NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

SELECTED MARINE CORPS RESERVE PERSONNEL STATUS FORECASTING

by

Cris A. Streetzel

December 2018

Thesis Advisor: Robert A. Koyak
Second Reader: Andrew Dausman,
Manpower & Reserve Affairs,
Headquarters Marine Corps

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE SELECTED MARINE CORPS RESERVE PERSONNEL STATUS FORECASTING	**5. FUNDING NUMBERS**
6. AUTHOR(S) Cris A. Streetzel	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) HQ USMC M&RA, Quantico, VA 22134	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

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

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

Select Marine Corp Reserve is a reserve component of the United States Marine Corps currently composed of approximately 32,000 Marines participating in weekend drills and active duty training, or activated for service. The retention and recruiting missions of this organization require an accurate population forecast with properly fit prediction intervals. We use decision trees to estimate the future population and variance of a cohort from individual personnel records. Our algorithm provides useful forecasts and prediction intervals up to 12 months into the future. Despite the success of the algorithm, seasonality remains an issue. We recommend further study to remove seasonality from this algorithm.

14. SUBJECT TERMS machine-learning, decision tree, USMC, Reserve, personnel, military attrition, military accession	**15. NUMBER OF PAGES** 75
	16. PRICE CODE

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

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

SELECTED MARINE CORPS RESERVE PERSONNEL STATUS FORECASTING

Cris A. Streetzel
Major, United States Army
BS, Florida Gulf Coast University, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 2018

Approved by: Robert A. Koyak
Advisor

Andrew Dausman
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Select Marine Corp Reserve is a reserve component of the United States Marine Corps currently composed of approximately 32,000 Marines participating in weekend drills and active duty training, or activated for service. The retention and recruiting missions of this organization require an accurate population forecast with properly fit prediction intervals. We use decision trees to estimate the future population and variance of a cohort from individual personnel records. Our algorithm provides useful forecasts and prediction intervals up to 12 months into the future. Despite the success of the algorithm, seasonality remains an issue. We recommend further study to remove seasonality from this algorithm.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION....1
A. BACKGROUND / LITERATURE REVIEW....1
1. Problem History....1
2. Problem Statement....4
3. Objective....4
B. SCOPE, LIMITATIONS AND ASSUMPTIONS....4
C. ORGANIZATION OF THESIS....5

II. PREVIOUS STUDIES....7

III. DATA AND METHODOLOGY....9
A. DATA SETS....9
1. MCRISS....9
2. TFDW....10
B. VARIABLES....11
1. Outcome Variable (Status Group Population)....11
2. Explanatory Variables....12
C. ALGORITHM DEVELOPMENT....12
1. System Characterization....12
2. Mathematical Relationships....14
3. Complexity Tuning....19
D. METHODOLOGY....21
1. Import and Initial Formatting....21
2. Feature Selection....21
3. Model Training....22
4. Forecast Computation....23

IV. ANALYSIS....25
A. IADT STATUS GROUP....25
1. Explanatory and Predicted Variable—Relationships....25
2. Output Analysis....30
B. SMCR STATUS GROUP....37
1. Explanatory and Predicted Variable—Relationships....37
2. Output Analysis....41
C. EFFECT OF APPLYING VARIANCE RATIOS....46

V. CONCLUSIONS AND RECOMMENDATIONS....51

A. CONCLUSIONS 51
B. POTENTIAL APPLICATIONS 51
C. TOPICS FOR FURTHER RESEARCH 52
1. Projection of the Unknown Population 52
2. Post-model Seasonality Adjustment 52

APPENDIX. STATUS CODES 53

LIST OF REFERENCES 55

INITIAL DISTRIBUTION LIST 57

LIST OF FIGURES

Figure 1. SMCR Accession and Attrition Process 3

Figure 2. Example of the Divergence of Variance between Training and Test Sets 20

Figure 3. IADT Explanatory Variable Relative Importance 26

Figure 4. IADT Predictions versus DODTCPG (One-Month Projection) 27

Figure 5. IADT Predicted by DODTCPG and Obligation Remaining (Three-Month Projection) 28

