and the studentized residuals are reassessed to ensure that they meet the assumption of non-constant variance. The final adequacy check is to detect influence and leverage points.

Influence and Leverage Points.

Regression models are desired to be representative of all data points available, but in some instances a subset of data points may influence the model [39]. An influence point in a regression model has a noticeable impact on the model coefficients by pulling the regression line in that direction [39]. A leverage point is an unusual x value and may control certain model properties [39].

When detecting influence points, it is valuable to consider both the location of the point in x space and the response variable [39]. Cook's D is a method to detect influential points by measuring the squared distance between the estimate of $\hat{\boldsymbol{\beta}}$ on all n points and the estimate obtained by removing the i^{th} point ($\hat{\boldsymbol{\beta}_{(i)}}$) [39]. Points with a Cook's D value of greater than 1 indicate an influential point and may have considerable influence on the estimates of the coefficients [39].

Detecting leverage points utilizes the diagonal elements of the hat matrix, h_{ii} [39]. The diagonal elements of the hat matrix is a standardized measure of the distance of a point from the center of the x space [39]. The values of h_{ii} are compared to the value of $2p/n$, where p is the number of regressors in the model and n is the number observations. Values that exceed this measure are considered leverage points. Some values that are identified as leverage points may not have a large impact on the model, so it is oftentimes beneficial to identify points that have both a large studentized residual and a large hat value to determine the impact on the model [39].

Model Validation.

Model validation is the process of determining if the model will function properly for its intended use [39]. The first step in assessing model validity is to analyze the regression coefficients and the predicted responses for correct sign and magnitude [39]. If the regression coefficients are either too large or have the opposite sign than expected the model may be in the incorrect form or multicollinearity is present. Additionally, if the predicted response has the wrong sign or magnitude this could indicate that the model has the incorrect form or incorrect estimations of the regression coefficients [39]. Another method to assess model validity is data splitting where the full data set is divided into an estimation set and a validation or test set [39].

When the data is split into the estimation set and test set, the estimation set is used to build the regression model and the data in the test are used to determine the model's performance in predicting the response [39]. There are multiple ways to split the data into the two sets and for this research the data is split based upon time. The data set used in this research contains data for RY2012-2017, so the estimation set contains the data for RY2012-2016 and the test set contains the data for RY2017. This method of splitting the data gives 3,360 observations in the estimation set and 720 observations in the test set.

When the regression model is built with the estimation set and validated on the test set, there are multiple measures to use to determine the model's validity. The first measure is the average squared prediction error shown in equation 22

$$\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N} \tag{22}$$

where N is the number of observations in the test set used to validate the model [39]. This measure is then compared to the residual mean square which is a measure of average variance of the residuals and if the value from equation 22 is larger than the residual mean square, this is an indication that model does not predict new responses as well as it fits responses in the estimation set [39].

A second method to assess model validity to use the R^2 and R^2_{pred} values presented previously. These values can be compared to measure the difference in prediction capability. A small loss in prediction capability indicates that the model is able to predict new values reasonably well in comparison to the actual model.

3.5 Conclusion

This chapter detailed the data gathering, cleaning, and organization process, the methodology used to formulate a model, and to ultimately validate the model's performance and adequacy. The subsequent section will apply this methodology to the data set built for this research with the goal of determining demographic or econometric factors that contribute to recruiting goaling.

IV. Results and Analysis

4.1 Introduction

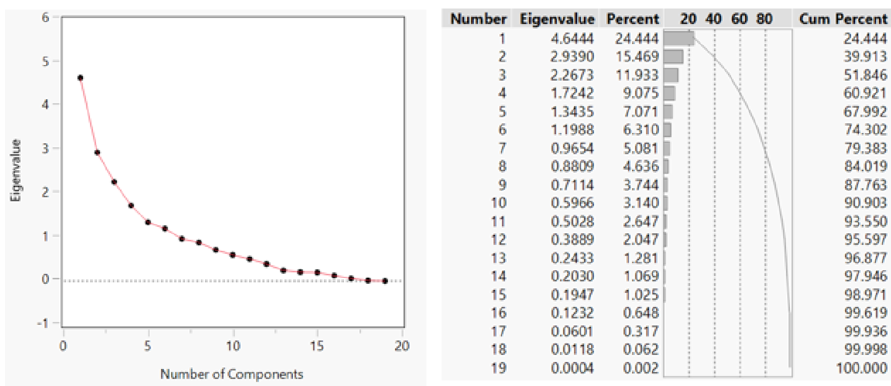
To address the research objective two linear regression models are used to identify characteristics that contribute to recruit goaling and production and used to create projections for the 360th Recruiting Group for RY2018-2021. Although these regressions were formulated using data at the flight level, the results and analysis in this section is conducted at the squadron level to compare goals and production for all enlisted recruiting squadrons in the 360th Recruiting Group. This chapter is organized into three sections: the first section will detail the formulation and results of the recruit goaling linear regression, section two will discuss the formulation and results of the recruit production model, and the final section will discuss the projections for both models.

4.2 Recruitment Goaling Model

Principal Components Analysis.

Beginning the model formulation process with 20 possible variables to include in the model, there was a risk of introducing multicollinearity in the model. To prevent multicollinearilty in the model formulation process, PCA was conducted on the final data set to determine how many dimensions can explain a preset amount of the variance in the data set and also to identify candidate regressors to include in the linear regression. For this iteration of PCA, the variables for the number of recruits goaled in a month (y_1) and the number of recruits produced in a month (y_2) were withheld. In PCA, the dependent variable is withheld from the analysis and it is logical to withhold the number of recruits produced as well because it is not reasonable to base the number of recruits goaled in a month to the number of recruits produced in that month as it is unknown at the time. A similar technique to PCA is factor analysis where the principal components are rotated to produce more interpretable results with the loadings now representing the correlation coefficient of the variable and the factor [41]. After conducting PCA and factor analysis, the variable loadings were not considerably different, so PCA was used to identify a subset of uncorrelated variables.

After conducting PCA in JMP, the number of components to utilize in the study was first determined using a scree plot and the component at which 80% of variance is explained. Analyzing the scree plot in Figure 5(a), shows a slight elbow at around the sixth principal component. This indicates that at around principal component six,



(a) Goal Scree Plot

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.6444	24.444		24.444
2	2.9390	15.469		39.913
3	2.2673	11.933		51.846
4	1.7242	9.075		60.921
5	1.3435	7.071		67.992
6	1.1988	6.310		74.302
7	0.9654	5.081		79.383
8	0.8809	4.636		84.019
9	0.7114	3.744		87.763
10	0.5966	3.140		90.903
11	0.5028	2.647		93.550
12	0.3889	2.047		95.597
13	0.2433	1.281		96.877
14	0.2030	1.069		97.946
15	0.1947	1.025		98.971
16	0.1232	0.648		99.619
17	0.0601	0.317		99.936
18	0.0118	0.062		99.998
19	0.0004	0.002		100.000

(b) Goal Cumulative Variance Table

Figure 5. Goaled Model PCA Results

there is a slight decrease in the amount of variance explained in the subsequent principal component. Additionally, considering Figure 5(b) indicates that the cumulative variance explained reaches 80% between components seven and eight. Considering these results, this research utilizes the first eight principal components to determine the candidate regressors to include in the linear regression model.

Extracting the first eight principal components and analyzing the loading matrix in Figure 6, candidate regressors are identified using the highest positively and negatively loaded variables, above a value of 0.3, on each principal component. This process was followed for all eight principal components with slight deviations if the variable with the highest loading was already included or if the information from this variable was already captured in a previously selected variable. For example, the highest loaded variable on the second principal component was the variable detailing the proportion of 15-24 year olds in an area and on the fourth principal component, the variable relating the proportion of 15-19 year olds was the highest loaded variable. In this instance, age of the population was already capture, so the variable of 15-19 year olds was not extracted from the principal component. This process identified 11 candidate regressors, shown in the red boxes in Figure 6, to include in the first iteration of the modeling process.

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
Previous Month Production	0.02665	0.03157	0.55393	-0.23740	-0.04938	0.43209	-0.17638	0.21378
Num Recruiters	-0.07329	0.07770	0.45141	-0.12628	0.10855	0.71197	0.13363	0.12451
QMA	-0.90711	0.15316	-0.02458	0.05400	-0.12497	-0.00631	0.16745	0.18954
Unemployment Rate	0.12594	0.38063	-0.29293	-0.40750	0.59323	-0.03037	0.09632	0.03204
Labor Force Participation Rate	-0.28361	-0.28932	0.51877	0.35028	0.33892	-0.24567	0.37372	0.06752
Total Pop	-0.87028	0.20201	-0.15810	-0.15315	-0.23263	0.04097	0.07445	0.14522
15-19 Proportion	0.26187	0.53488	0.01574	0.58394	0.23659	0.18204	-0.10987	0.19910
20-24 Proportion	0.21138	0.80154	0.19815	0.31210	-0.27693	-0.10759	0.05709	-0.09524
15-24 Proportion	0.26417	0.81561	0.15518	0.46868	-0.11301	-0.00802	-0.00197	0.00870
Proportion Vet	0.44941	-0.19541	0.56694	0.03309	0.35433	-0.18393	0.39297	-0.09916
Zone Area	0.75118	-0.09151	-0.11620	-0.09786	-0.29075	0.05204	0.30377	0.43089
Obesity Raw	0.80193	0.17862	-0.00845	-0.17153	0.17250	-0.04554	-0.29414	-0.10595
HS Raw	0.15111	-0.57835	0.04075	0.28126	-0.14308	0.37046	0.10031	-0.20911
SC Raw	-0.66527	-0.25725	0.35721	0.26768	-0.15769	-0.25565	-0.16536	0.10585
VC NUM	-0.54399	0.48555	-0.04469	-0.25035	0.33993	-0.07951	0.07963	0.27580
Voting Participation	0.24003	-0.41413	0.35769	0.11679	0.15191	-0.22620	-0.50526	0.44341
Sponsor	0.16296	0.29965	0.62950	-0.35227	-0.28173	-0.14323	0.05131	-0.19618
JROTC Det	-0.01772	0.23401	0.51184	-0.52884	-0.20672	-0.21039	-0.06999	-0.02924
Zone/Rec	0.72363	-0.12015	-0.24525	-0.08801	-0.31846	-0.18431	0.28078	0.36688

Figure 6. Goal PCA Loading Matrix

Model Formulation.

The 11 candidate regressors identified in PCA and the 59 indicator variables representing each individual flight were used as the candidate regressors in the first linear regression model. Mixed step-wise regression followed by backwards elimination at a significance level of $\alpha = 0.05$ identified which variables were statistically significant in the model. After an initial residual analysis of the studentized residuals, the studentized residuals displayed an outward opening funneling pattern indicating non-constant variance. Applying the Box-Cox method to transform the dependent variable, identified a quarter root transformation would minimize the sum of squared errors. This transformed dependent variable, $\sqrt[4]{y_1}$, was applied in the mixed stepwise regression followed by backwards elimination procedure to identify the variables signficant in the model.

This process identified five independent variables statistically significant at $\alpha = 0.05$. These five variables included the obesity rate (x_6), the high school graduation rate (x_7), the voting participation rate (x_{10}), the number of recruiters in a flight (x_{13}), and the number of JROTC detachments in an area (x_{18}). Analyzing the leverage plots in JMP for these 5 variables indicated that there were non-linearities in the obesity rate and the voting participation rate. To account for these non-linearities, the square of these variables were included in the model and found significant at $\alpha = 0.05$. Table 7 shows these variables with their respective coefficient, p-value,

Table 7. Goal Model Parameters ($\sqrt[4]{y_1}$)

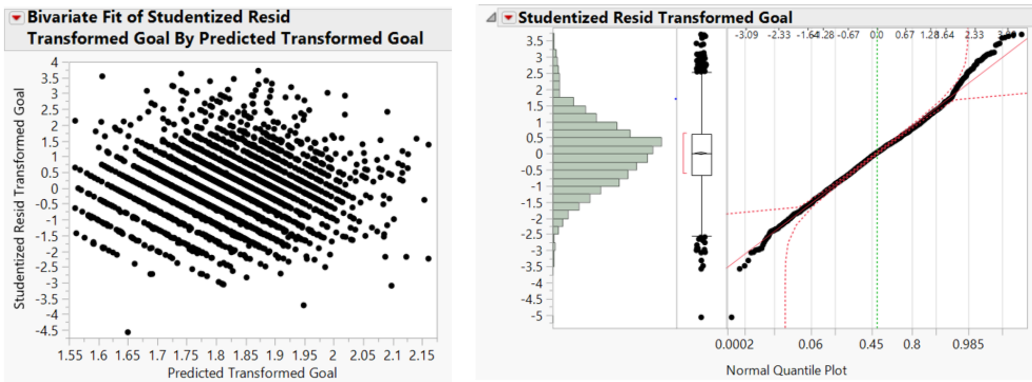
Variable	Coefficient	P-Value	VIF
x_6	0.627	< 0.0001	2.98
x_6^2	-7.381	0.0029	2.88
x_7	0.296	< 0.0001	2.46
x_{10}	0.146	0.005	2.62
x_{10}^2	-1.074	0.0082	4.52
x_{13}	0.026	< 0.0001	1.45
x_{18}	0.0056	< 0.0001	2.09

and VIF. In addition to these variables, various indicator variables representing each flight were found signficant giving an individual regression equation for each flight by manipulating the intercept of the linear regression model. These were not included in the Table 7, but the individual model results are shown in Appendix C.

The coefficients in Table 7 provide insight into how the mean value of the number of goals will change based on a one unit change for each of these variables when all other varaibles are held constant. Interpreting the signs of these coefficients can identify which factors are related to increased goals or decreased goals. The coefficient for the obesity rate is positive, but the coefficient for the squared obesity is negative and much higher in magnitude. When this is applied in the linear regression model, the overall affect is negative indicating that areas with higher obesity rates are correlated with lower monthly goals. The result is similar for voting participation rates. These coefficients indicate that for an increase in the voting participation rate, the number of recruits goaled in that zone would decrease. The high-school graduation rate, number of recruiters, and the number of JROTC detachments in an area have a positive affect on the number of recruits goaled in a zone indicating that higher levels of these variables are correlated with higher monthly goals for a recruiting flight. Including the intercept for each flight gives a final regression model for each flight allowing comparisons between flights and squadrons in the 360th Recruiting Group.

Model Adequacy.

With a model formulated for recruit goaling, model adeqeuacy was assessed to ensure that the generated model met all assumptions of linear regressions shown in Table 6. The first check was to conduct residual analysis on the studentized residuals of the model using residual plots to determine if the model meets assumptions 3 and 5 of regression analysis. Figure 7(a) shows the studentized residual vs. predicted goal plot. This figure indicates that the residuals display constant variance without a pattern associated with them which satisfies assumption 3 of regression analysis. Figure 7(b) shows the normal probability plot of the studentized residuals. This plot shows that the studentized residuals are reasonably symmetric and normally distributed. This plot shows a slight tail at the upper end, but still falls on the diagonal line



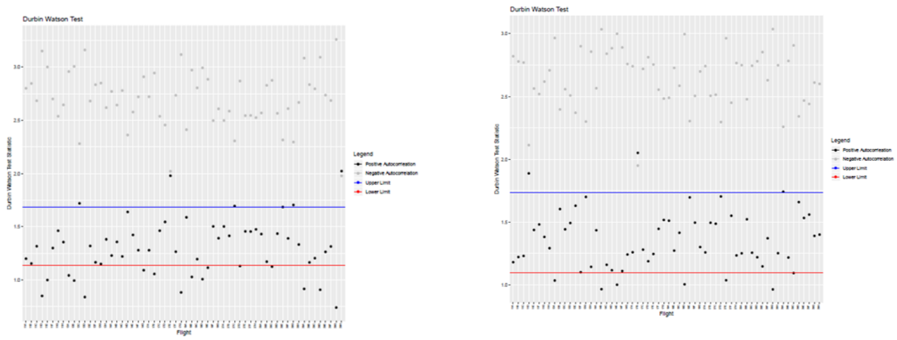
(a) Studentized Residuals vs. Predicted Produced

(b) Normal Probability Plot

Figure 7. Goaled Model ($\sqrt[4]{y_1}$) Residual Analysis

showing that the residuals are normally distributed satisfying assumption 5 of linear regression.

The final assumption that is checked using residual analysis is assumption 4 indicating whether the error terms are uncorrelated. The Durbin-Watson test for autocorrelation was used to determine if there were any flights that had correlated error terms. Because there are 60 flights included in this analysis, the residuals for each flight had to be broken out and a Durbin-Watson test statistic calculated for each flight. This was accomplished using the R function durbinWatsonTest found in the "car" package [42]. In order to quickly visualize and perform the Durbin Watson test, each value for the test statistic, d, was plotted with lines indicating the upper and lower limits for the test. Figure 8(a) shows the Durbin-Watson test for the originial model with the solid black dots indicating the test statistic values of d, the grey dots indicating test statistic value for $4 - d$, the red line representing the lower limit for the test at 1.134, and the blue line showing the upper limit for the test at 1.685 [43]. Where the values of d fall below the red line indicate that there is autocorrelation present in the residuals and the values of d that fall above the blue line indicate that



(a) Durbin Watson Test (b) Durbin-Watson Test with y_{t-1}

Figure 8. Durbin-Watson Test

there is no autocorrelation in the model. Where the d statistic falls in between the two limits indicates that the test is inconclusive. For all flights with values of d either in between the two limits or above the blue line, this research considers autocorrelation to not exist among the residuals. There is presence of autocorrelation in many of the flights which does not satisfy assumption 4. To correct this issue, the lagged-1 response variable, y_{t-1}, is included in the model. To determine the value of ϕ_1, the flight's indicator variable is crossed with this variable.

Figure 8(b) shows a similar chart to the original Durbin-Watson test, but with the test statistics, d, from the model including y_{t-1}. This figure indicates that all but a few flights demonstrate autocorrelation in the residuals and for the purpose of this research this is considered sufficient to satisfy assumption 4 of linear regression.

The final model adequacy determination is to determine if there are any influence or leverage points within the data. Using Cook's D value to identify any influence points and comparing to the test value of 1, this model does not display any influential points in the data. There are some values of Cook's D that are less than and greater than the rest of the Cook's D values; however, the corresponding residual and hat value are not excessively large indicating that this point is not considered an influence point. Using the diagonal values of the hat matrix and comparing to the value of $2p/n$, there are some values that are greater than this test value, but the residuals for these points are not much greater than the other residuals, concluding that these should not be considered leverage points.

Model Fit and Validation.

After ensuring the model meets all assumptions of linear regression, model fit and validation are assessed to ensure that the model performs as desired. To assess the model fit, a variety of model summary statistics are used including R^2, R^2_{adj}, R^2_{pred}, the VIF values, and the mean square error. These model summary statistics are shown in Table 8. The different R^2 values for this model are all relatively low. Although

Table 8. Goal Model Summary Statistics

Statistic	Value
R^2	0.44
R^2_{adj}	0.42
R^2_{pred}	0.41
MSE	.011

high values of the R^2 values are desired, research studies pertaining to individuals and demographic data are difficult to have R^2 that are high due to the nature of the noisy data and difficulty in modeling studies involving individuals.

The VIF values for each of the independent variables included in the study are reported in Table 7. These values are all less than 10, indicating that multicollinearity is not present in the model. Although not shown in the table, the VIFs for the indicator variables representing the flights and the lag variables, y_{t-1} are all under 10 indicating that multicollinearity is not present in the model.

The data used to build the model included all data for RY2012-2016, with the data for RY2017 withheld as a validation set for model. Using the average squared prediction error shown in equation 22 for the RY2017 data and comparing to the MSE for the data used to create the model gives an indication of how well the model performs. The average squared prediction error for the data in RY2017 is 0.0099 and the MSE for the built model is 0.01. These values are very close indicating that the model performs well on data that is not included in the model building set. As a second measure of model validation, the R^2 and R^2_{pred} are compared to determine the loss in prediction power between the two data sets. The loss in prediction ability is small with the difference in the two values around 0.03. This indicates that while the validation set does lose some prediction ability it is not drastic, concluding that the model performs well outside the scope of the original data.

Model Results.

The model presented in the previous sections determines goals based on characteristics of each recruiting zone. This predicted goal from the model allows for comparisons between each of the 360th Recruiting Group's squadrons to determine which squadron has a potential for higher goals than the other squadrons based on the composition of the zone. Figure 9 shows the predicted goals from the model presented in the research with the actual goals from RY2012-2017.

The predicted goals from the model presented in this research are shown in the red line. This model shows a much more consistent goaling procedure throughout the months associated with this research. The actual goals, shown in black, are the actual goals that were levied on the squadrons. There is much more variability in these goals as shown by the spikes throughout the years. The proposed model reduces the variability in the goaling procedure as the goals are determined by the demographics of the zone which do not change drastically from month to month allowing a more consistent goaling procedure.

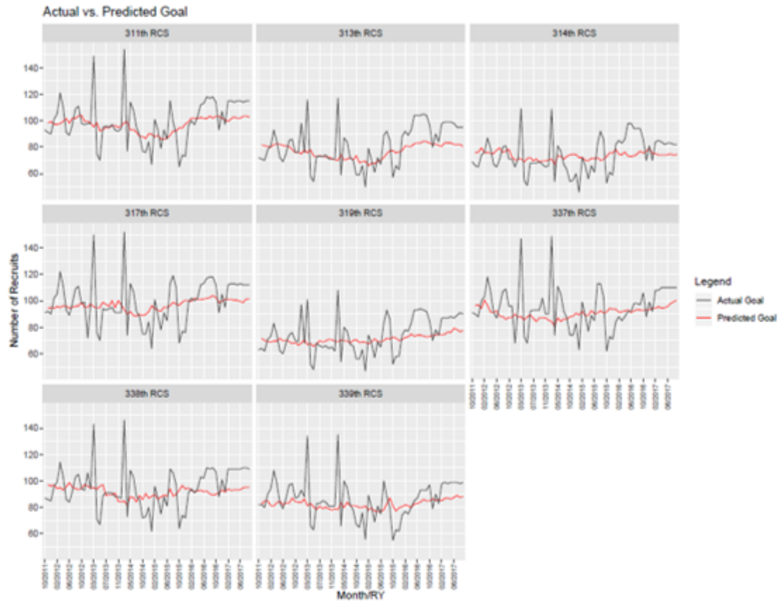


Figure 9. Actual vs. Predicted Goals for RY2012-2017 by Squadron

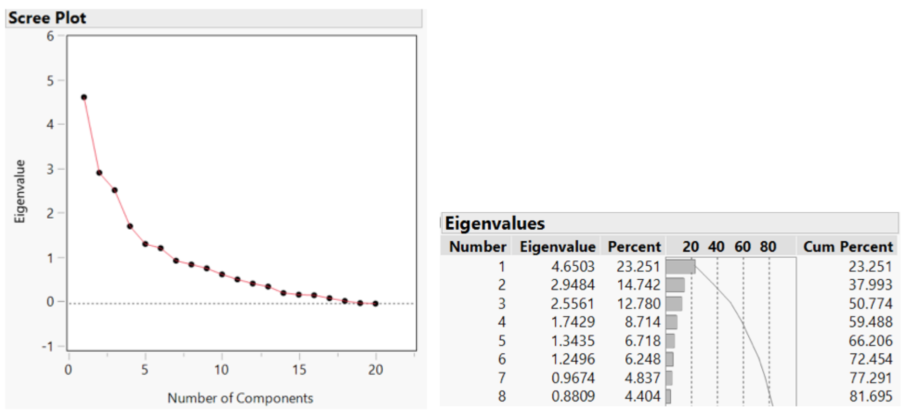
This chart also shows which squadrons have a higher goaling potential due to the composition of their zone. Figure 9 shows that the 311th, 317th, and 337th Recruiting Squadrons have a higher potential for goals than the other squadrons in the group. The goaling procedure based on zone demographics shows that these squadrons are consistently at higher goaling potential than the other squadrons with less variability. Additionally, from Figure 9 the 360th Recruiting Group can identify areas that have consistently lower goaling potential including 311th RCS and the 319th RCS. These squadrons are lower goaled throughout the months included in this study, but with smaller spikes in the data. This goaling distribution method allows the 360th Recruiting Group to identify which squadrons should receive a larger or smaller monthly goal based on zone characteristics.

This analysis can also be applied to the lower levels of the recruiting echelons. A demographic goaling model allows each squadron to compare the flights to determine which flight has a higher goaling potential. This analysis is not included in this chapter, but graphs showing the predicted goals for each flight is shown in Appendix A. To determine how these goals affect recruiting production for the 360th Recruiting Group, a similar modeling methodology is applied to analyze recruit production.

4.3 Recruitment Production Model

Principal Components Analysis.

This research develops a model of recruit production to identify demographic factors of zones that contribute to recruit production and to determine how the developed recruit goaling model affects recruit production within the 360th Recruiting Group. PCA was first used to reduce the number of dimensions and identify candidate regressors to include in the model. The PCA for the production model included all variables except the dependent variable of recruit production, y_2. The results of



Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	4.6503	23.251	23.251
2	2.9484	14.742	37.993
3	2.5561	12.780	50.774
4	1.7429	8.714	59.488
5	1.3435	6.718	66.206
6	1.2496	6.248	72.454
7	0.9674	4.837	77.291
8	0.8809	4.404	81.695

(a) Produced Scree Plot (b) Goal Cumulative Variance Table

Figure 10. Produced Model PCA Results

PCA for this model yielded very similar results to the PCA for the goaling model with between 6-8 principal components extracted using the scree plot and cumulative variance chart. Figure 10 shows the scree plot and cumulative variance chart for the produced PCA.

Using these first eight principal components identified similar variables as candidate regressors as the PCA for the goal model. The exception is that the variable

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
Goaled	0.08895	0.14695	0.63137	-0.14506	-0.00217	-0.28500	-0.06780	0.00339
Previous Month Production	0.03583	0.08558	0.54459	-0.17037	-0.04937	-0.27351	-0.11091	0.21014
Num Recruiters	-0.06011	0.13797	0.52357	-0.11270	0.10705	-0.65697	0.16373	0.12394
QMA	-0.90656	0.15525	-0.02217	0.04727	-0.12507	-0.00644	0.16192	0.18998
Unemployment Rate	0.12345	0.35483	-0.29137	-0.43212	0.59336	0.03101	0.09151	0.03249
Labor Force Participation Rate	-0.27869	-0.24669	0.47573	0.42028	0.33988	0.28327	0.36950	0.06782
Total Pop	-0.87055	0.19541	-0.12929	-0.17680	-0.23286	-0.05712	0.06998	0.14558
15-19 Proportion	0.26254	0.52608	-0.08474	0.57436	0.23572	-0.21967	-0.10835	0.19914
20-24 Proportion	0.21539	0.80781	0.06185	0.34523	-0.27657	0.11896	0.05598	-0.09526
15-24 Proportion	0.26743	0.81678	0.01248	0.48977	-0.11309	0.00149	-0.00219	0.00870
Proportion Vet	0.45650	-0.14769	0.53083	0.11359	0.35546	0.26014	0.39700	-0.09913
Zone Area	0.74948	-0.10372	-0.10191	-0.11193	-0.29090	-0.05131	0.30087	0.43139
Obesity Raw	0.80380	0.17746	-0.01328	-0.17051	0.17265	0.04154	-0.29470	-0.10627
HS Raw	0.15066	-0.57016	0.10326	0.26020	-0.14429	-0.38591	0.10924	-0.20900
SC Raw	-0.66235	-0.22613	0.33616	0.31557	-0.15692	0.26097	-0.17412	0.10602
VC NUM	-0.54252	0.48382	-0.06741	-0.24635	0.34031	0.09500	0.07782	0.27589
Voting Participation	0.24443	-0.38010	0.37457	0.15793	0.15252	0.21249	-0.51868	0.44367
Sponsor	0.17533	0.36113	0.59897	-0.26455	-0.28047	0.23966	0.05817	-0.19634
JROTC Det	-0.00667	0.28862	0.51409	-0.45402	-0.20537	0.29093	-0.07019	-0.02916
Zone/Rec	0.71801	-0.14963	-0.25154	-0.10373	-0.31807	0.17228	0.26868	0.36755

Figure 11. Produced PCA Loading Matrix

for the monthly number of recruits goaled, y_1, was heavily loaded on the third principal component indicating that it would be included as a candidate regressor in the model. Additionally, the variable for voting participation which was included in the goal modeling process and identified as a significant regressor in the goaled model was not selected as a candidate regressor for the produced model. This iteration of PCA also identified 11 candidate regressors to include in the modeling process, shown in Figure 11.

Production Model Formulation.

The 11 candidate regressors identified in PCA and the 59 indicator variables representing each individual flight were included as the regressors in this model formulation. Mixed stepwise regression followed by backwards elimination at an $\alpha = 0.05$ significance level was again used to create a model of recruit production. This process identified three independent variables and a variety of the indicator variables signficant in the model. The three zone characteristic variables that were identified in this model are the monthly number of recruits goaled, y_1, the unemployment rate, x_4, and the number of JROTC detachments in a zone, x_{18}. The coefficients, p-values, and VIFs for these three regressors are included in Table 9. The coefficients for

Table 9. Produced Model Parameters

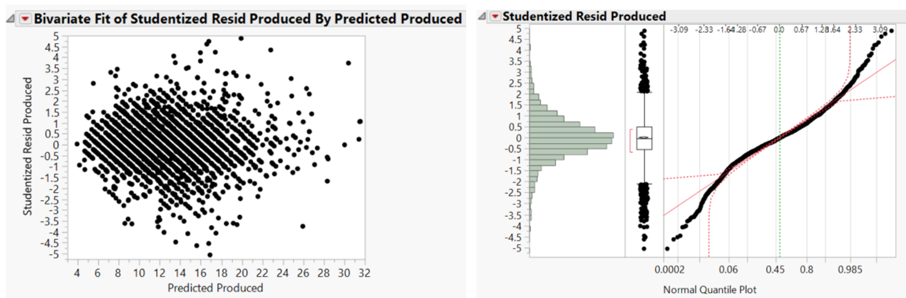
Variable	Coefficient	P-Value	VIF
y_1	0.929	< 0.0001	1.50
x_4	-6.04	0.0414	1.35
x_{18}	0.06	0.0003	3.18

each individual flight are not included in this table, but these indicator variables will affect the intercept of the regression line giving an individual production model for each flight. The full model results are included in Appendix C.

Table 9 identifies the relationship between the mean number of recruits produced in a month and the demographic or zone characteristics for each flight. This table identifies that the monthly number of recruits goaled is positively related to the number of recruits produced. The coefficient for this parameter is very close to one, which indicates that with all other variables held constant a one goal increase will increase recruit production by almost 1. This indicates the importance that the monthly goal has in recruit production. The number of JROTC detachments in an area is also positively related, although to a lesser degree. This variable is also positively related to the number of recruits goaled each month indicating that it is correlated with both aspects of recruiting. The unemployment rate is inversely related to recruit production indicating that as unemployment increases, recruit production decreases with all other factors held constant.

Model Adequacy.

The model for recruit production was checked against the linear regression assumptions in Table 6 in a similar manner to the model of recruit goaling. The first method is to examine a plot of the studentized residuals against the predicted values in Figure 12(a). This plot shows relatively constant variance with the residuals able to fit between two horizontal lines. The normal probability plot is Figure 12(b) indi-



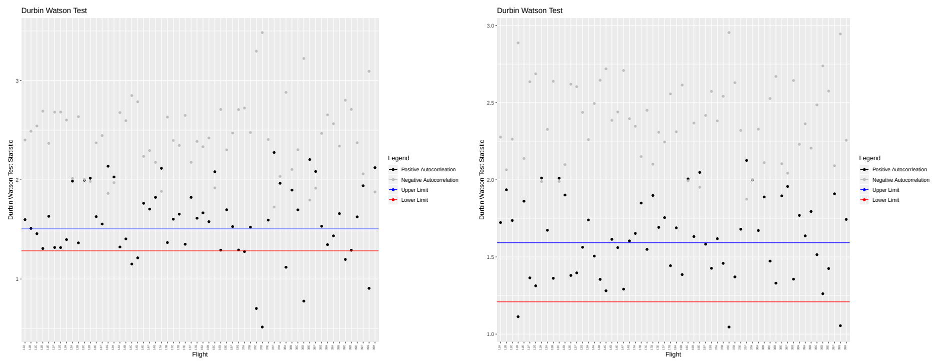
(a) Studentized Residuals vs. Predicted Value

(b) Normal Probabiliy Plot

Figure 12. Produced Model Residual Plots

cates that the residuals are symmetric and reasonably normally distributed. There is a slight tail at the lower and upper end, but the residuals follow a normal distribution throughout. These plots shows that this model satisfies assumptions 2,3, and 5 from Table 6.

Examining the residuals to determine how well the model satisfies assumption 4 of Table 6, the Durbin-Watson test is again used. The Durbin-Watson test statistics for positive and negative autocorrelation were calculated in the same manner as the test statistics for the goaled model producing the two charts in Figure 13. Similar to the figures for the Durbin-Watson Test in the goaled model, the figures show the test statistic, d, for positive and negative autocorrelation represented by the solid black dots and the grey dots, respectively. The red line on the chart indicates the lower limit and the blue line indicates the upper limit for the Durbin-Watson Test. This research fails to reject the null hypothesis of this test indicating that there is no autocorrelation if the values of d fall in between the two limits or above the blue line. Figure 13(a) shows a slight problem with autocorrelation. A variable representing



(a) Durbin-Watson Test

(b) Durbin-Watson Test with y_{t-1} and y_{t-2}

Figure 13. Produced Durbin-Watson Test

the lagged response variable, y_{t-1} for one lag was introduced into the model and the Durbin-Watson test did not change considerably, so a response variable with two lags, y_{t-2} was introduced as well. The introduction of these two lag variables reduced the autocorrelation significantly with only three flights now exhibiting autocorrelation, as shown in Figure 13(b). The coefficient for these lag variables was captured by interacting the indicator variable with each of the lag variables producing ϕ_1 and ϕ_2. These values are included for each flight in the full model results table in Appendix C. This model adequacy procedure determined that the model for recruit production satisfies assumptions 2-5 of linear regression.

The final model adequacy determination is to assess whether there are influence or leverage points within the data used to build the model. Using Cook's D value as a measure of influence, no points are greater than one indicating that there are no influence points in the data. Additionally, where Cook's D values are much larger than the others, the studentized residuals are not excessively large or small in comparison to the other points indicating that this point is not exhibiting influence. Examining the diagonals of the hat matrix to determine if there are any leverage points, the hat values are compared to the value of $2p/n$, which is about 0.08 for this model. There are some hat values that are larger than this test value, but the studentized residuals for these points do not indicate that they may be leverage points as they are in the range of the other residuals.

Model Fit and Validation.

To determine the performance of the recruit production model, the same summary statistics and validation procedure used in the goaling model were used. The summary statistics for this model, shown in Table 10, indicate that this model provides a good fit to the data used to build the model. The R^2 and R^2_{adj} are both pretty high

Table 10. Produced Model Summary Statistics

Statistic	Value
R^2	0.72
R^2_{adj}	0.71
R^2_{pred}	0.69
MSE	5.8

indicating that aboutn 72% of the variance in the data is explained in this model. The MSE for this model is 5.8 which shows the estimated variance of the residuals is about 6 recruits.

The second measure of model validation is to ensure that the independent variables are not correlated with each other. Using the VIFs from Table 9, there is no multicollinearity in the independent variables in the model. Additionally, the VIFs for the indicator variables and both lag variables did not exhibit any values greater than ten indicating that they are not correlated with each other. The final determination of model validation is to measure how well the model predicts new data that was not included in the model building process.

The MSE from Table 10 is used for model validation and compared to the average squared prediction error for the data that was withheld from the modeling procedure.

The data for RY2017 was withheld to test the performance of this model in predicting production on data that was outside of the model building data. The average squared prediction error for the data in RY2017 was 15.3. This is much larger than the MSE which indicates that this model may not predict new data very accurately. To properly validate the model, the R^2_{pred} is also compared to the R^2 of the original model. The difference between these two R^2 values is relatively small, so the loss in prediction ability between the data used to fit the model and the data withheld was deemed acceptable for this research.

Model Results.

This model of recruit production is used to determine the performance of each squadron within the 360th Recruiting Group while using both the models developed in this research. The predicted goals generated from the recruit goaling model were input into the recruit production model to predict how many recruits each unit would produce with the proposed goaling model. This method of comparing each squadron can allow the 360th Recruiting Group to determine if the squadron performance is consistent with the goaling model and to determine higher recruit producing squadrons.

In analyzing the recruiting production model, the predicted number of recruits produced while using the predicted goals is compared to the actual number of recruits produced to determine if the consistent pattern of recruit production is mirrored in this model. Figure 14 shows the predicted number of recruits produced using the predicted goals in blue and the actual number of recruits produced in black. This chart indicates that the predicted goals from the recruit goaling model produces a more consistent number of recruits throughout the years included in the study. This indicates that a goaling procedure based on demographics would reduce the variability

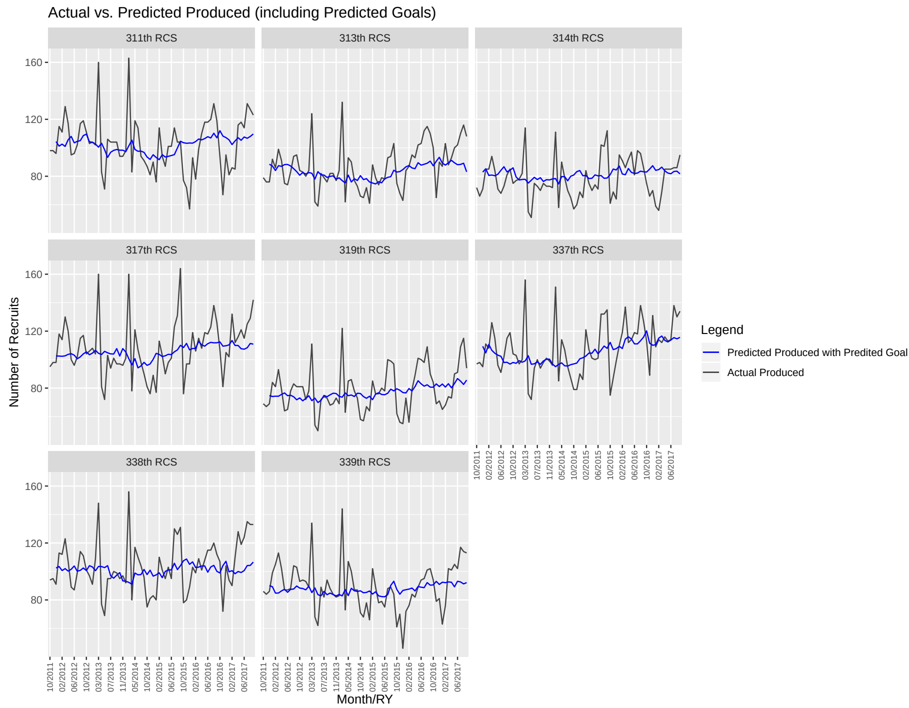


Figure 14. Predicted Recruit Production using Predicted Goals (RY2012-2017)

in the number of recruits produced by the 360th Recruiting Group from month to month.

Additionally, Figures 14 and 15 allow the 360th Recruiting Group to compare each squadron to determine how each squadron would perform with the predicted goals. Figure 14 indicates that the results from the recruit goaling model are consistent with recruiting production because the 311th RCS, 317th RCS, and 337th RCS are consistently higher producing squadrons than the other units. Figure 15 shows a comparison between each squadron with both recruit goaling and recruit production. The relationships between the squadrons continue with the 311th, 317th, and the 337th showing consistently higher goaling and production potential than the other squadrons. Using these models, the goals and production for each squadron can be projected into the future to determine how the relationship between squadrons appears in future months.

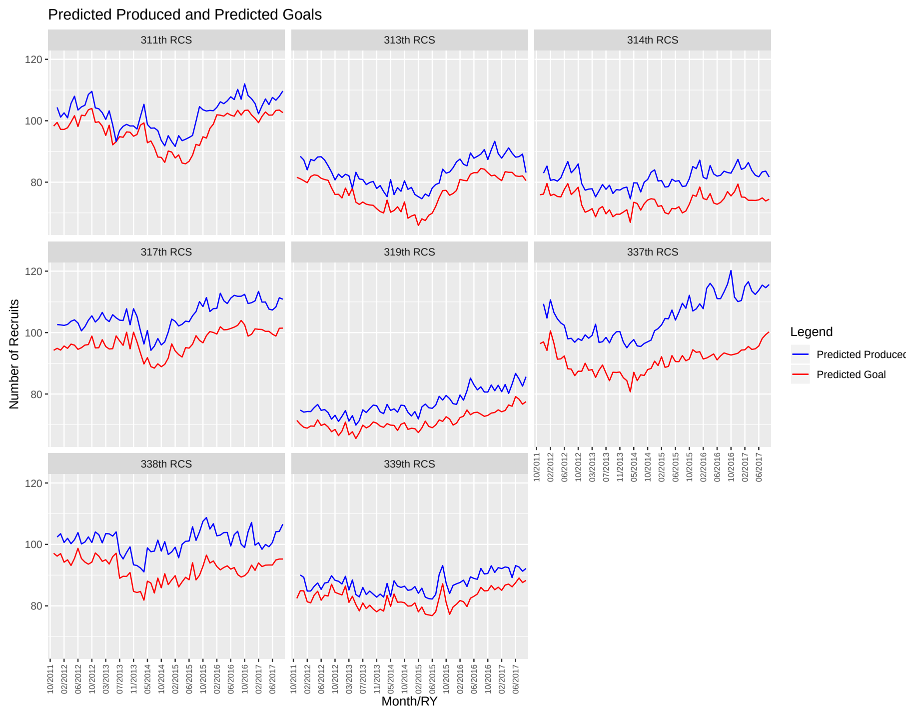


Figure 15. Predicted Goals and Predicted Production (RY2012-2017)

4.4 Recruit Goaling and Production Projections

For each model, the independent variables found significant in the model were forecasted out 48 months from the original data. This produced monthly values for RY2018-2021. Projecting out the monthly goals and production enables comparisons between the squadron to be drawn for future months and determines if the relationships that the goal and production models identified continues.

A simple moving average was used to forecast the values for each independent variable. A simple moving average is a forecasting technique that weights the most recent observations $\frac{1}{N}$, where N is the number of time periods in the span [30]. A simple moving average, M_t, of length N is given by equation 23 [30].

$$M_t = \frac{y_t + y_{t-1} + \ldots + y_{t-N+1}}{N} = \frac{1}{N} \sum_{t=T-N+1}^{T} y_t \tag{23}$$

The span N is determined by the data and which value of N will model the data in the best manner. For this research, each independent variable was plotted against time and simple moving averages with varying spans were plotted over the time series to determine the span. The spans for the independent variables ranged from two months to four months depending on which simple moving average fit the data better.

Forecasts for each flight were generated for each independent variable with a 95% confidence region using the R package "smooth" [44]. The MSEs for each forecast for each flight were very small and approaching zero indicating that these forecasts were good fits to the data. The forecasts and 95% confidence region for each flight are shown in Appendix B. The confidence region for each forecast increases as the time periods increase. The larger the confidence region in the forecasts indicates that there is more variability in the forecasts and that the true point estimate of the forecast could take on a larger range of values. The forecasts for the independent variables were then input into the recruit goaling and recruit production models.

Recruiting Goals Forecast.