Figure 6. IADT Predicted by DODTCPG and Obligation Remaining 29

Figure 7. IADT Predicted by Fiscal Year (FY) (Seven-Month Projection) 30

Figure 8. IADT Prediction Interval Width by Projection Period 31

Figure 9. Training and Hindcast Errors by FY 33

Figure 10. IADT Error by Projection Period 34

Figure 11. Sample IADT Hindcast Projections 36

Figure 12. IADT Errors versus Calendar Month 37

Figure 13. SMCR Explanatory Variable Relative Significance 38

Figure 14. SMCR Predicted by DODTCPG & EAS Remaining (One-Month Projection) 39

Figure 15. SMCR Predicted by RCOMPCODE & Obligation Remaining (Eight-Month Projection) 40

Figure 16. SMCR Prediction Interval Width by Projection Period 42

Figure 17. SMCR Training and Hindcast Errors by SEQ 43

Figure 18. SMCR Error by Projection Period 44

Figure 19. Sample SMCR Hindcast Projections 45

Figure 20. SMCR Error versus Calendar Month 46

Figure 21. Variance Ratio (VR) by Projection Horizon 47

Figure 22. Comparison of Prediction Interval Adjustments.....48

LIST OF TABLES

Table 1. Selected MCRISS Fields....9

Table 2. TFDW Fields....10

Table 3. Status Groups....11

Table 4. Constructed Variables....12

Table 5. Effect of Tree Fit on the Bias of Variance Estimates of a Binomial Distribution....17

Table 6. Final Explanatory Variables....22

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The Selected Marine Corps Reserve (SMCR) encompasses approximately 32,000 Marines. Manpower and Reserve Affairs (M&RA)- Headquarters Marine Corps requires accurate manpower projections to manage the training pipeline and recruiting mission, and to maximize end-strength. Previous projection models have exhibited shortcomings in accuracy or seasonality. These models have been limited to a few explanatory variables and unable to make full use of the wealth of data available on current Marines and recruits. We construct an algorithm that projects future population values with improved accuracy and valid prediction intervals in order to support M&RA decision making.

The future population of the SMCR can be broken into two categories: 1) known individuals in the recruiting pipeline or reservists continuing in SMCR to the projected timeframe, and 2) individuals who do not yet exist in SMCR personnel systems but will access into SMCR prior to the projected timeframe. We construct a decision-tree model to forecast the accession and attrition of known personnel and estimate the forecast uncertainty. Used properly, decision trees, which fall under the category of machine–learning techniques, provide a means for extracting information that has the most predictive power from SMCR personnel systems while minimizing potential noise variables. With these methods, we make maximum use of actionable demographic and administrative records available for each Marine in the population to project the future population of a cohort.

We also develop a novel method to extract and calibrate the uncertainty estimates from the decision tree itself. We do this by estimating the binomial variance of the decision tree leafs and calibrating it with variance data obtaining during training. This yields prediction intervals that are relatively insensitive to the fit of the decision tree. While our results are not perfect, the algorithm adequately estimates the projection uncertainty in historical projection tests. This modeling capability provides improved projections for decision making in the areas of recruiting, retention, and budgeting.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I thank all the instructors of the NPS Operations Research program. It is only through their brilliance and patience that we (eventually) learned to add. Most of all, I thank Alex, Selena, and DK. Sorry for all the missed walks and playtimes….

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND / LITERATURE REVIEW

1. Problem History

The Selected Marine Corps Reserve (SMCR) is the largest sub-component of the Marine Corps Reserve. The Reserve was formed on 29 Aug 1916 in preparation for U.S. involvement in World War I. Since then, U.S. Marine Corps Reservists have participated in every U.S. conflict, both as part of reserve units and as individual augmentees to active duty units. As of 2017, the SMCR was composed of approximately 32,000 Marines participating in weekend drill, active duty training, or activated for service. In any particular month, approximately 1,400 SMCR recruits are attending 13 weeks of Initial Active Duty Training (IADT) and varying durations of job-related training. In order to fully man a ready and trained pool of Marines to meet its force and budget requirements, Manpower and Reserve Affairs (M&RA - Headquarters Marine Corps is tasked to forecast the number of available Reserve Marines to support decisions related to recruiting targets, Initial Active Duty Training dates, bonuses, manpower budgeting, and more. An accurate population model is critical for all of these decisions.

New entries into the military are termed accessions while exit from the military is termed attrition. M&RA employs two forecasting models to predict the evolution of SMCR accessions and one model to predict SMCR attrition rates. Manpower analysts combine the projections of the attrition model and one of the accession models to predict the aggregate population totals that are needed for decision making. In operational use, shortcomings occur in all three models. As implemented, none of the models provides prediction intervals to quantify their uncertainty. Based on analyst feedback, the one accession model has unacceptably high error rates in the spring and the other model is too inaccurate for operational use. Finally, the attrition model has higher error rates than is desired by M&RA. Due to the large uncertainties inherent in the projections, M&RA must manage the population levels, budgetary requirements, and recruiting goals more conservatively than otherwise so as not to exceed statutory or budgetary limits (S. Norton, Maj USMC-M&RA,

personal communication, Apr 27, 2017). M&RA requested this study to improve the accuracy of the aggregate forecast and allow them to manage the population with lower safety margins.

The future SMCR population comes from two distinct subsets:

- The known population, which consists of personnel who are in the Delayed Entry Program or are already drilling Marines. These individual are represented in the Marine Corps Recruiting Information Support System (MCRISS) or Total Force Data Warehouse (TFDW) databases with varying levels of detail;

- The unknown population, which consists of personnel who are not part of the known population, but who will become part of the SMCR during the period of interest. No information on these individuals is available at the time of forecasting