The forecasts for variables in Table 7 were included into the model to develop point forecasts with a 95% confidence region for the number of recruits goaled each month. The 95% confidence region indicates that the actual value for the number of recruits goaled in a month will be within the upper and lower limits with 95% certainty. The point forecasts and 95% upper and lower limits are shown in Figure 16. The point forecast is shown in the red line surrounded by the confidence region.

Figure 16 is used to compare how each squadron performs in reference to the others. This chart shows the same pattern as Figure 9 with the 311th, 317th, and 337th RCS having a higher level of goaling than the other squadrons. The confidence regions can also be compared between all the squadrons indicating the variability of

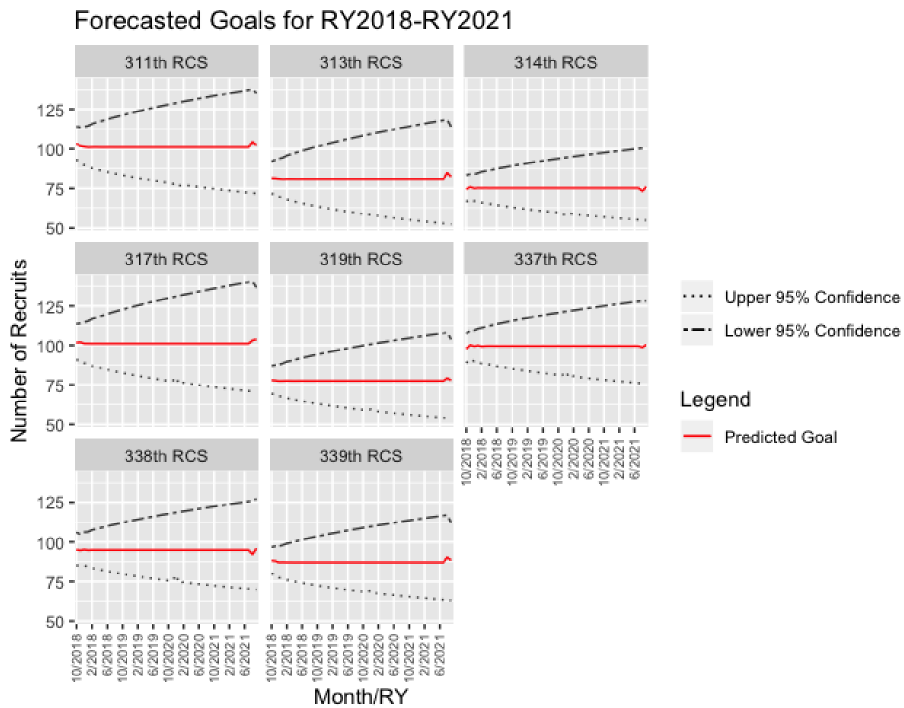


Figure 16. Forecasted Recruiting Goals

the forecasted level of goals. Among the three higher goaling potential squadrons, the 317th RCS has the largest interval at 95% confidence. This indicates that there is a larger range of values that the actual number of goals could take. The 337th has the smallest confidence interval indicating that the true values of monthly goals have less variability and are more consistent throughout the forecasted months. The forecasted goals shown in Figure 16 were used to create projections of recruiting production.

Recruiting Production Forecast.

The zone characteristic variables in Table 9 were projected out 48 months to identify point forecasts and a 95% confidence region for recruiting production. The results of these projections identify which squadrons have a higher potential for recruit production and the variability associated with each squadron. Figure 17 shows the relationships between each squadron's production and how the 95% confidence intervals compare.

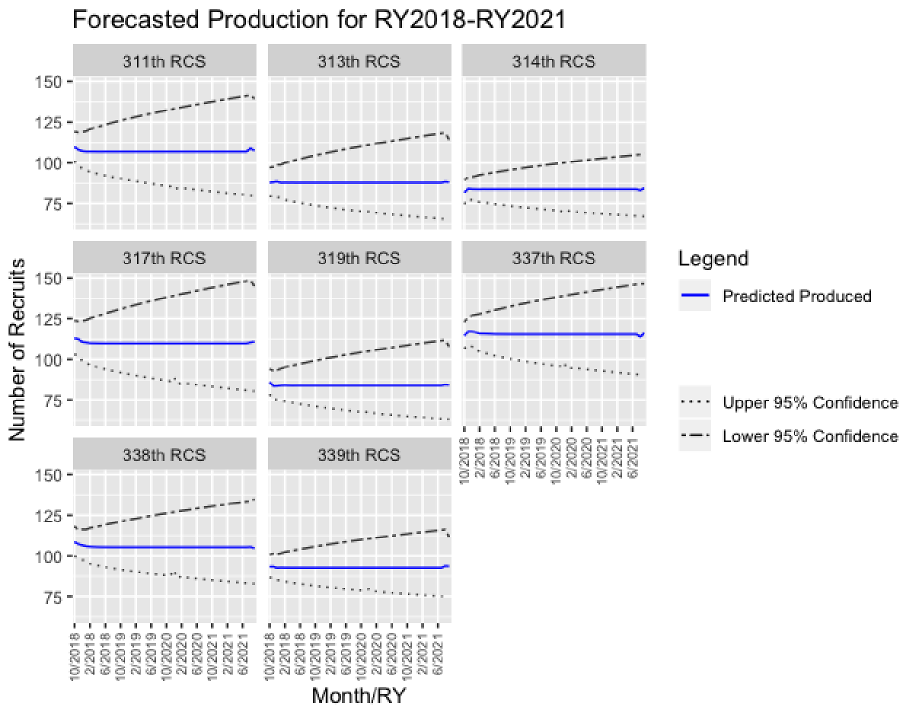


Figure 17. Forecasted Recruiting Production

Comparing each of the squadrons projected production identifies a similar relationship to that Figure 14 where the 311th, 317th, and 337th have a higher production potential in comparison to each of the other squadrons. This chart shows that based on the zone characteristics identified in the recruit production model, these squadrons would have a consistently higher production using the recruit goaling model proposed in this research. Additionally, Figure 17 shows which of the squadrons would have a lower production potential based on the characteristics of their zone. The 314th RCS and 319th RCS have lower projected production than the other squadrons.

In Figure 17, it is important to analyze the size of the 95% confidence region and compare these amongst the squadrons as well. Among the higher production potential squadrons identified in the figure, the 317th has the widest confidence region indicating that the true value of the recruit production could take on a wider range of values. The 337th RCS has a smaller confidence region than both the 317th and 311th indicating that the true value of their recruit production would be more consistent throughout the forecasted period. The recruiting projections indicate that a goaling procedure based on zone characteristics or demographics enables a more consistent recruit production and facilitates differentiation between squadrons.

V. Conclusion

Currently, the 360th Recruiting Group distributes monthly squadron goals utilizing a historic propensity measure and manning correction factor based on the number of recruiters available. They are interested in determining a more equitable approach to distribute recruiting goals to their flights using characteristics of each of the zones. This research utilized open source data comprised of 20 economic and demographic variables to describe each zone at the recruiting flight level.

The data describing each recruiting flight's zone was used to develop a multiple linear regression model to determine zone characteristics that correlate with recruiting goals and recruiting production. This research identified five independent variables that correlate with recruiting goals and three independent variables correlated with recruit production. Additionally, an individual regression model was developed for each individual flight using indicator variables to compare each flight's performance to a baseline flight. The two linear regression models developed enables identification of higher or lower goaling potential squadrons and flights and enables a more equitable placement of goals through the recruiting group.

Finally, this research utilized the recruit goaling and recruit production models to project recruitment goals and recruitment production for 48 months of data outside the original data. Projecting these models allows the recruiting group to compare how each squadron will perform against one another using these models and to examine the variability among each of the squadrons projected goaling and production potential.

Comparing the model results in this research to the results included in previous studies shows consistency among many of the results. Although much of the previous work identified in the literature review is focused on recruit production as opposed to recruit goals, comparing the variables identified and the associated relationships for both models gives an indication of the consistency of this work. The relationships identified between recruiting behavior and the obesity rate, high school graduation rate, number of recruiters, and the monthly number of goals are all consistent with previous work in this area and identify similar correlations with recruiting.

The unemployment rate was identified in various studies examining military recruiting, but the relationship differed between studies. This could be attributed to the studies examining different branches of the military. There were some studies indicating a positive relationship between military recruiting and the unemployment rate, while others indicate a negative relationship as shown in this study. This indicates that the negative relationship in this study is logical as previous research has also identified a negative relationship. The relationship between recruiting and the voting participation rate has been included in studies, but a relationship was not identified. Although not directly included in previously discussed recruiting studies, the relationship identified between JROTC presence and recruiting is consistent among the literature. Studies identified that recruiter interaction with high school students [9] had a positive relationship with military recruiting. Although JROTC is not explicitly a recruiting interaction, it is a potential interaction with youth and the military. This shows that relationship between JROTC presence in a zone and recruiting is positively correlated and is consistent with previous work. Comparing the relationships identified in this research with previous work provides an indication that the proposed goaling model in this research provides a consistent estimate of recruiting performance with previous work.

This research finds that the obesity rate and voting participation rate of a recruiting flight's zone is inversely correlated with recruit goaling; while a zone's high school graduation rate, the number of recruiters, and the number of JROTC detachments are positively correlated with recruit goaling. Additionally, this research found that the monthly number of goals and the number of JROTC detachments are positively correlated with recruiting production while the unemployment rate is inversely related to recruiting production. This research showed that using zone characteristics to levy monthly squadron goals allows a more consistent goaling and production pattern across the recruiting year. This research also identified higher goaling and production potential squadrons within the 360th Recruiting Group. The projections presented in this research also identify that the relationships among the squadrons persist in the projected data and identifies the squadrons with higher variability within these projections.

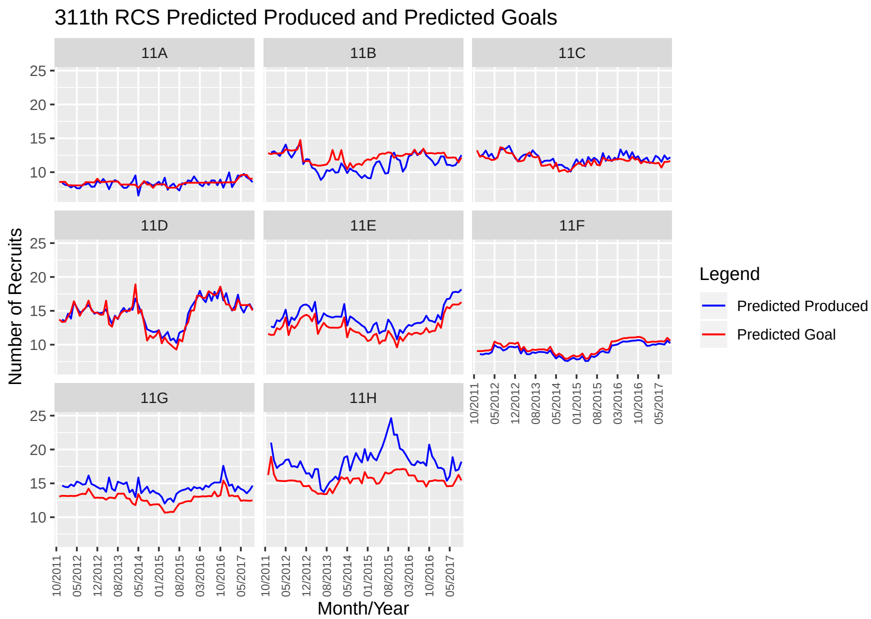
5.1 Future Research

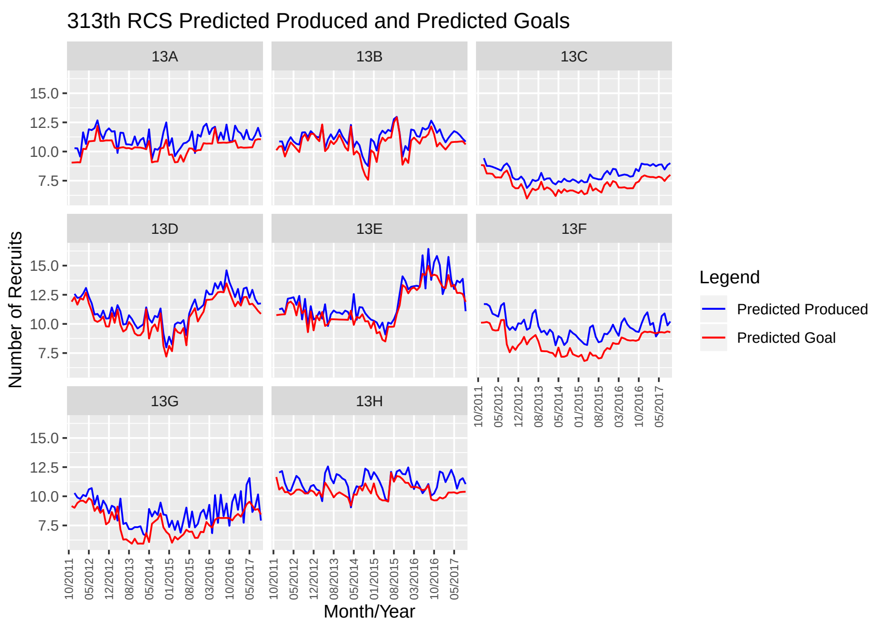
While military recruiting has been studied extensively, this research identified three areas of potential future research. The first area is the affect of high-school programs on military recruiting, specifically JROTC. This research identified the number of JROTC detachments significant in both recruit goaling and recruit production showing that it is positively correlated to both aspects of recruiting. Data availability only allowed for the number of detachments to be included in this study, but further JROTC data such as enrollment numbers for each detachment and historical figures for the number of detachments and enrollment would allow a more thorough examination of the affects of JROTC programs on recruiting. Additionally, other high-school programs such as Civil Air Patrol or other extracurricular programs could be examined to identify areas of potential recruiting interaction for high school students.

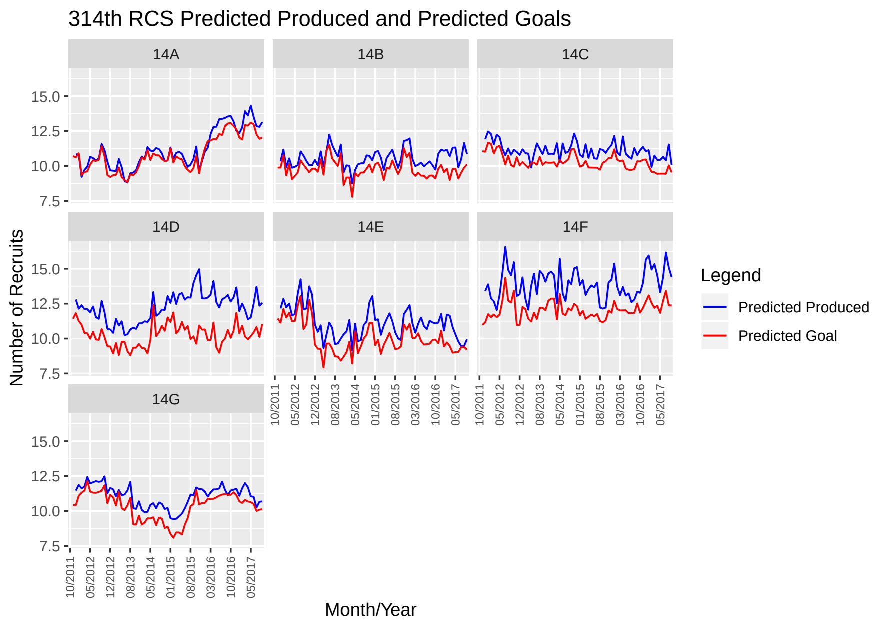
The second area of future research is to apply this methodology to the other recruiting groups within the AFRS. This would identify consistencies within the United States that contribute to Air Force recruiting or if each group has different characteristics that relate more heavily to each group. This would allow the AFRS to compare the recruiting potential of each group across the United States.

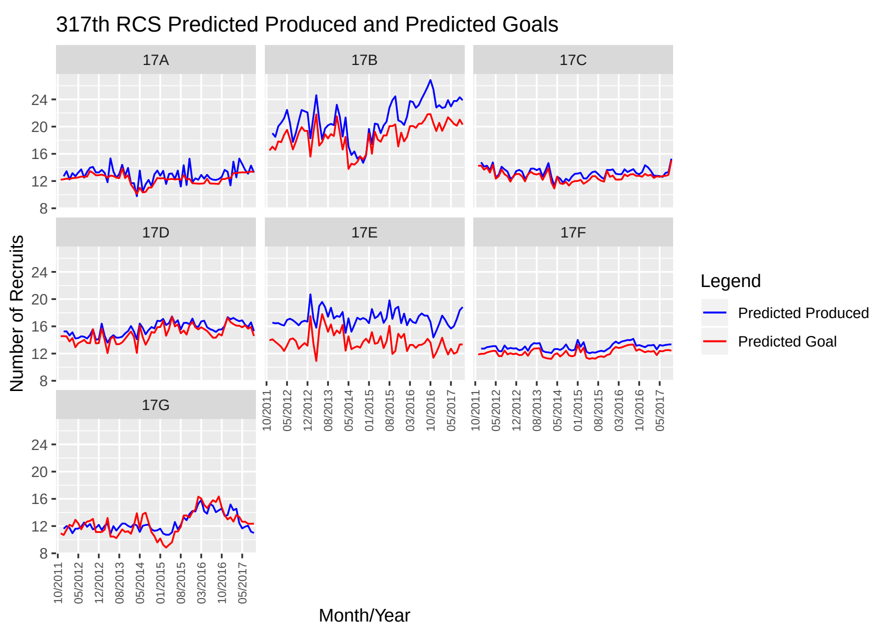
The final area of potential future research would be to apply this methodology across all recruiting flights around the United States. This would enable the AFRS to distribute goals based on zone characteristics and not on recruiting propensity as they have. Additionally, this would identify an equitable distribution of annual goals for each of the recruiting groups, squadrons, and flights.

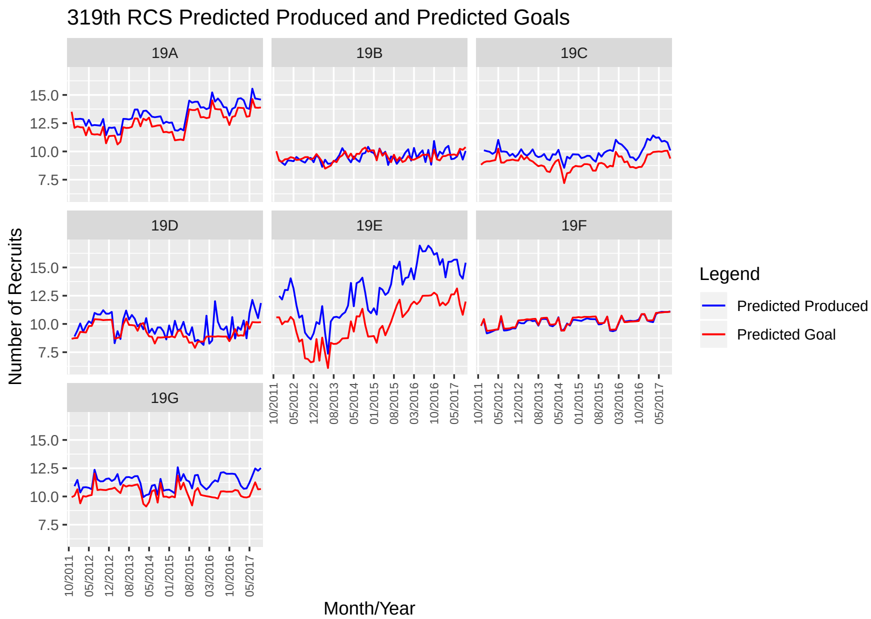
Appendix A. Recruit Production and Goals for Flights



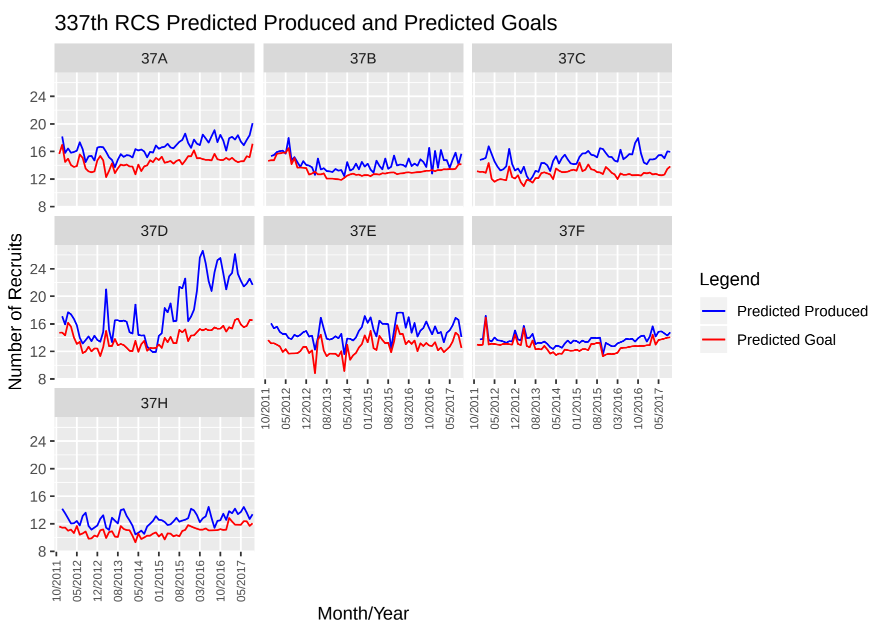



314th RCS Predicted Produced and Predicted Goals
14A
14B
14C
14D
14E
14F
14G
Number of Recruits
Month/Year
Legend
Predicted Produced
Predicted Goal


317th RCS Predicted Produced and Predicted Goals
17A
17B
17C
17D
17E
17F
17G
Number of Recruits
Month/Year
Legend
Predicted Produced
Predicted Goal



319th RCS Predicted Produced and Predicted Goals
19A
19B
19C
19D
19E
19F
19G
Number of Recruits
Month/Year
Legend
Predicted Produced
Predicted Goal



337th RCS Predicted Produced and Predicted Goals
37A
37B
37C
37D
37E
37F
37H
Number of Recruits
Month/Year
Legend
Predicted Produced
Predicted Goal

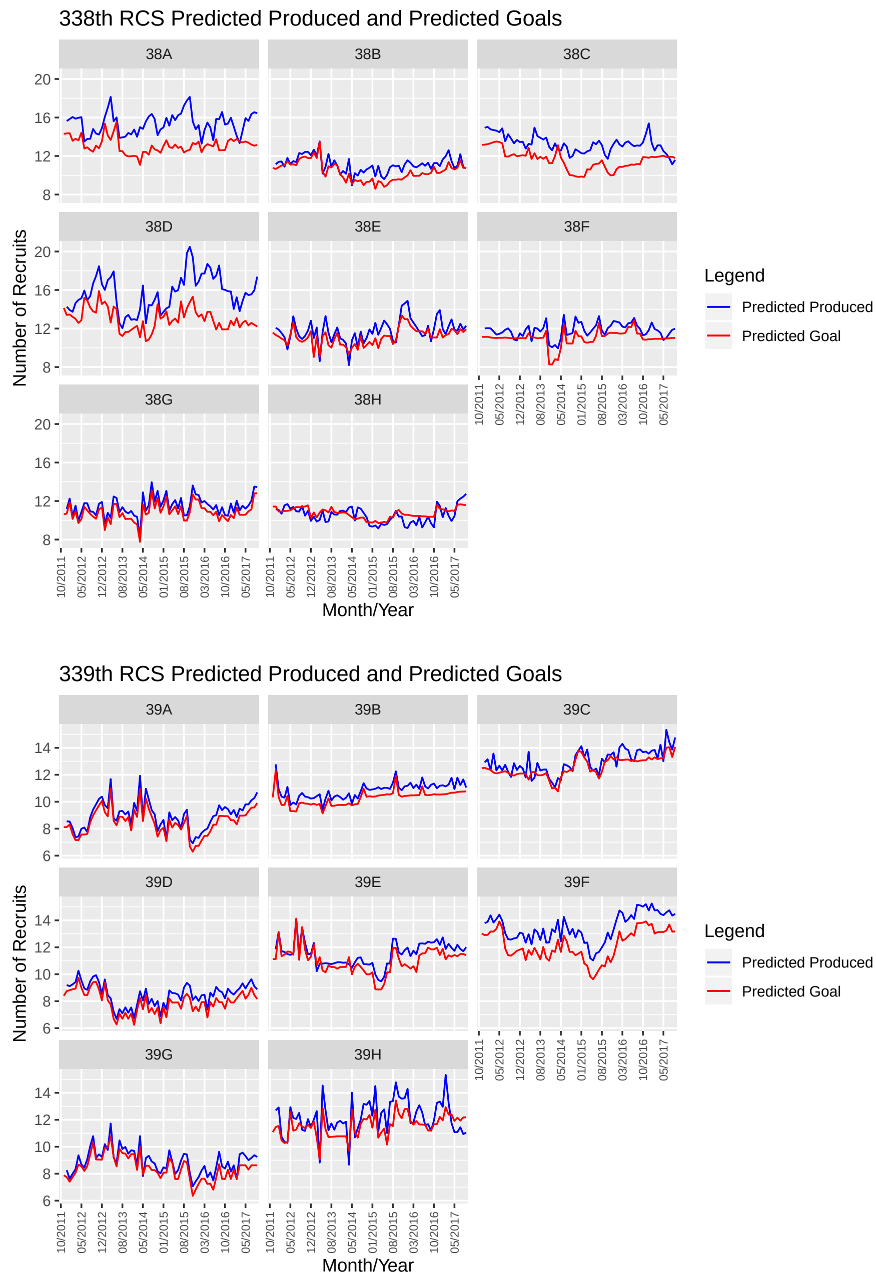
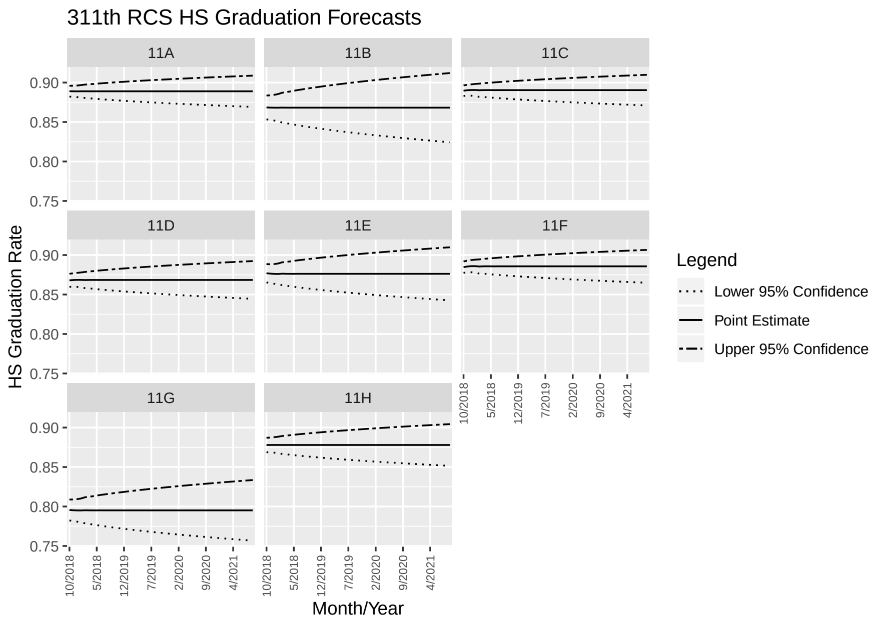
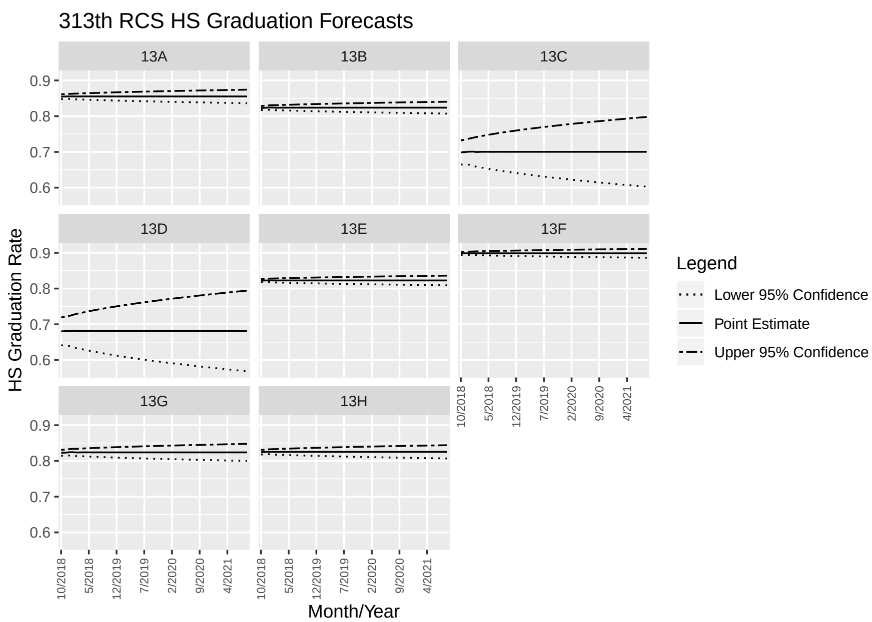
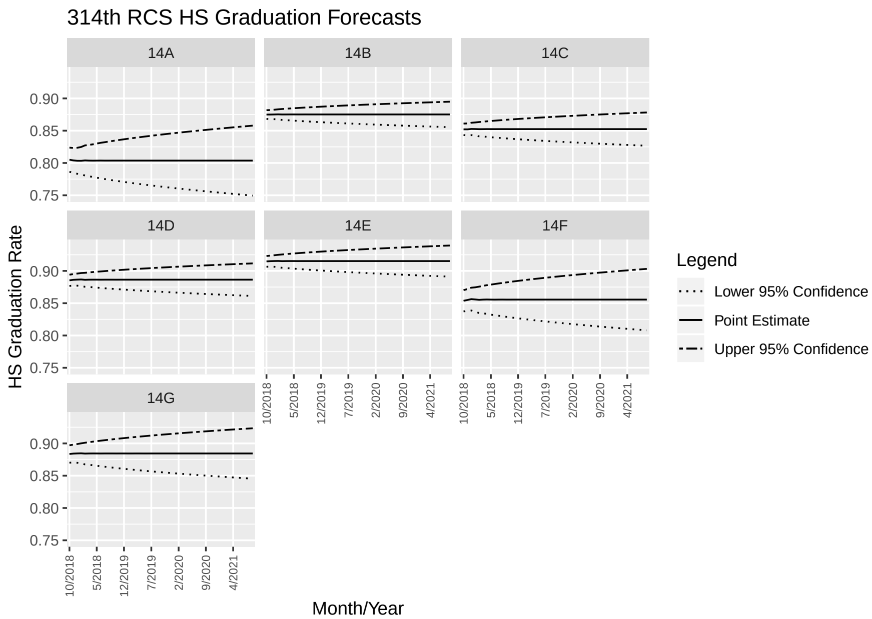


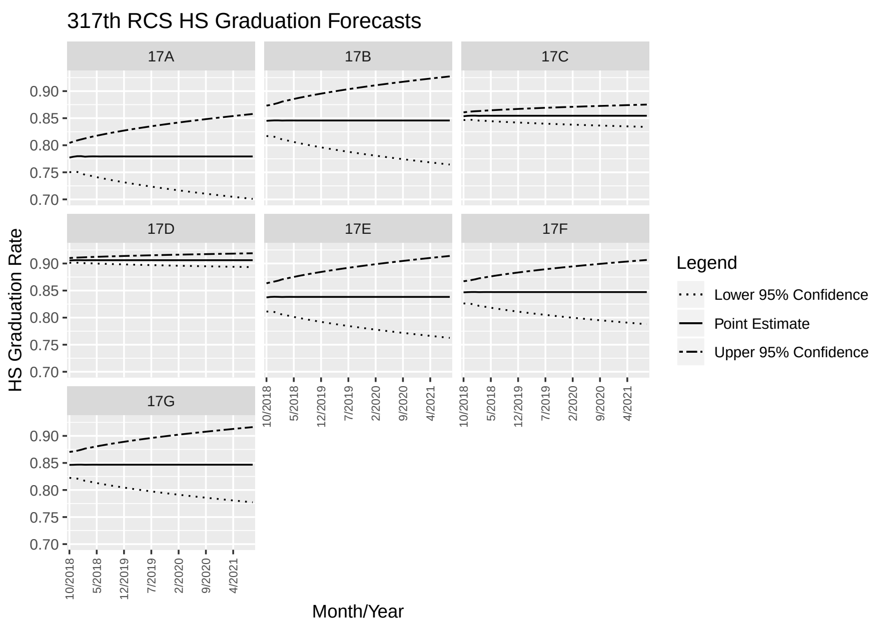
Figure 18. Predicted Produced and Predicted Goals RY2012-2017

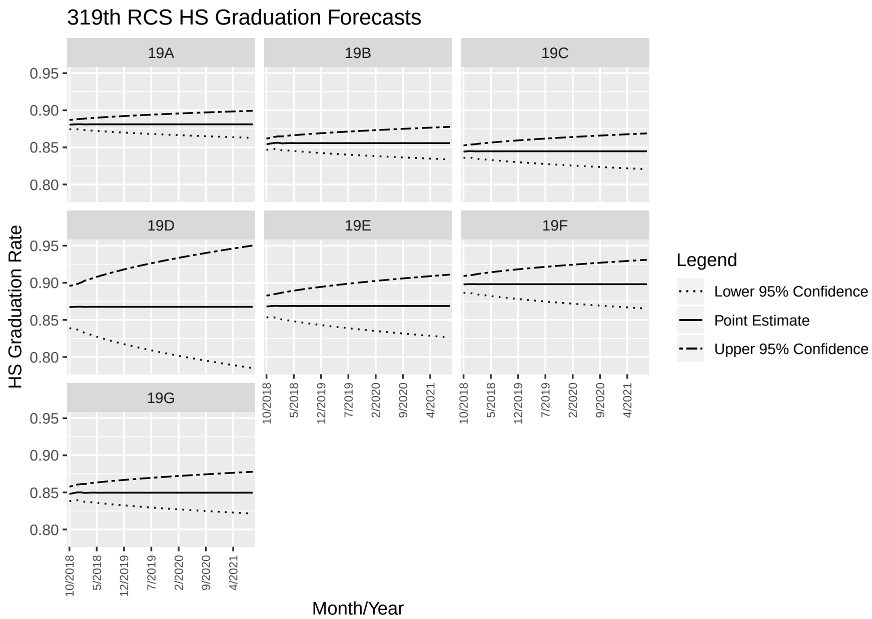
Appendix B. Independent Variables Forecast

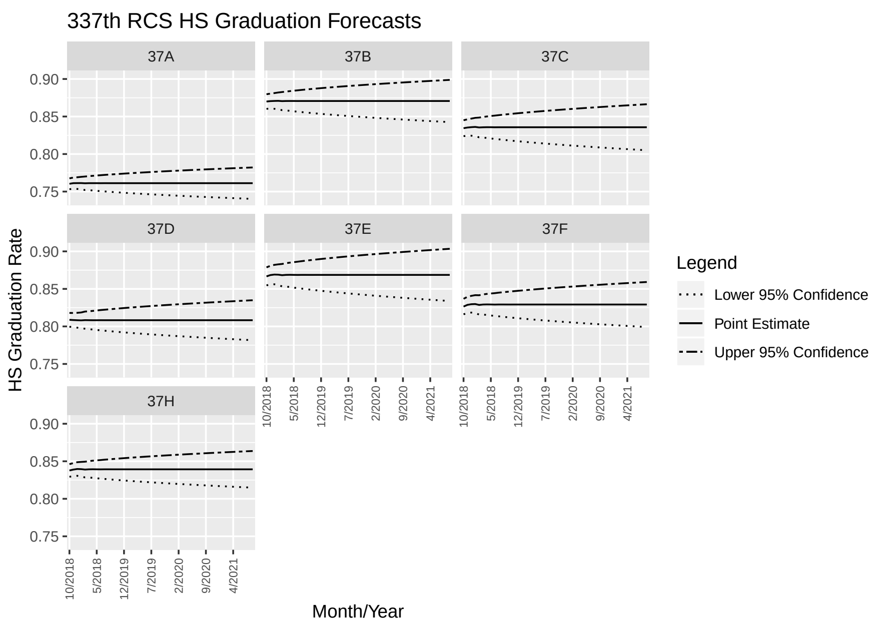





314th RCS HS Graduation Forecasts
14A
14B
14C
14D
14E
14F
14G
HS Graduation Rate
0.90
0.85
0.80
0.75
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


317th RCS HS Graduation Forecasts
17A
17B
17C
17D
17E
17F
17G
HS Graduation Rate
0.90
0.85
0.80
0.75
0.70
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


319th RCS HS Graduation Forecasts
19A
19B
19C
19D
19E
19F
19G
0.95
0.90
0.85
0.80
HS Graduation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


337th RCS HS Graduation Forecasts
37A
37B
37C
37D
37E
37F
37H
0.90
0.85
0.80
0.75
HS Graduation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

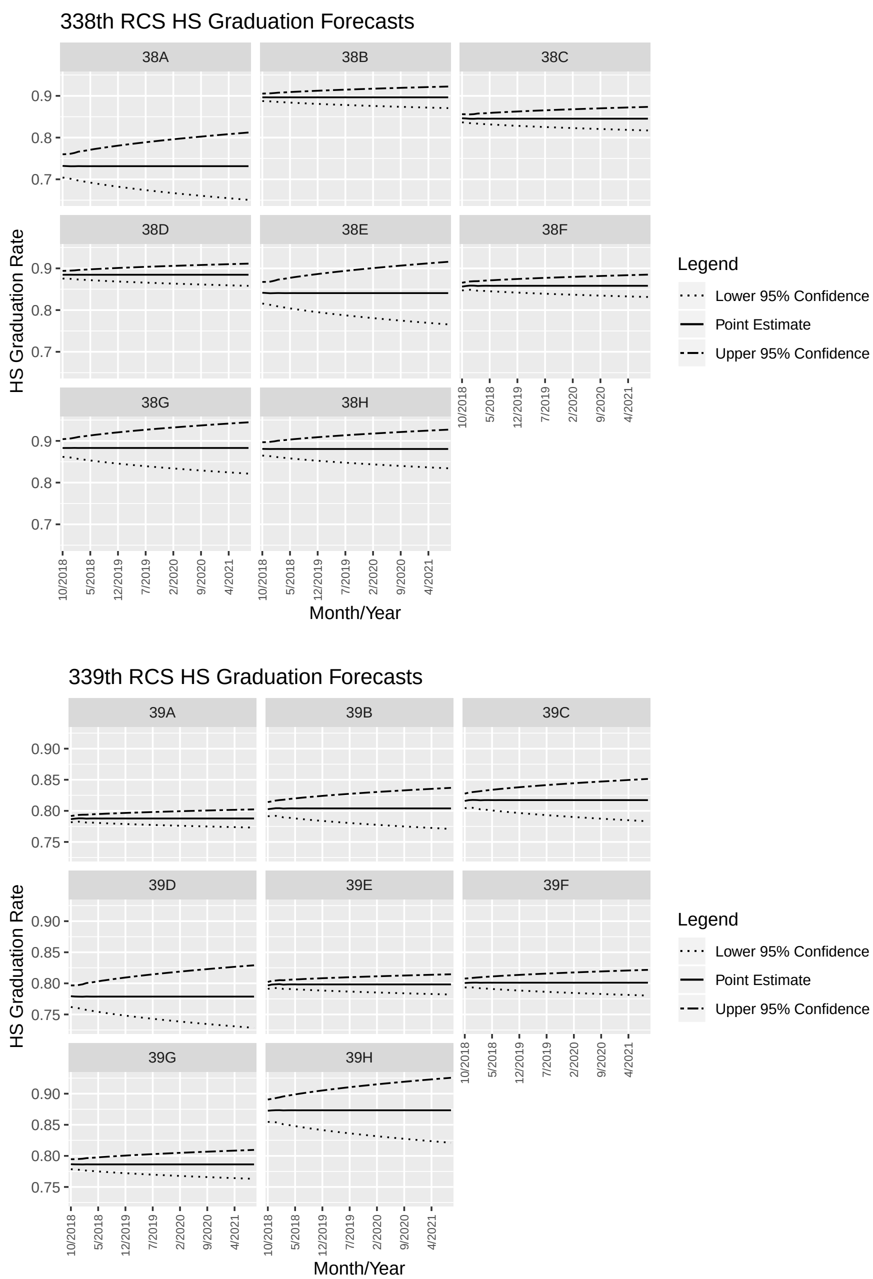
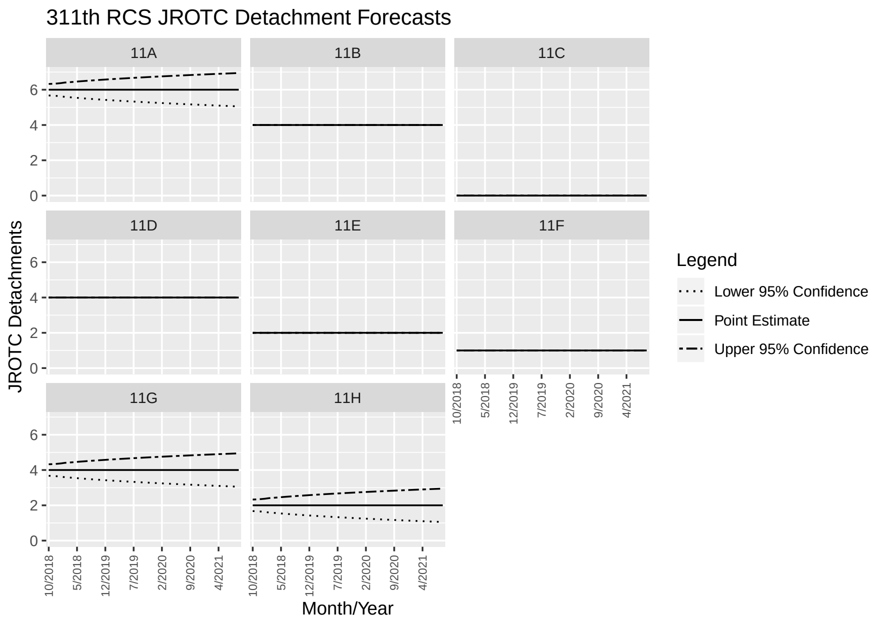
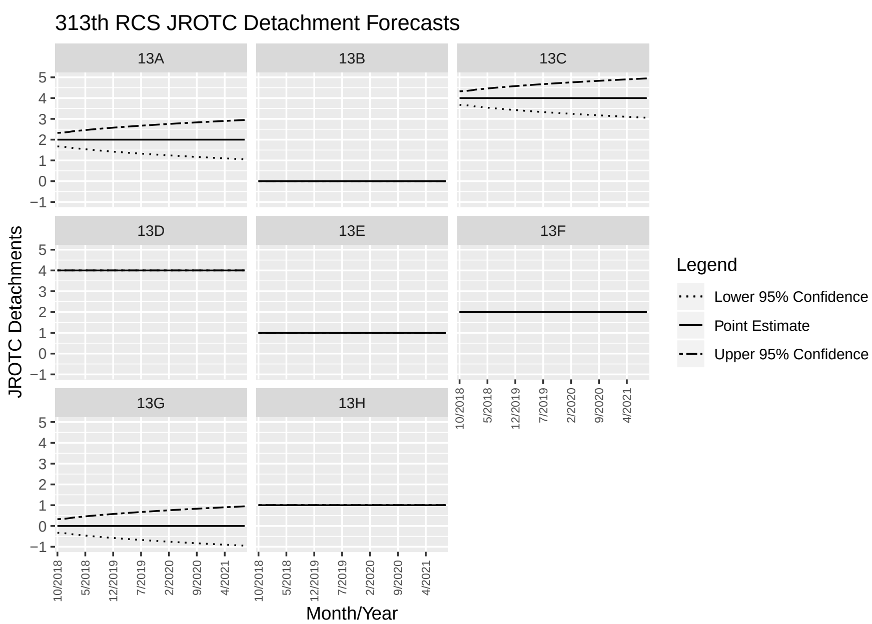
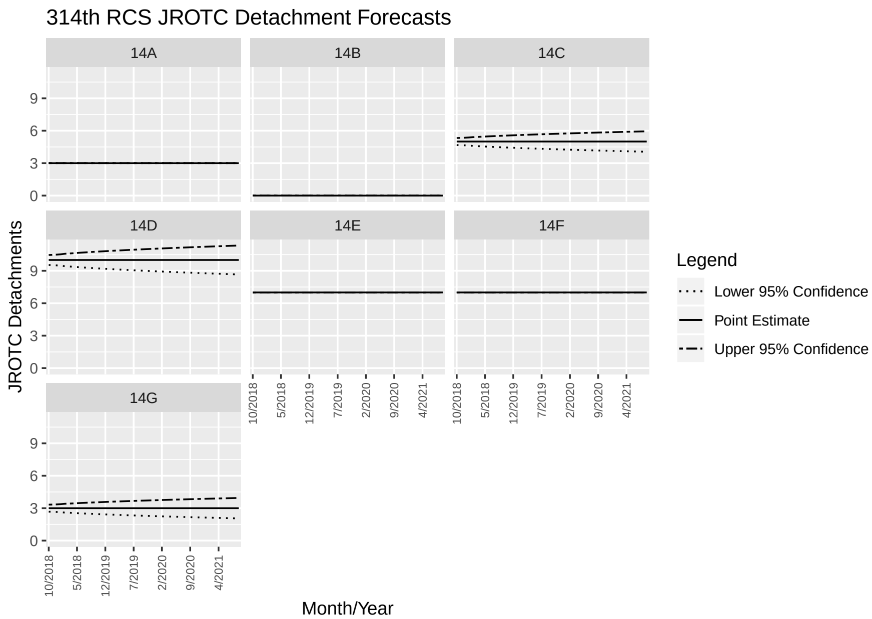
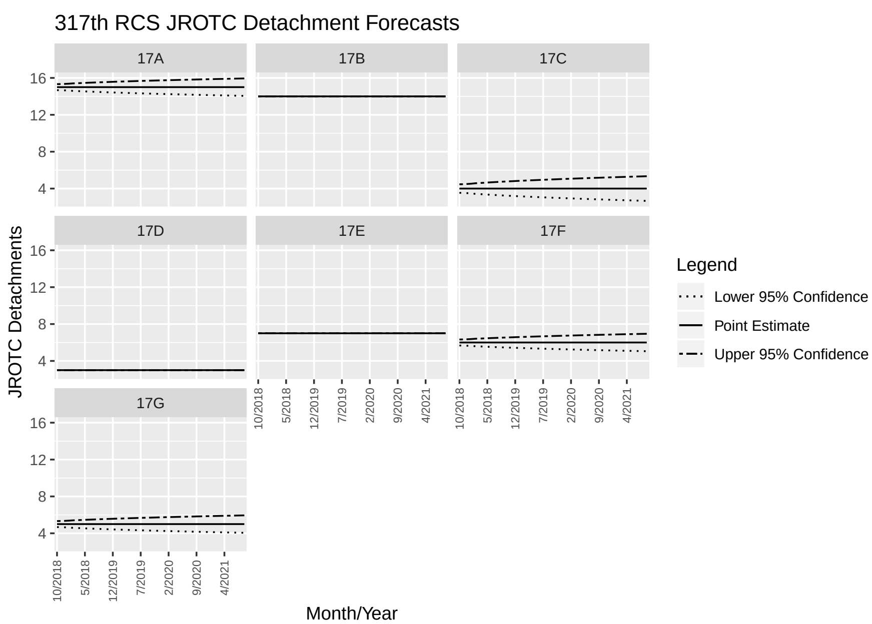