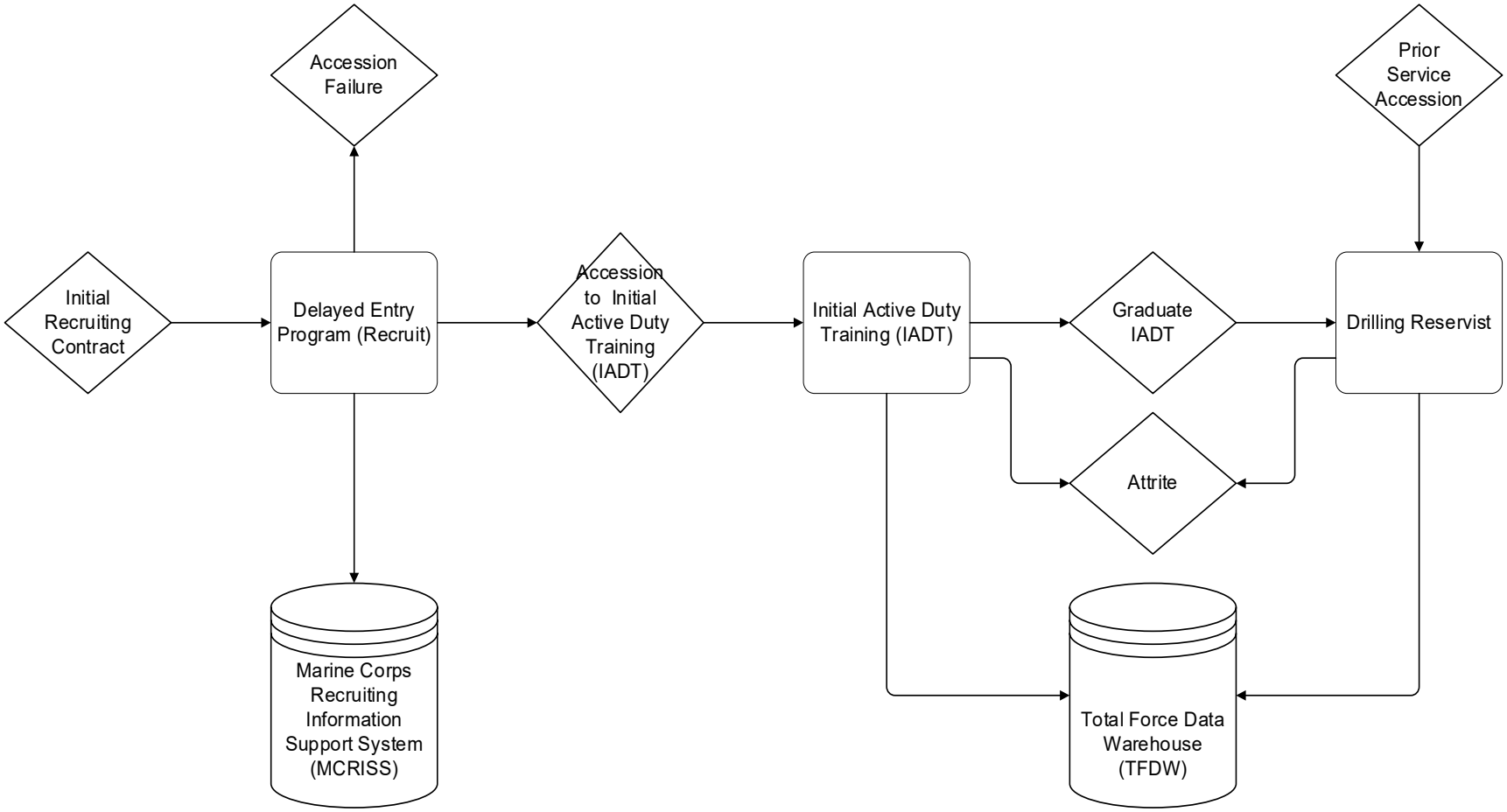

Figure 1. SMCR Accession and Attrition Process

Figure 1 illustrates the SMCR accession and attrition process. New recruits are recorded in the MCRISS database as they enter the Delayed Entry Program (DEP). They may attrite out of DEP or remain in it until they are accessed into Initial Active Duty Training (IADT). Once at IADT, their records are maintained in the Total Force Data Warehouse (TFDW) until their discharge from the SMCR. If they do not attrite in IADT, they graduate and transition to become a drilling Reservist. They remain in that group until they attrite, either by discharge or retirement. In addition to this process, personnel may transition directly from active duty into the SMCR without undergoing the training or recruiting process.

2. Problem Statement

How can the M&RA forecast models be improved? M&RA requires accurate forecasting to support its decision points and planning. It lacks a forecast model that has an acceptable level of error across multiple time-spans. The SMCR also requires reliable models to quantify the forecast uncertainty.

3. Objective

We construct an algorithm that provides improved accuracy and valid prediction intervals in order to forecast the evolution of accessions and attrition of the aggregate known population of the SMCR over future periods ranging from one to twelve months. Our goal is to produce a production algorithm that predicts the size of the known population of Marines at multiple time-scales. It is critical to M&RA decision process that we generate proper prediction intervals on our estimates. The final algorithm should be retrainable in a production environment with minimal upkeep or specialty knowledge. Finally, the algorithm should be able to adapt to future changes in population characteristics.

B. SCOPE, LIMITATIONS AND ASSUMPTIONS

The proposed algorithm performs analysis on Selected Marine Corps Reserve (SMCR) personnel records as extracted from current databases, with no requirement for manual pre-processing or cleanup. The algorithm provides the following output products:

- For each Marine in the current population, an estimate of the probability of SMCR membership at each projection horizon.
- The forecast of the size of the visible SMCR population of the known cohort for each projection horizon with a 95% prediction interval.

This study limits all predictions to the known population (the Marines listed in SMCR databases) at the time of the prediction. We make no provision for estimating the unknown population.

C. ORGANIZATION OF THESIS

We approach this problem sequentially. First, we generate a sequence of models to estimate each prediction horizon for each status group. Then we estimate the variance of the prediction and of the prediction errors. We then use the resulting values to predict the size of a validation dataset.