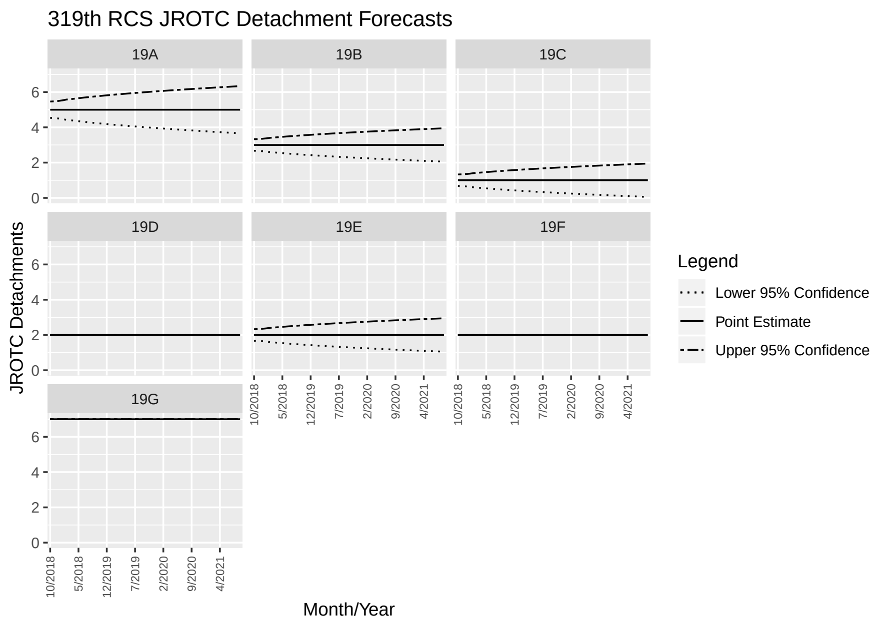
Figure 19. HS Graduation Rate Forecasts

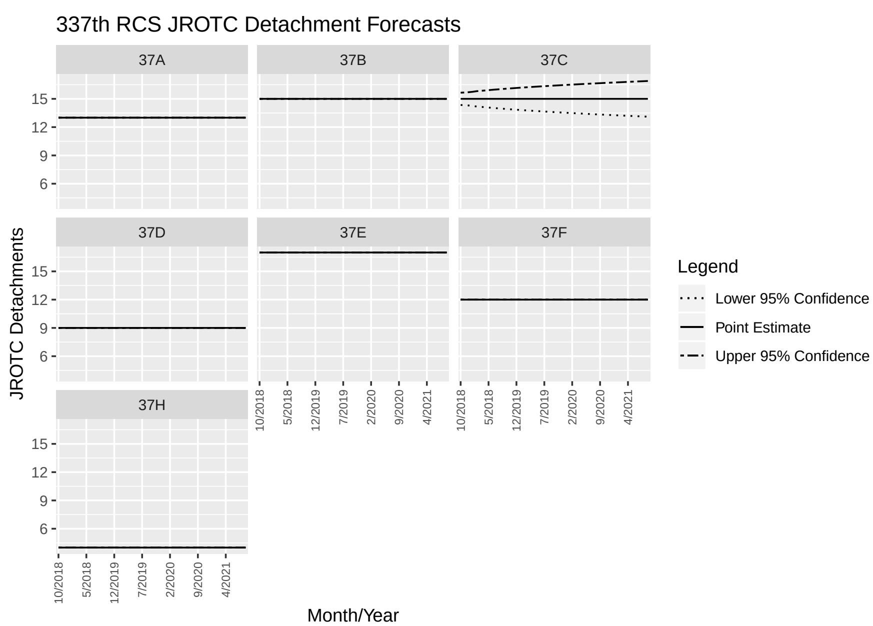

311th RCS JROTC Detachment Forecasts
11A
11B
11C
11D
11E
11F
11G
11H
JROTC Detachments
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

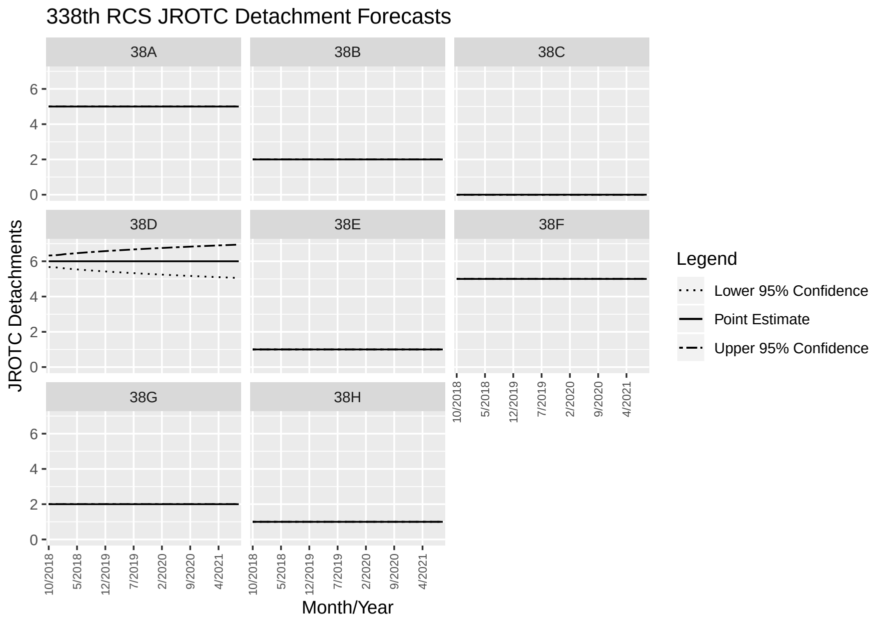

313th RCS JROTC Detachment Forecasts
13A
13B
13C
13D
13E
13F
13G
13H
JROTC Detachments
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


314th RCS JROTC Detachment Forecasts
14A
14B
14C
14D
14E
14F
14G
JROTC Detachments
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


317th RCS JROTC Detachment Forecasts
17A
17B
17C
17D
17E
17F
17G
JROTC Detachments
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


319th RCS JROTC Detachment Forecasts
19A
19B
19C
19D
19E
19F
19G
JROTC Detachments
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


337th RCS JROTC Detachment Forecasts
37A
37B
37C
37D
37E
37F
37H
JROTC Detachments
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence



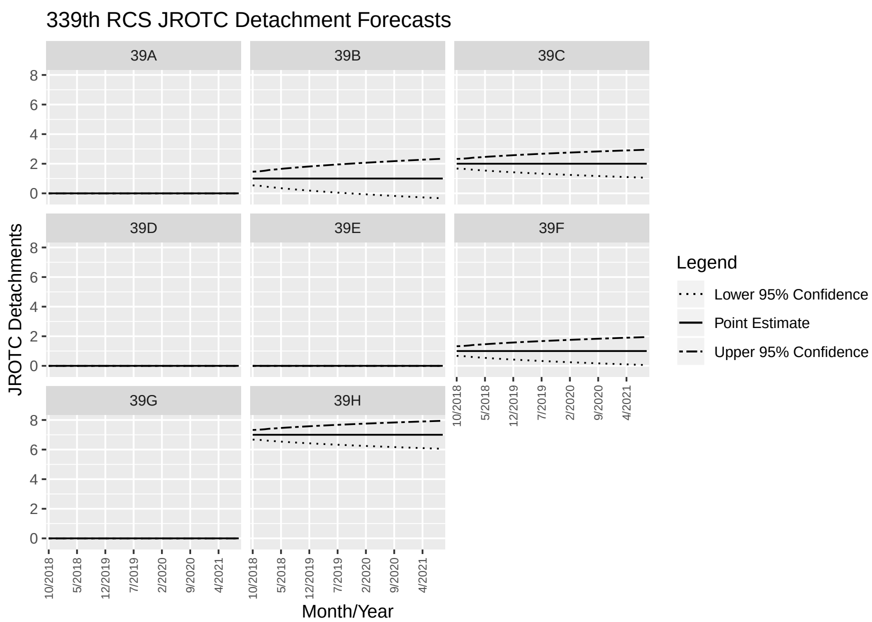
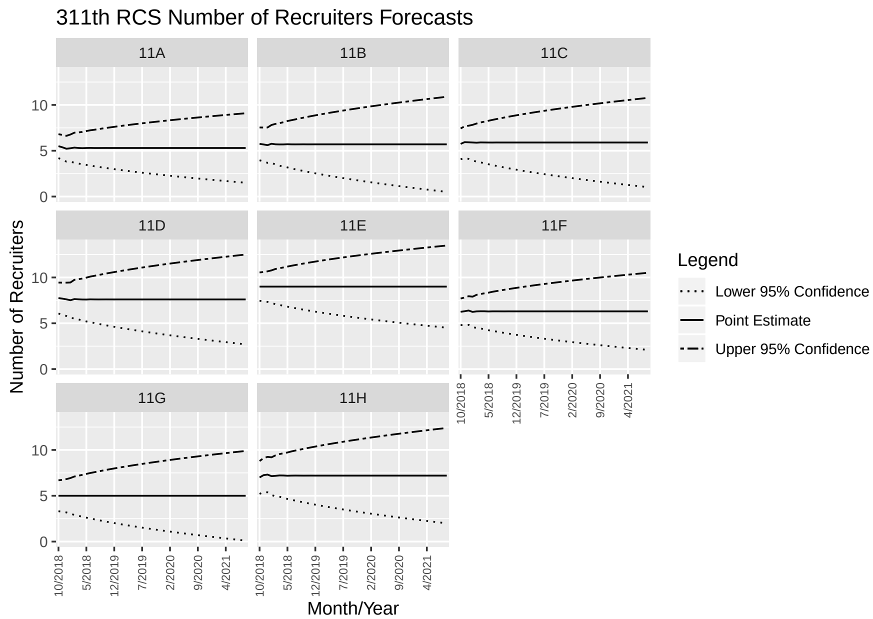
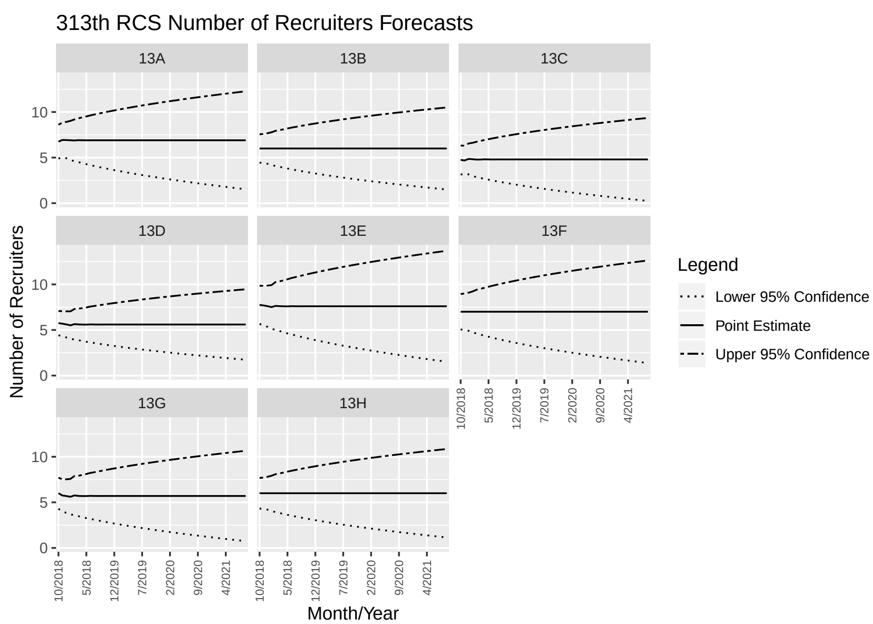
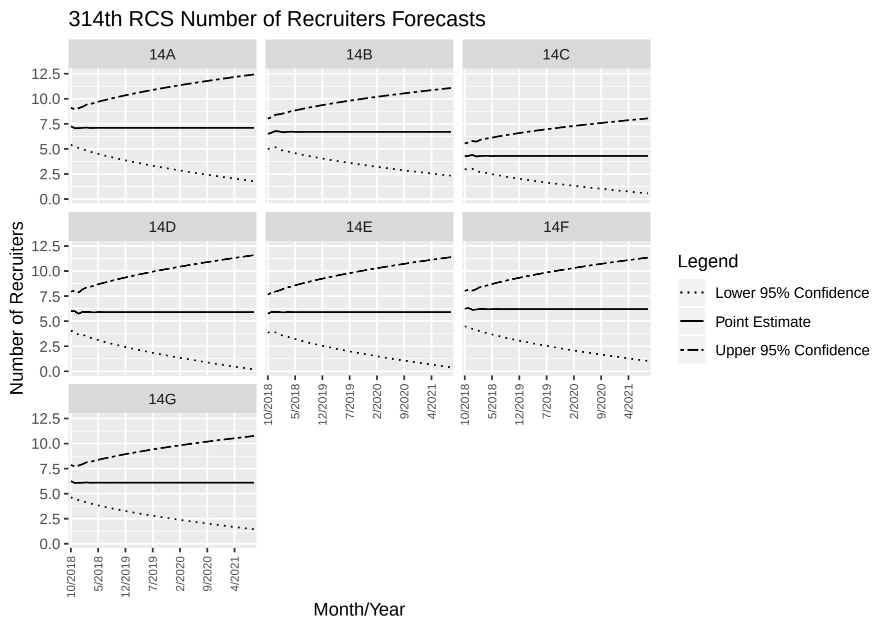
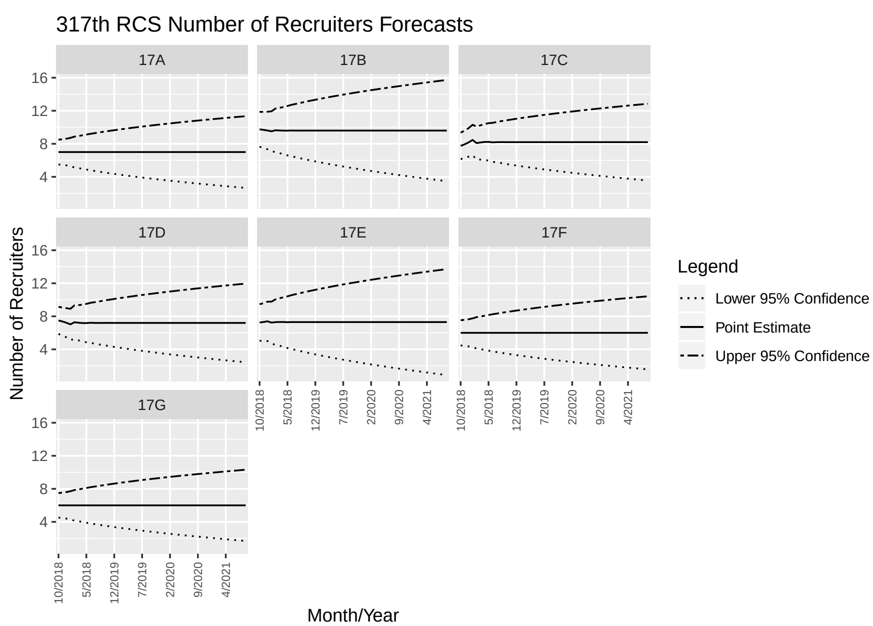


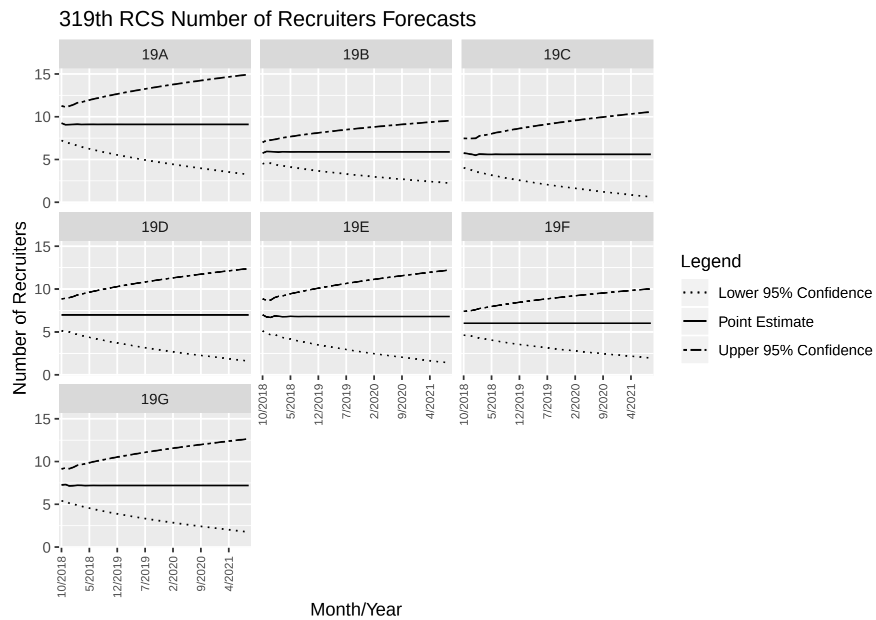
Figure 20. JROTC Detachment Forecasts

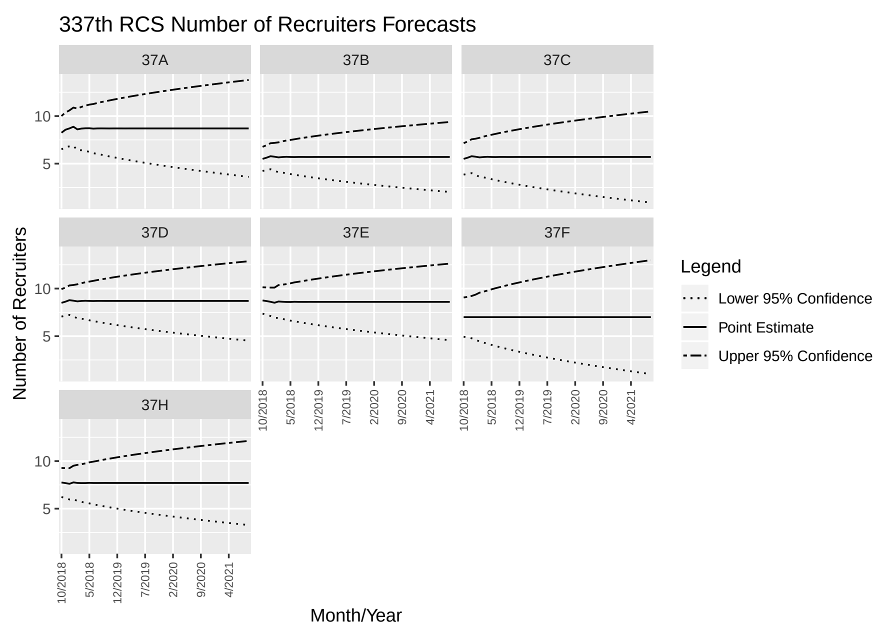

311th RCS Number of Recruiters Forecasts
11A
11B
11C
11D
11E
11F
11G
11H
Number of Recruiters
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