THIS PAGE INTENTIONALLY LEFT BLANK

II. PREVIOUS STUDIES

There have been many studies of general military attrition behaviors and several that attempt to develop SMCR-specific models. Buddin (1984) studies the factors influencing early attrition behavior (defined as the first six months of service). He uses a multivariate regression model to analyze survey from the year 1979 and personnel data of active duty recruits of all service branches. Buddin found a strong correlation with pre-enlistment work history. Buddin finds that high school graduates have much lower attrition rates. Most importantly, Buddin finds that older recruits have higher attrition rates in the first six months while having lower attrition rates at the 36-month mark.

Baykiz (2007) utilizes logistic regression to analyze attrition among military recruits. Baykiz finds that female recruits, unmarried recruits, recruits with no children, and recruits with lower AFQT scores exhibited higher attrition behaviors in this phase. Interestingly, Baykiz also finds enlistment that occurred late in a month incurred higher attrition and that high school seniors attrited more in March and April of each year.

Emery (2010) focuses on forecasting the loss component of the SMCR end-strength forecast. He updates SMCR's existing weighted moving-average model with an exponentially smoothed error adjustment. He applies his technique to projecting each loss category instead of a single aggregated population and is able to achieve slight improvement over the legacy model.

Erhardt (2012) attempts to apply a Markov model to forecasting SMCR losses. Erhardt finds that losses exhibited significant seasonality, which prevents an annually aggregated model from achieving stationarity. He recommends a monthly aggregated Markov model in future studies.

Dausman (2016) applies Markov models to both SMCR accessions and losses, but this time with more success. Dausman developed separate Markov models for each Marine Military Occupational Specialty (MOS) based on average and weighted-average transition rates. His result is the current model in use for SMCR population forecasting, but the model

has suffered from seasonality errors since adoption (S. Norton, Maj USMC-M&RA, personal communication, Apr 27, 2017).

Based on our review of available literature, there are clear differences in the methodology of SMCR-specific prediction studies and general military attrition studies. These studies develop explanatory models that link the attributes of individual service-members to their attrition rate. The SMCR-specific studies have focused on the aggregate totals of the population or sub-population to make predictions. As a result, we find that previous predictive modeling studies do not take full advantage of the explanatory variables available for individual recruits that previous analytical studies found to be significant. Use of such data could improve the ability of forecast models to adapt to changing input distributions as they influence accession and attrition rates. While we are not aware of any study that directly analyzed influences on SMCR individual attrition, the studies of active component service-members are a useful starting point in determining which explanatory variables an improved model might incorporate. The aggregated and Markovian methodologies used in previous predictive studies precluded them from subdividing the data by all the explanatory variables available. Such an approach would make the models unmanageable and more sensitive to sampling errors. As a result, the models did not solve seasonality issues or incorporate all relevant individual data (S. Norton, Maj USMC-M&RA, personal communication, Apr 27, 2017).

III. DATA AND METHODOLOGY

In order to understand our approach to this modeling effort, we start with a brief review of the datasets available for incorporation into this model. We then discuss which variables survive screening and the outputs that the model is trained against. In Section III.C, we delve into the mathematics of our model and the complexities of determining prediction intervals from such a model. Finally, we demonstrate the application of the algorithm to our problem set.

A. DATA SETS

The datasets that contain SMCR Recruit and Marine records are extracted from two sources: the Marine Corps Recruiting Information Support System (MCRISS) and the Total Force Data Warehouse (TFDW). Each database is backed up on the last day of each month, a file that we term a "snapshot." Each snapshot consists of approximately 16,000 MCRISS records and 34,000 TFDW records. Arraying the snapshots across time allows us to profile the evolution of the records and the transitions the Marine moves through in a defined period.

1. MCRISS

The MCRISS dataset contains manually entered data on all initial recruits contracted for accession with SMCR. It does not contain any data on individuals pending transfer from active duty components to the SMCR. MCRISS has approximately 219,000 record snapshots available from 31 Oct 2005 to 30 Sep 2017. Table 1 lists the MCRISS fields available from M&RA.