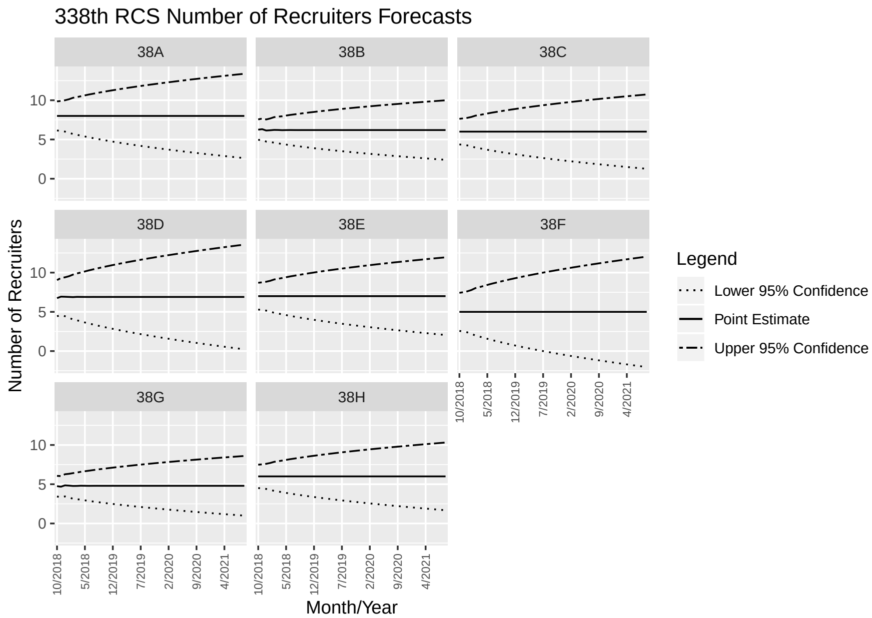

313th RCS Number of Recruiters Forecasts
13A
13B
13C
13D
13E
13F
13G
13H
Number of Recruiters
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


314th RCS Number of Recruiters Forecasts
14A
14B
14C
14D
14E
14F
14G
Number of Recruiters
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


317th RCS Number of Recruiters Forecasts
17A
17B
17C
17D
17E
17F
17G
Number of Recruiters
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


319th RCS Number of Recruiters Forecasts
19A
19B
19C
19D
19E
19F
19G
Number of Recruiters
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021


337th RCS Number of Recruiters Forecasts
37A
37B
37C
37D
37E
37F
37H
Number of Recruiters
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021



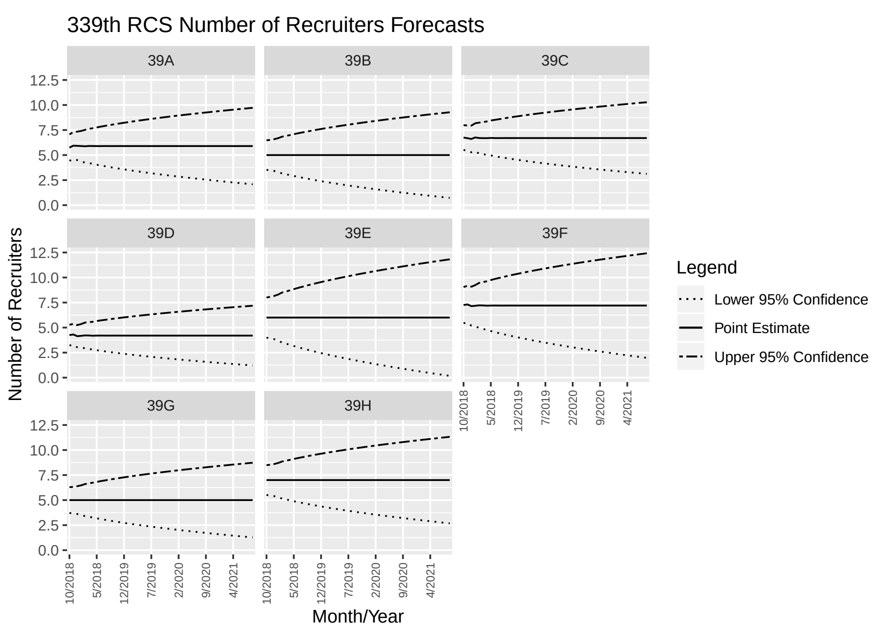
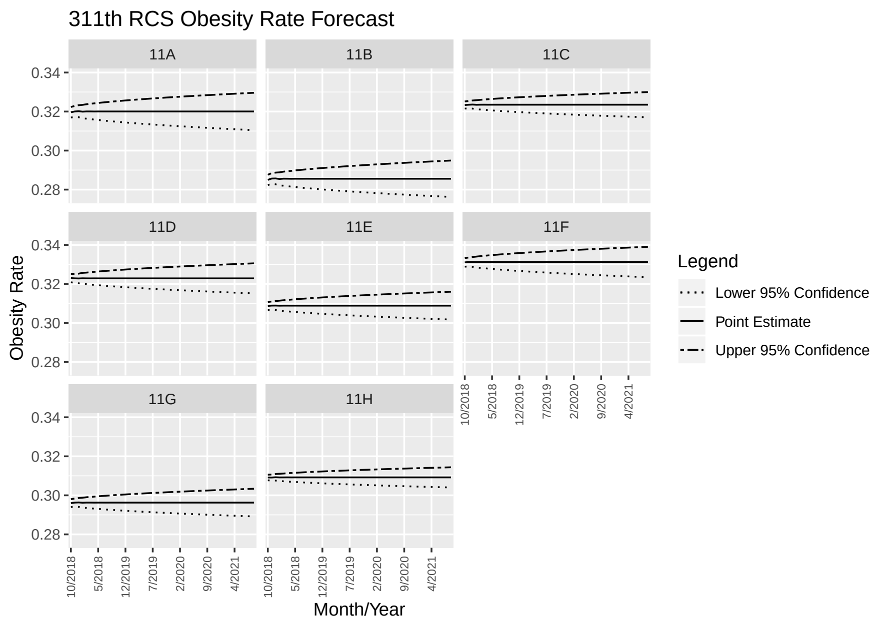
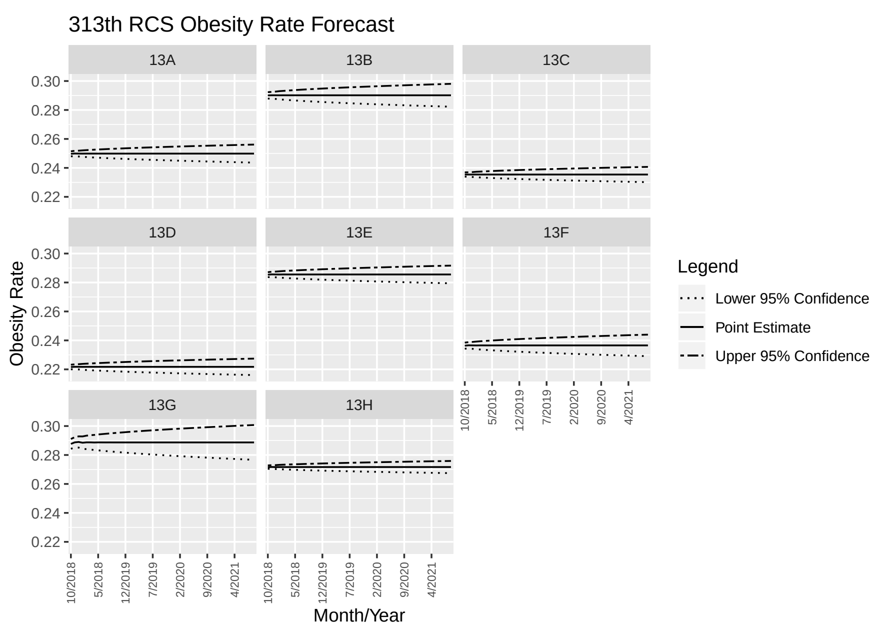
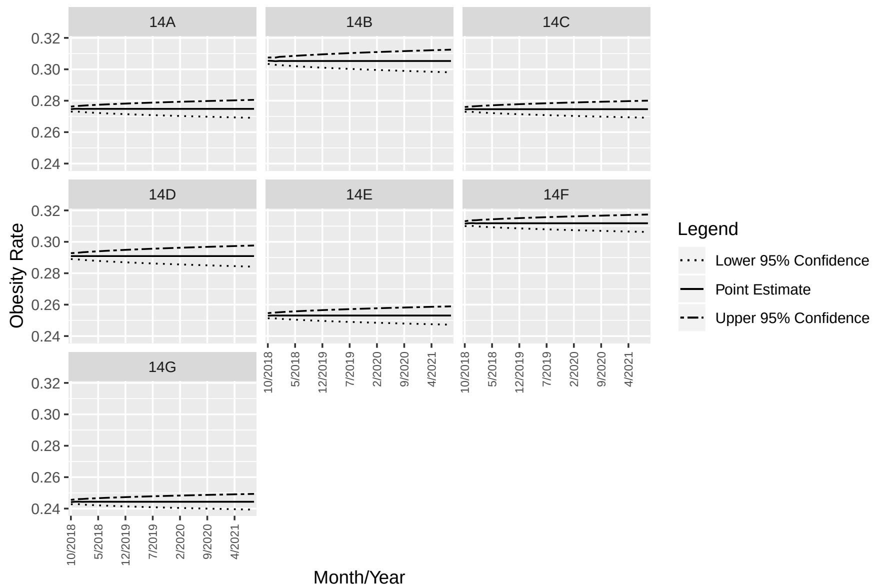


Figure 21. Number of Recruiters Forecasts

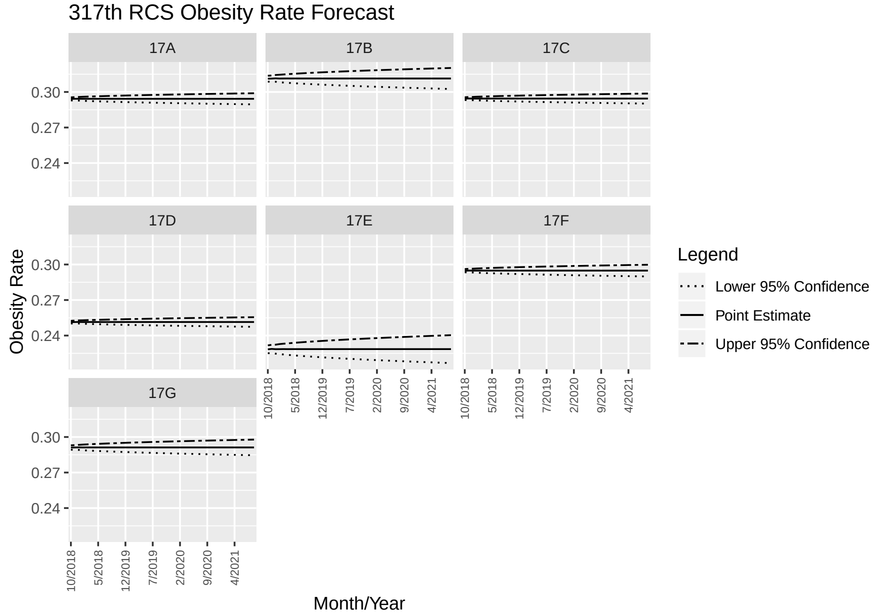

311th RCS Obesity Rate Forecast
11A
11B
11C
11D
11E
11F
11G
11H
Obesity Rate
0.34
0.32
0.30
0.28
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

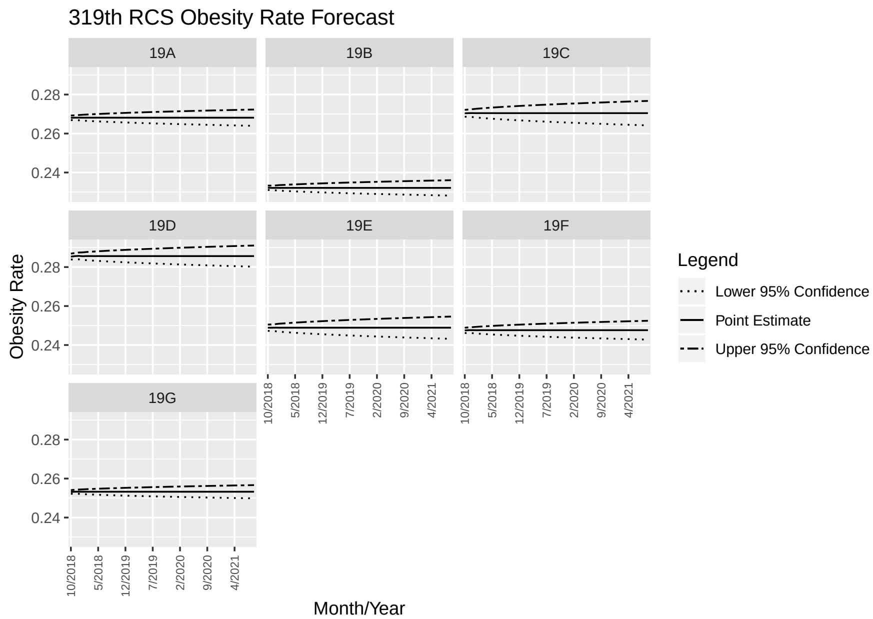

313th RCS Obesity Rate Forecast
13A
13B
13C
13D
13E
13F
13G
13H
Obesity Rate
0.30
0.28
0.26
0.24
0.22
Month/Year
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

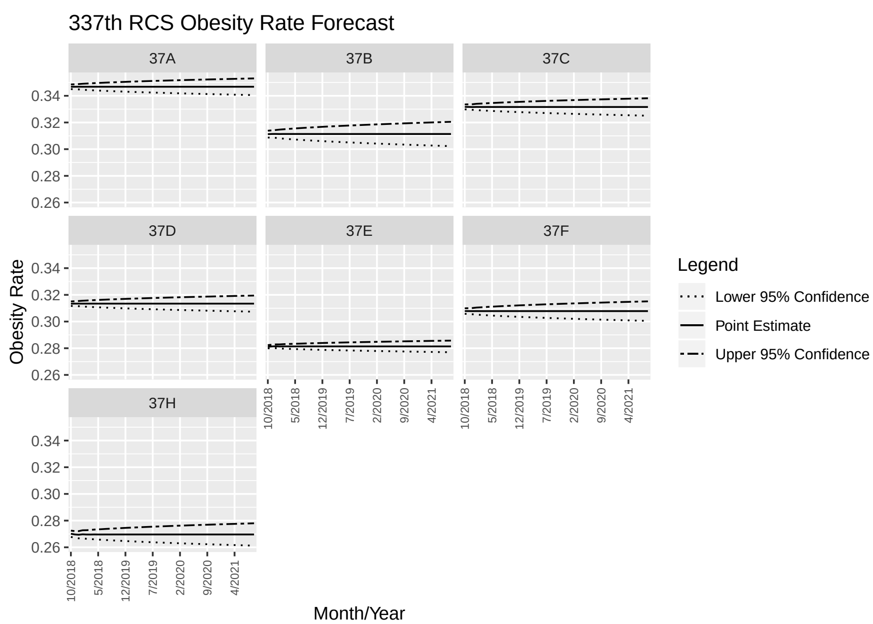
314th RCS Obesity Rate Forecast

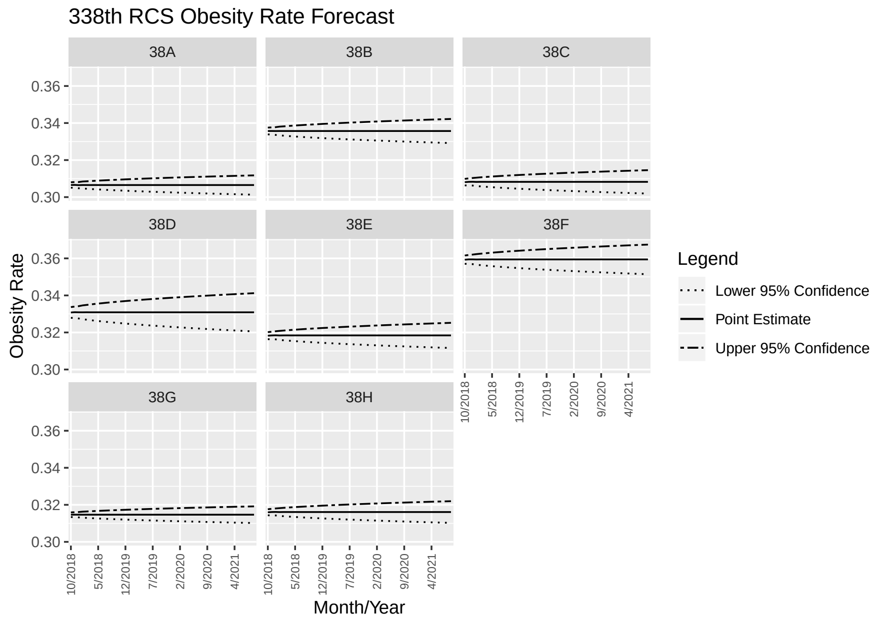


317th RCS Obesity Rate Forecast




319th RCS Obesity Rate Forecast
19A
19B
19C
19D
19E
19F
19G
0.28
0.26
0.24
Obesity Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


337th RCS Obesity Rate Forecast
37A
37B
37C
37D
37E
37F
37H
0.34
0.32
0.30
0.28
0.26
Obesity Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence



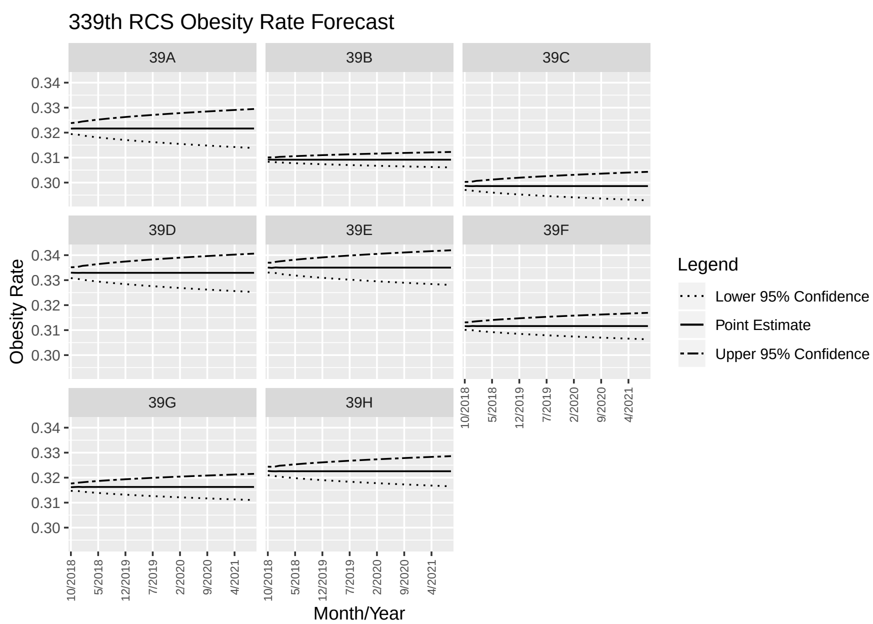
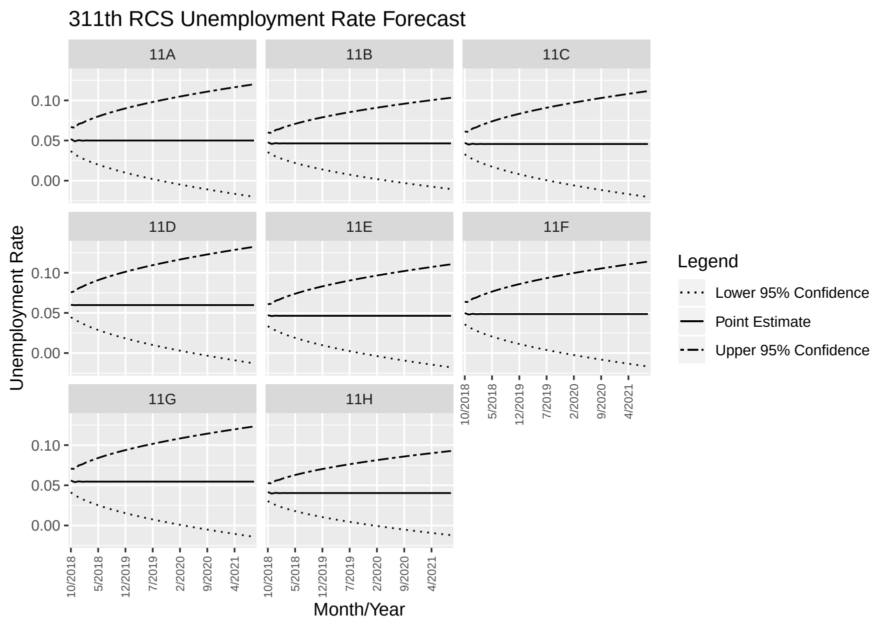
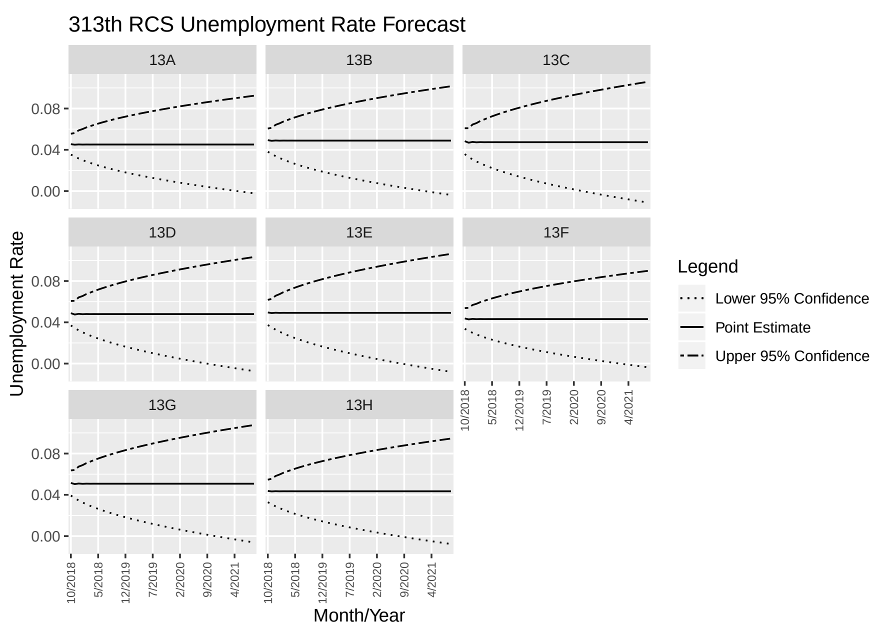
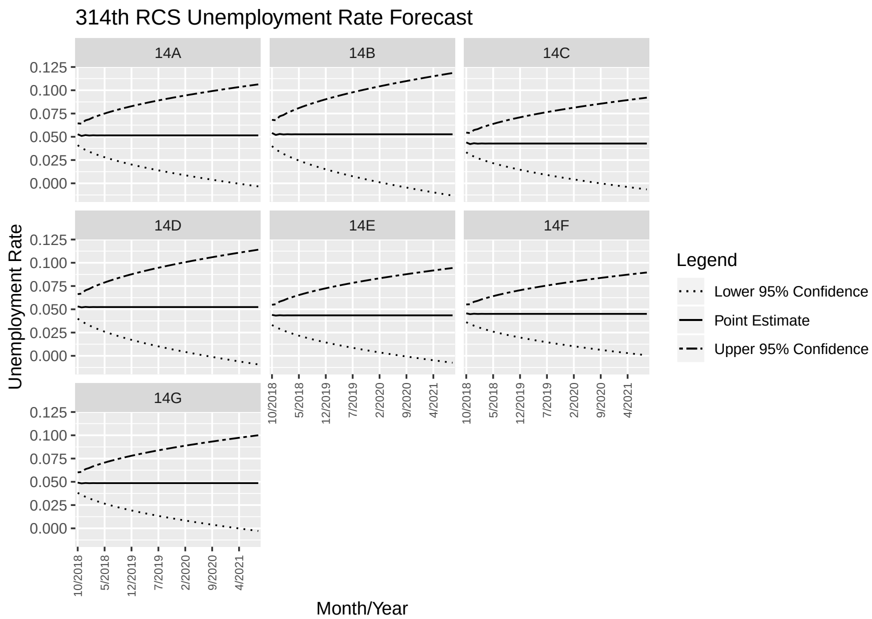
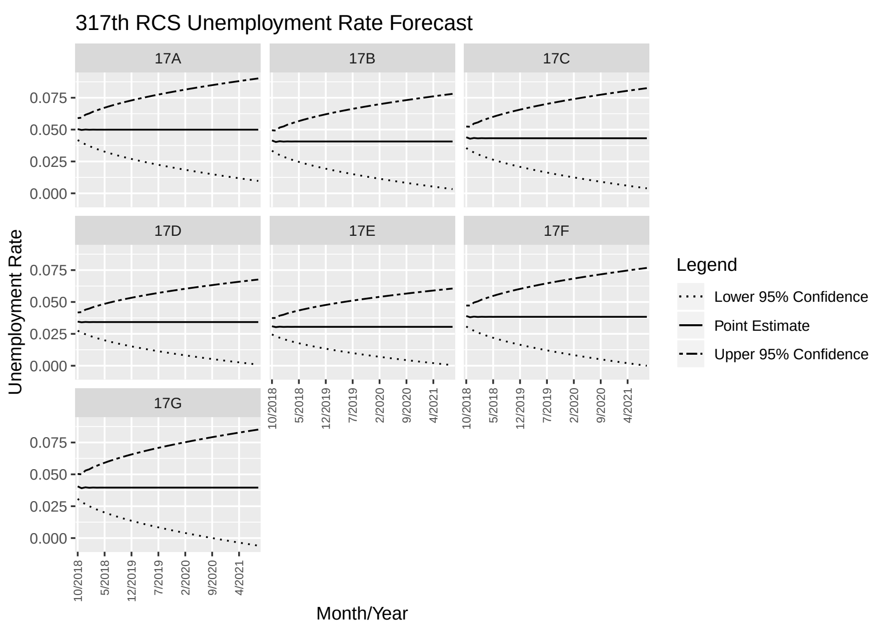


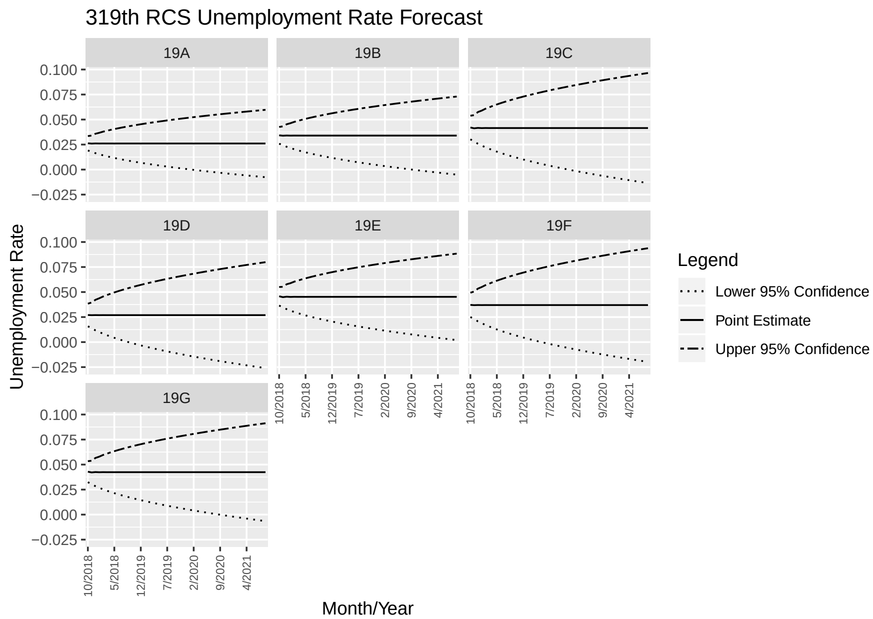
Figure 22. Obesity Rate Forecasts

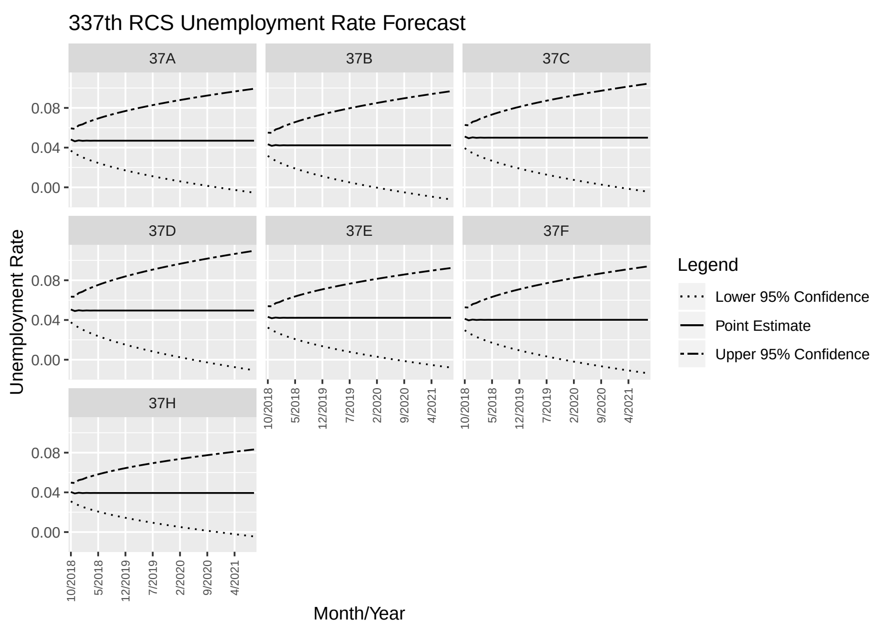

311th RCS Unemployment Rate Forecast
11A
11B
11C
11D
11E
11F
11G
11H
0.10
0.05
0.00
Unemployment Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

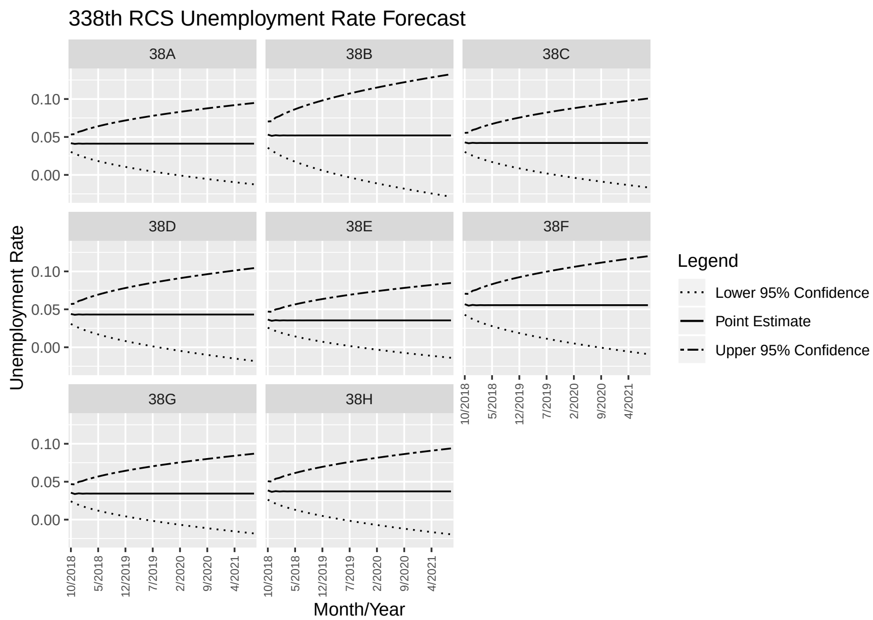

313th RCS Unemployment Rate Forecast
13A
13B
13C
13D
13E
13F
13G
13H
0.08
0.04
0.00
Unemployment Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


314th RCS Unemployment Rate Forecast
14A
14B
14C
14D
14E
14F
14G
Unemployment Rate
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


317th RCS Unemployment Rate Forecast
17A
17B
17C
17D
17E
17F
17G
Unemployment Rate
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


319th RCS Unemployment Rate Forecast
19A
19B
19C
19D
19E
19F
19G
0.100
0.075
0.050
0.025
0.000
−0.025
Unemployment Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


337th RCS Unemployment Rate Forecast
37A
37B
37C
37D
37E
37F
37H
0.08
0.04
0.00
Unemployment Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence



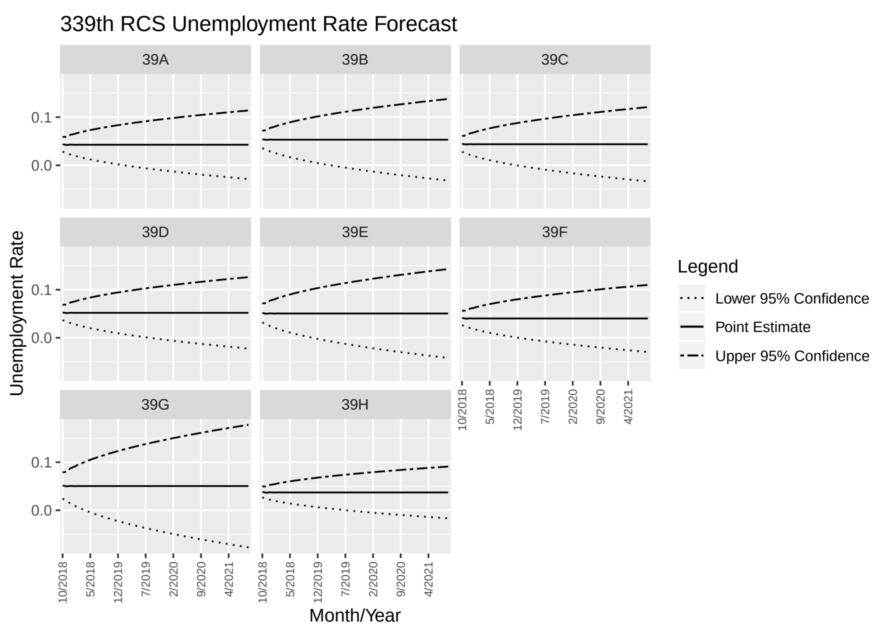
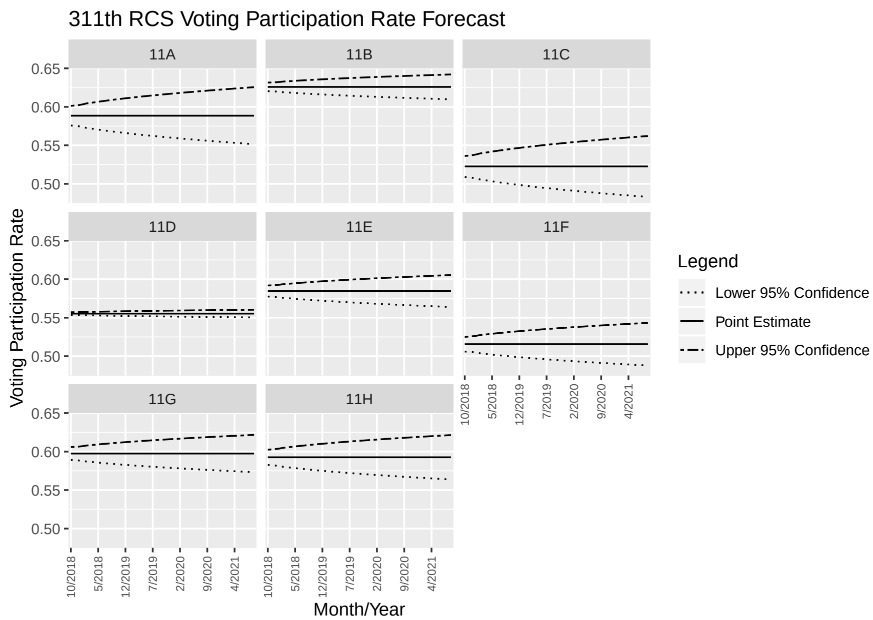
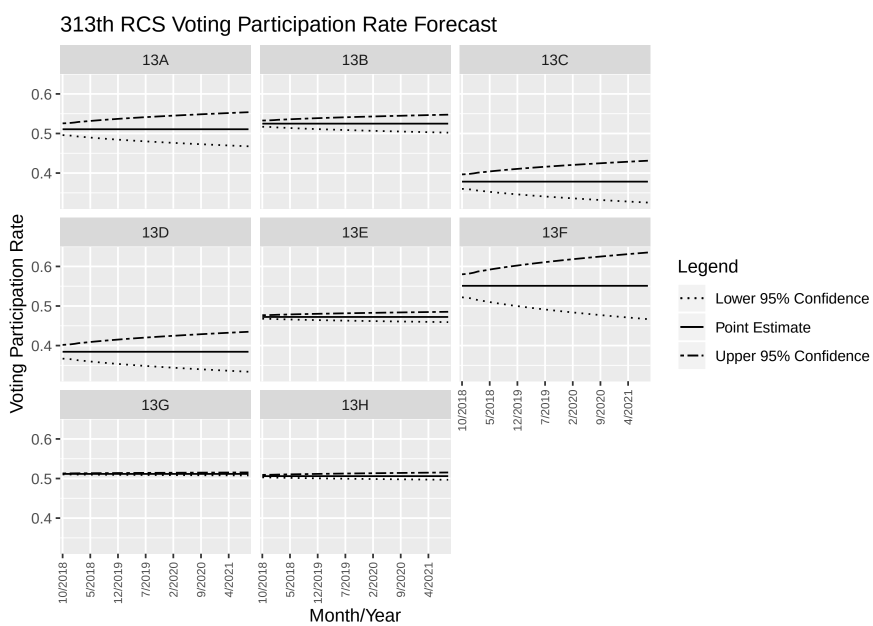
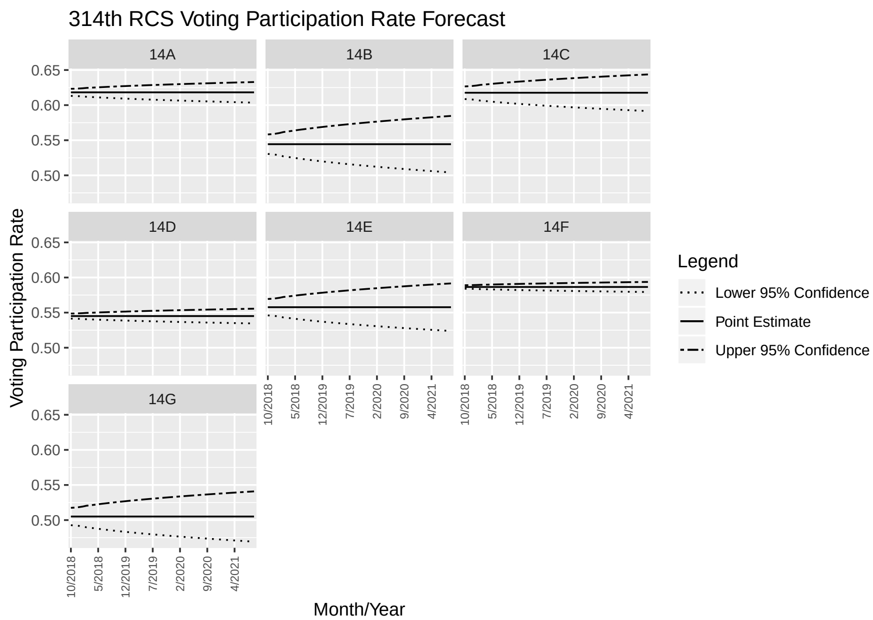
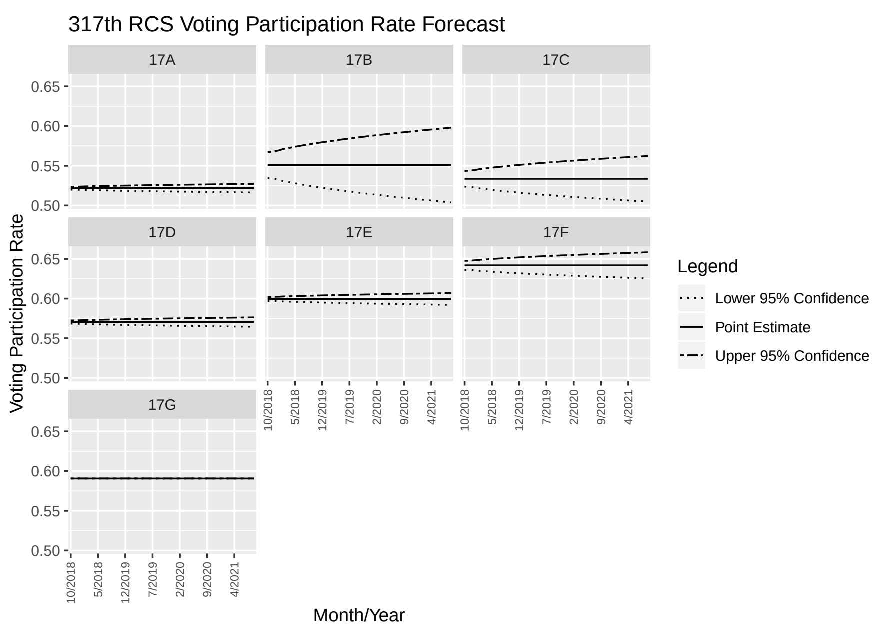