Table 1. Selected MCRISS Fields

FIELD	DESCRIPTION
RSEQ	Sequence Code associated with the month the database snapshot was taken
START_WK	Date of initial entry into MCRISS
DISCHARGE_CODE	Reason Code for discharge from Delayed Entry Program
OCC_FLD	Occupational Field

FIELD	DESCRIPTION
MOS	Military Occupational Specialty
SHIP_DATE	Scheduled date for departure to Initial Active Duty Training (IADT)
ENLIST_TERM	Length of Initial Enlistment (Years)
PEBD	Pay Entry Base Date
RUC	Reserve Unit Code
SPLIT_I	Flag for indicate of recruit will split IADT attendance
QSN_PIVOT_DATE	Date of split IADT Training
EDIPI	Electronic Data Interchange Personal Identifier

2. TFDW

The TFDW dataset is an extract from the USMC Total Force Data Warehouse and is comprised of data on Marines currently in SMCR drilling Reserve and training statuses. At the initiation of this study, TFDW contained 4.9 million record snapshots available from 31 OCT 2005 to 30 SEP 2017. Table 2 lists all available SMCR data resident in the TFDW.

Table 2. TFDW Fields

FIELD	DESCRIPTION
SEQ	Sequence ID associated with the month the database snapshot was taken
DODTCPG	Current Service Status
COMPCODE	Identifies SMCR Members in a mobilized status
RCOMPCODE	Sub-Statuses to DODTCPG
RRECSTAT	The reporting status of SMCR Members
CIVILIAN_EDUC_LEVEL_CODE_1	Number of years of education completed
DATE_ENLISTMENT_OR_ACCEPTANCE	Date of enlistment contract
DATE_OF_BIRTH	Self-explanatory
EAS	End of Active Service (for training)
PMOS	Primary Military Occupations Specialty (MOS)
MARITAL	Marital Status
NUMBER_OF_DEPENDENTS	Number of family members
RUC	Reserve Unit Code
DATE_INITIAL_ENTRY_RESERVE	Date of entry into SMCR
EDIPI	Electronic Data Interchange Personal Identifier
HOR_ZIP	Home of Record Zip Code
HOR_STATE	Home of Record State
HOR_CITY	Home of Record City
AFQT_SCORE	Armed Forces Qualification Test score
CBT_FITNESS_SCORE_QY	Most recent Combat Fitness Test score

FIELD	DESCRIPTION
PHYS_FIT_SCORE_QY	Most recent Physical Fitness Test score
CURRENT_CITY	Current residence City
CURRENT_STATE	Current residence State
CURRENT_ZIP	Current residence Zip Code
MAND_DRILL_PARTIC_STOP_DT	Date that a Marine's service obligation ends
RACE	Racial self-identification
ETHGRP	Ethnic self-identification
SEX	Gender

B. VARIABLES

We divide the data attributes into two categories: an outcome variable, which we use to fit our model and assess errors; and explanatory variables, which provide descriptive information on Marines so that we may categorize them.

1. Outcome Variable (Status Group Population)

The TFDW DODTCPG field provides the current status of SMCR members. We group the DODTCPG statuses into sponsor-specified status categories. The predicted variable of each model is comprised of a binary encoding of one of the status groups. We match the status group value at a future time to a set of explanatory variables from the current time. We provide additional information on SMCR personnel status codes in the Appendix. In this study, we use the status groups listed in Table 3.

Table 3. Status Groups

STATUS GROUP	ASSOCIATED DODTCPG CODES	DESCRIPTION
IADT_Proj	UF, UP	Undergoing IADT
SMCR_Proj	SA, UQ, UX	Drilling Marines
Total_Proj	SA, UQ, UX, UF, UP	A superset of all Marines

2. Explanatory Variables

Each explanatory variable provides information on a Marine at a given time. We consider all available data in the MCRISS and TFDW for suitability as explanatory variables, except for equal opportunity-related data fields. As stated in the goals of this study, part of the output of this algorithm is the probability of an individual's future status (i.e., their ability to succeed or continue) with the intent that action might be taken to alter the outcome. However, actions taken based on equal opportunity-protected categories would violate regulation and practice. Omitting EO category-related predictors ensures that the algorithm will not violate these norms.