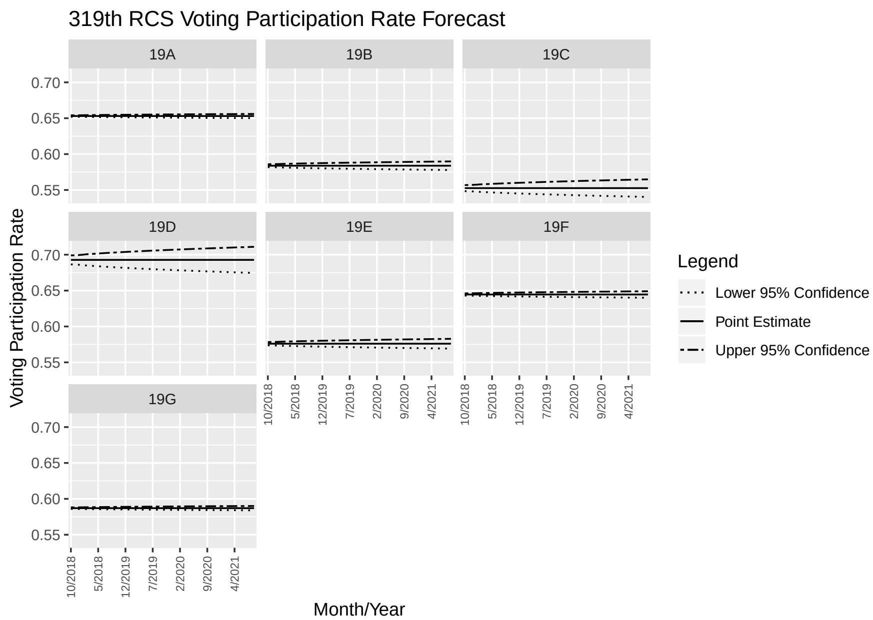
Figure 23. Unemployment Rate Forecasts


311th RCS Voting Participation Rate Forecast
11A
11B
11C
11D
11E
11F
11G
11H
0.65
0.60
0.55
0.50
Voting Participation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

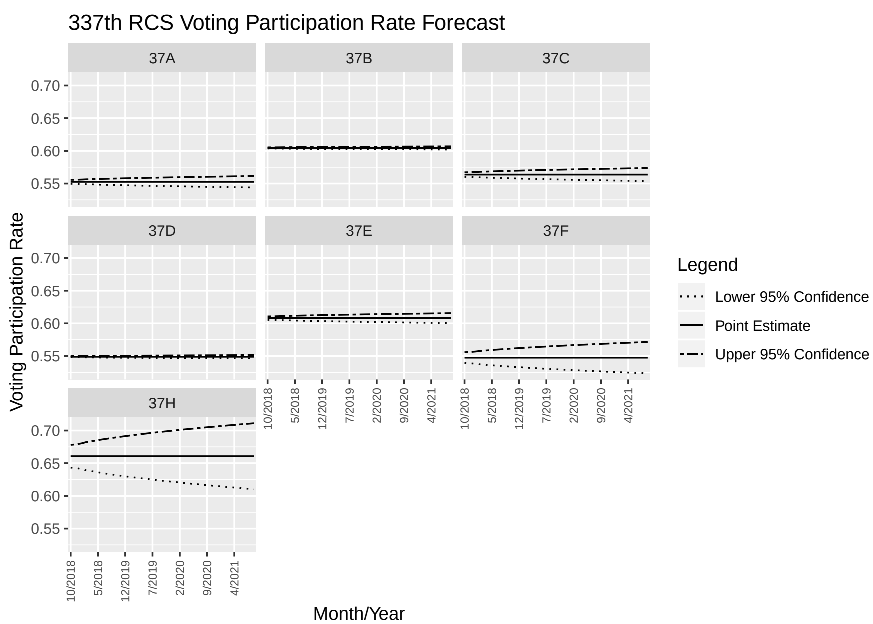

313th RCS Voting Participation Rate Forecast
13A
13B
13C
13D
13E
13F
13G
13H
0.6
0.5
0.4
Voting Participation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence


314th RCS Voting Participation Rate Forecast
14A
14B
14C
14D
14E
14F
14G
0.65
0.60
0.55
0.50
Voting Participation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence

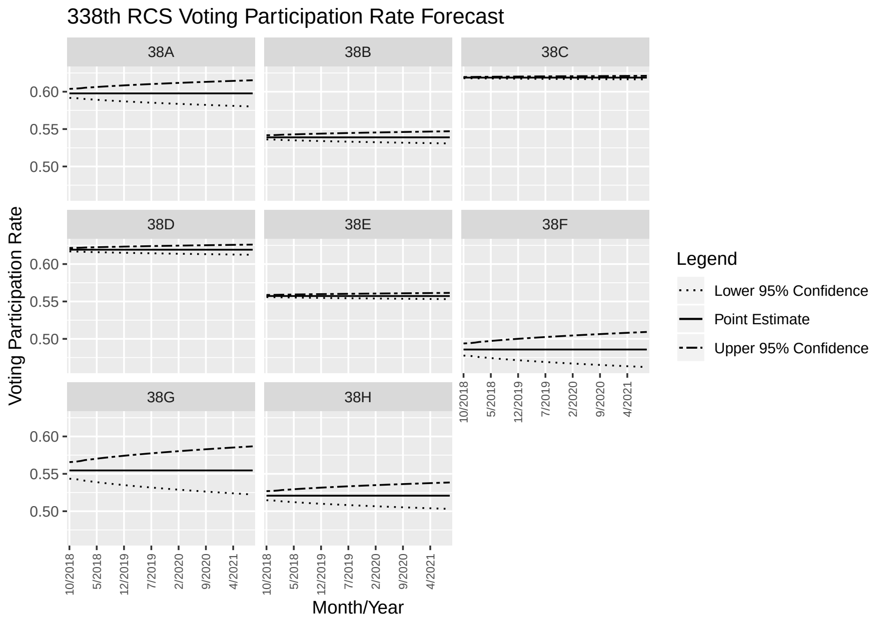

317th RCS Voting Participation Rate Forecast
17A
17B
17C
17D
17E
17F
17G
0.65
0.60
0.55
0.50
Voting Participation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence



319th RCS Voting Participation Rate Forecast
19A
19B
19C
19D
19E
19F
19G
0.70
0.65
0.60
0.55
Voting Participation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence



337th RCS Voting Participation Rate Forecast
37A
37B
37C
37D
37E
37F
37H
0.70
0.65
0.60
0.55
Voting Participation Rate
10/2018
5/2018
12/2019
7/2019
2/2020
9/2020
4/2021
Month/Year
Legend
Lower 95% Confidence
Point Estimate
Upper 95% Confidence



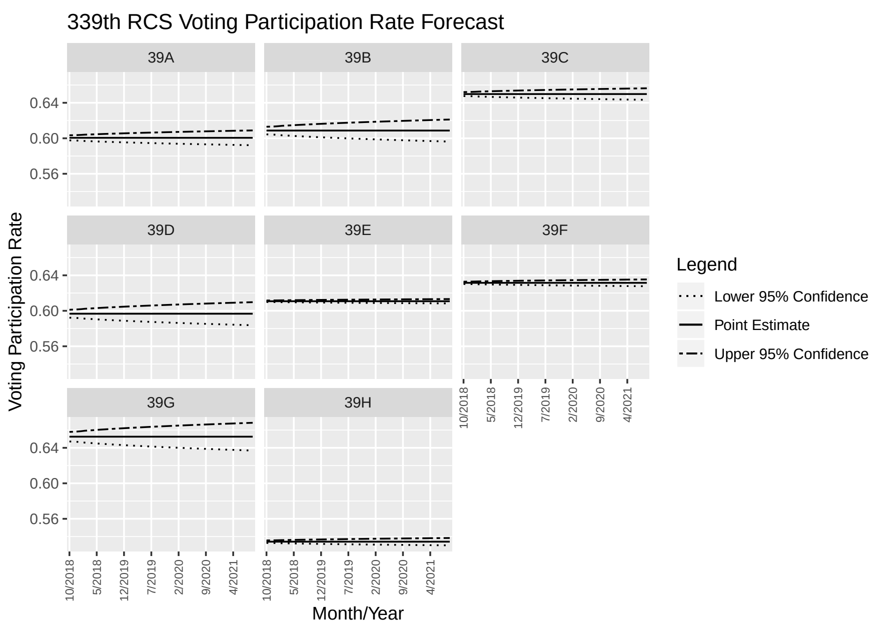


Figure 24. Voting Participation Rate Forecasts

Appendix C. Full Model Results

Table 11. Goal Model Results

Flight ID	β_o	β_{x_6}	$\beta_{x_6^2}$	β_{x_7}	$\beta_{x_{10}}$	$\beta_{x_{10}^2}$	$\beta_{x_{15}}$	$\beta_{x_{20}}$	$\phi_{y_{t-1}}$
11A	1.028	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.28
11B	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.335
11C	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.179
11D	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.641
11E	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.505
11F	1.105	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.367
11G	1.223	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.256
11H	1.256	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.259
13A	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.263
13B	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.573
13C	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.184
13D	1.316	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.628
13E	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.435
13F	1.082	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.200
13G	1.073	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.489
13H	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.157
14A	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.400
14B	1.073	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.023
14C	1.103	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.234
14D	1.061	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.023
14E	1.099	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	-0.031
14F	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.162
14G	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.368
17A	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.258
17B	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.594
17C	1.191	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.149
17D	1.318	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.029
17E	1.321	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	-0.187
17F	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.367
17G	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.597
19A	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.265
19B	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.389
19C	1.118	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.349
19D	1.073	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.259
19E	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.553
19F	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.292
19G	1.101	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.240
37A	1.236	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.146
37B	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.329
37C	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.189
37D	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.516
37E	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	-0.151
37F	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.341
37H	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.132
38A	1.209	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.162
38B	1.107	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.438
38C	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.351
38D	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.053
38E	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	-0.005
38F	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.266
38G	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	-0.121
38H	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.235
39A	1.105	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.619
39B	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.307
39C	1.234	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.224
39D	1.106	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.609
39E	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.421
39F	1.186	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.390
39G	1.087	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	0.635
39H	1.159	0.627	-7.38	0.296	0.146	-1.07	0.026	0.006	1.159

Table 12. Produced Model Results

Flight ID	β_o	β_{y_1}	β_{x_4}	$\beta_{x_{18}}$	$\phi_{y_{t-1}}$	$\phi_{y_{t-2}}$
11A	0.198	0.929	-6.041	0.061	0.244	-0.119
11B	0.384	0.929	-6.041	0.061	0.355	0.175
11C	1.633	0.929	-6.041	0.061	0.347	-0.050
11D	1.633	0.929	-6.041	0.061	0.043	0.0100
11E	2.519	0.929	-6.041	0.061	0.252	0.058
11F	0.607	0.929	-6.041	0.061	0.216	0.033
11G	2.586	0.929	-6.041	0.061	0.171	0.137
11H	2.315	0.929	-6.041	0.061	0.291	0.269
13A	1.633	0.929	-6.041	0.061	0.031	0.0630
13B	1.633	0.929	-6.041	0.061	0.165	-0.006
13C	1.633	0.929	-6.041	0.061	0.230	0.012
13D	1.633	0.929	-6.041	0.061	0.155	0.107
13E	1.633	0.929	-6.041	0.061	0.071	0.215
13F	1.633	0.929	-6.041	0.061	0.128	-0.114
13G	1.633	0.929	-6.041	0.061	-.0217	0.193
13H	1.633	0.929	-6.041	0.061	0.078	-0.095
14A	1.633	0.929	-6.041	0.061	0.242	0.134
14B	1.633	0.929	-6.041	0.061	0.194	-0.060
14C	1.633	0.929	-6.041	0.061	0.323	-0.080
14D	2.704	0.929	-6.041	0.061	0.339	0.155
14E	1.633	0.929	-6.041	0.061	0.118	-0.068
14F	2.723	0.929	-6.041	0.061	0.004	-0.022
14G	1.633	0.929	-6.041	0.061	0.117	0.006
17A	0.976	0.929	-6.041	0.061	0.025	0.188
17B	1.556	0.929	-6.041	0.061	0.295	0.198
17C	1.633	0.929	-6.041	0.061	0.179	-0.034
17D	1.633	0.929	-6.041	0.061	0.311	0.018
17E	2.887	0.929	-6.041	0.061	0.340	0.155
17F	1.633	0.929	-6.041	0.061	0.203	0.016
17G	0.932	0.929	-6.041	0.061	-0.008	-0.056
19A	1.633	0.929	-6.041	0.061	0.215	0.017
19B	0.438	0.929	-6.041	0.061	0.061	0.057
19C	1.633	0.929	-6.041	0.061	0.127	0.008
19D	1.633	0.929	-6.041	0.061	0.422	-0.114
19E	3.729	0.929	-6.041	0.061	0.361	0.033
19F	0.949	0.929	-6.041	0.061	0.220	0.051
19G	1.633	0.929	-6.041	0.061	0.287	0.065
37A	1.897	0.929	-6.041	0.061	0.361	0.079
37B	1.633	0.929	-6.041	0.061	0.034	0.146
37C	1.633	0.929	-6.041	0.061	0.424	0.143
37D	1.576	0.929	-6.041	0.061	0.582	0.203
37E	2.970	0.929	-6.041	0.061	0.246	-0.131
37F	1.633	0.929	-6.041	0.061	0.112	0.0397
37H	2.757	0.929	-6.041	0.061	0.142	-0.185
38A	2.522	0.929	-6.041	0.061	0.435	0.051
38B	1.633	0.929	-6.041	0.061	0.206	-0.118
38C	1.672	0.929	-6.041	0.061	0.021	-0.065
38D	2.020	0.929	-6.041	0.061	0.373	0.283
38E	1.633	0.929	-6.041	0.061	0.0269	-.0068
38F	1.633	0.929	-6.041	0.061	0.143	-0.84
38G	1.633	0.929	-6.041	0.061	0.132	0.070
38H	1.404	0.929	-6.041	0.061	0.245	0.223
39A	1.633	0.929	-6.041	0.061	0.271	0.014
39B	1.633	0.929	-6.041	0.061	0.238	-0.018
39C	1.633	0.929	-6.041	0.061	0.383	-0.085
39D	1.633	0.929	-6.041	0.061	0.216	0.060
39E	1.633	0.929	-6.041	0.061	0.112	0.014
39F	2.575	0.929	-6.041	0.061	0.211	-0.032
39G	1.633	0.929	-6.041	0.061	0.292	-.046
39H	1.633	0.929	-6.041	0.061	1.633	1.633

Bibliography

1. "Air Force Recruiting Service Groups/Squadrons." https://www.recruiting.af.mil/About-Us/Groups-Squadrons/. Accessed 26 July 2019.

2. R. V. Cooper, *Military Manpower and the All-Volunteer Force.* Santa Monica California: RAND Corporation, 1977.

3. "Air Force Recruiting Service." https://www.recruiting.af.mil/About-US/Fact-Sheets/Display/Article/714536/air-force-recruiting-service/. Accessed 4 June 2019.

4. "The Enlisted Path: Get Airman Ready." https://www.airforce.com/how-to-join/process/enlisted. Accessed 26 July 2019.

5. U.S. Department of the Air Force, "Air Force Recruiting Service Procedural Guidance." `https://mypers.af.mil/ci/fattach/get/10009912/1557761459/redirect/1/filename/AFRS%20Procedural%20Guiance%20PSDG%2013%20May%202019.pdf`, 13 May 2019. Accessed 17 May 2019.

6. Air Force Recruiting Service, "FY19 AMMO Metrics," 1 October 2018.

7. "360th Recruiting Group." https://www.recruiting.af.mil/About-Us/Fact-Sheets/Display/Article/714506/360th-recruiting-group/. Accessed 26 July 2019.

8. M. Murrary and L. McDonald, *Recent Recruiting Trends and Their Implications for Models of Enlistment Supply.* Santa Monica California: RAND Corporation, 1999.

9. B. Orvis and B. Asch, *Military Recruiting: Trends, Outlook, and Implications.* Santa Monica California: RAND Corporation, 2001.

10. J. Dertouzos, *Recruiter Incentives and Enlistment Supply.* Santa Monica California: RAND Corporation, 1985.

11. C. J. Simon and J. T. Warner, "Managing the all-volunteer force in a time of war," *The Economics of Peace and Security Journal*, vol. 2, no. 1, pp. 20 – 29, 2007.

12. B. C. Intrater, J. K. Alt, S. E. Buttrey, and J. B. House, "Understanding the Impact of Socioeconomic Factors on Navy Accessions," *Military Operations Research*, vol. 23, no. 1, pp. 31 – 47.

13. J. L. McDonald, E. D. White, R. R. Hill, and C. Pardo, "Forecasting US Army enlistment contract production in complex geogrphical marketing areas," *Journal of Defense Analytics and Logistics*, vol. 1, no. 1, pp. 69 – 87.

14. J. L. Williams, "Identifying Demand Indicators for Air Force Recruiting Service With Discriminant Analysis," Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, 2001.

15. B. M. Fulton, "Determining Market Categorization of United States Zip Codes for Purposes of Army Recruiting," Master's thesis, Naval Postgraduate School, Monterey, CA, 2016.

16. E. M. Monaghan, "Estimating the Depth of the Navy Recruiting Market," Master's thesis, Naval Postgraduate School, Monterey, CA, 2016.

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

18. J. McDonald, "Analysis and Modeling of U.S. Army Recruiting Markets," Master's thesis, Air Force Insitute of Technology, Wright-Patterson AFB, OH, 2016.

19. J. Dertouzos and S. Garber, *Human Resource Management and Army Recruiting: Analyses of Policy Options*. Santa Monica California: RAND Corporation, 2006.

20. W. E. Alley, "Effect of Air Force Recruiting Incentives on Volunteer Enlistment," tech. rep., Air Force Human Resources Lab, Lackland AFB, TX, 1976.

21. M. R. Kilburn and B. Asch, *Recruiting Youth in the College Market: Current Practices and Future Policy Options*. Santa Monica California: RAND Corporation, 2003.

22. U.S. Census Bureau, "Explore Census Data." https://data.census.gov/cedsci/all?q=populationhidePreview=falsetable=DP05tid =ACSDP1Y2018.DP05lastDisplayedRow=15. Accessed 10 Sepetember 2019.

23. U.S Bureau of Labor Statistics, "Local Area Unemployment Statistics (One-Screen Data Search." https://data.bls.gov/PDQWeb/la. Accessed 20 June 2019.

24. County Health Rankings and Roadmap, "Rankings Data and Documentation." https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation. Accessed 30 June 2019.

25. T. McGovern, "US County Level Election Results 08-16." `https://github.com/tonmcg/US_County_Level_Election_Results_08-16`. Accessed 11 September 2019.

26. "The Target Population for Military Recruitment: Youth Eligible to Enlist Without a Waiver." `https://dacowits.defense.gov/Portals/48/Documents/General%20Documents/RFI%20Docs/Sept2016/JAMRS%20RFI%2014.pdf?ver=2016-09-09-164855-510`. Access 31 October 2019.

27. W. Trahan, “AFRISS undergoes system upgrade.” https://www.recruiting.af.mil/News/Article-Display/Article/697778/afriss-undergoes-system-upgrade/. Accessed 14 October 2019.

28. Military One Source, “Demographic Profiles.” https://www.militaryonesource.mil/reports-and-surveys/demographic-profiles. Accessed 9 August 2019.

29. “WINGS GSU Locator.” `https://holmcenter.com/locator/`. Access 4 December 2019.

30. D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2015.

31. The National Bureau of Economic Research, “County Distance Database.” https://www.nber.org/data/county-distance-database.html. Accessed 25 September 2019.

32. U.S. Census Bureau, “ZIP Code Tabulation Areas (ZCTAs).” https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html. Accessed 12 August 2019 .

33. U.S. Census Bureau, “Explanation of the 2010 ZIP Code Tabulation Area (ZCTA) to County Relationship File.”

34. U.S. Census Bureau, “2010 ZCTA to County Relationship File.”

35. UDS Mapper, “ZIP Code to ZCTA Crosswalk.” https://www.udsmapper.org/zcta-crosswalk.cfm. Accessed 19 September 2019.

36. D. T. Larose and C. D. Larose, *Data Mining and Predictive Analytics*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2 ed., 2015.

37. W. R. Golden and M. Goldstein, *Multivariate Analysis: Methods and Applications*. John Wiley & Sons, Inc., 1 ed., 1984.

38. M. Brems, “A One-Stop Shop for Principal Component Analysis.” https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c. Accessed 20 October 2019.

39. D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc., 5 ed., 2012.

40. B. L. Bowerman, R. T. O’Connel, and A. B. Koehler, *Forecasting, Time Series, and Regression*. Brooks/Cole, 4 ed., 2005.

41. "A Practical Introduction to Factor Analysis: Exploratory Factor Analysis." https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/. Accessed 22 January 2020.

42. J. Fox and S. Weisberg, *An R Companion to Applied Regression*. Thousand Oaks CA: Sage, third ed., 2019.

43. "Durbin-watson significance tables." https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf. Accessed 10 Jan 2020.

44. I. Svetunkov, *smooth: Forecasting Using State Space Models*, 2019. R package version 2.5.4.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
26–03–2020	Master's Thesis	Sept 2018 — Mar 2020

4. TITLE AND SUBTITLE

Analysis and Forecasting of the 360th Air Force Recruiting Group Goal Distribution

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Spangler, Tyler, Capt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-20-M-172

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

360th Recruiting Group
Bldg 54, Bay 5, Door 14
New Cumberland, PA 17070-5097
DSN 771-6897, COMM 717-770-6897
Email: timothy.goodroe@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

360 RCG

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A:
Approved for Public Release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This research utilizes monthly data from 2012-2017 to determine economic or demographic factors that significantly contribute to increased goaling and production potential in areas of the 360th Recruiting Group's. Using regression analysis, a model of recruiting goals and production is built to identify squadrons within the 360 RCG's zone that are capable of producing more or fewer recruits and the factors that contribute to this increased or decreased capability. This research identifies that a zone's high school graduation rate, the number of recruiters, and the number of JROTC detachments in a zone are positively correlated with recruiting goals and that a zone's obesity rate and voting participation rate are inversely related to goaling. Additionally, this research found that the monthly number goals and the number of JROTC detachments in a zone are positively correlated with recruit production and the unemployment rate is inversely related with production. Using these two linear regression models, recruiting goals and recruiting production are projected into 48 months of new data to identify higher goaling and production potential squadrons.

15. SUBJECT TERMS

Recruiting, Forecasting, Statistical Analyses

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON / 19b. TELEPHONE NUMBER *(include area code)*
U	U	U	U	121	Dr. Raymond R. Hill, AFIT/ENS; (937) 255-3636, x7469; rhill@afit.edu