In order to decrease the correlation between date variables, we subtract each data variable from a selected base, typically the data of the projection or the date of entry. We also add construct the fiscal year as an input to allow the algorithm to distinguish time-based relationships. The constructed variables are listed in Table 4.

Table 4. Constructed Variables

CONSTRUCTED VARIABLE	DESCRIPTION
Obligation_Remaining	# months remaining in a Marines service obligation
EAS_Remaining	# months remaining in a Marines current training status
TIS_Now	# months since entry into SMCR
FY	The fiscal year of the current SEQ
Month	The calendar month of the current SEQ

C. ALGORITHM DEVELOPMENT

1. System Characterization

The SMCR is a system in which Marines enter, transition through multiple states, and then exit the system at a future time. Two previous studies (Erhardt, 2012; Dausman, 2016) have applied Markov processes to modeling this system with some success. Ultimately, the approach was of limited effectiveness because the Markov transitions were of limited complexity (e.g., using only a few explanatory variables) and because the Markov transitional probabilities are not constant with respect to time. In order to improve

the process, a model must allow variable complexity of status transitions to ensure that it considers all relevant information without overfitting the less-complex transitions, particularly as those relationships change over longer projection horizons.

We model each Marine's outcome as a Bernoulli random variable with an unknown probability of success. A Bernoulli random variable is similar to the outcome of a simple coin flip in which "heads" signifies success, but in this case, each Marine has a different probability of success. If a subgroup of Marines has similar probabilities, we describe the number of successes among them as a random variable with an approximately binomial distribution. In this way, we can estimate the expected value of each individual in a group, even though each individual only has one outcome to observe. Once we estimate the binomial parameter for each group, we can estimate the aggregate outcome as the sum of the expected values of all binomial parameters.

A variety of modeling techniques exist that could be applied to this study. Most explanatory models use regression to describe variable relationships, but modeling requires parametric relationships between input and output that may not apply to our study data. Markov modeling techniques have been used in previous SMCR modeling studies, but they are limited by the need for relationships to remain constant over time. We also consider neural networks or "deep learning" techniques to be too computationally demanding for our purposes. Of the remaining machine learning techniques, decision trees and their extension, random forests, exhibit the most promise for this application. Exploratory testing between random forests and decision trees yields minor differences between them in this application. Because decision trees also allow much finer control of the complexity, provide more interpretability, and require significantly less computation, we select decision trees for implementation of our algorithm. For additional information of the other techniques we mention, we refer the reader to Faraway (2016).

The idea of decision trees is to group similar cases with similar attributes and outcomes based on the values of their explanatory variables together into a decision node called a "leaf." Some researchers have used the number of intervening branches in a tree as a measurement of distance between cases (Buttrey & Whitaker, 2016). Under that definition of distance, cases sharing the same leaf have a zero distance and are the most

similar to each other. Once estimated, the parameters for splitting the cases into different leaves provide a path to assign future cases to a leaf, with the leaf providing an estimated probability of success for the new case based on the average of the training data.

2. Mathematical Relationships

Rather than directly estimating the population level, we first categorize and estimate using a decision tree to induce grouping of the Marines into a binomial variable. The proportion of each type of outcome on a leaf provides the estimated probability of each Marine remaining in the population. We then sum the expected values to estimate the Marine population remaining at a future time. By characterizing the remaining Marine population as a sum of Bernoulli random variables, we can also use the known variance of Bernoulli random variables to produce prediction intervals that adjust to changing input distributions.

Specifically, we index Marines in a cohort as $m = 1 \ldots M$. We define Y_m to be equal to one if Marine m is in the projected status group, and zero otherwise. Because Y_m is unobservable due to it occurring in the future, we describe it as a Bernoulli random variable with parameter $P_m = P(Y_m = 1)$. The following facts are immediate:

$$E(Y_m) = P_m \tag{1}$$

$$Var(Y_m) = P_m(1 - P_m) \tag{2}$$

Our goal is to predict T, the number of known Marines in the population at a future point, which we define as

$$T = \sum_{m=1}^{M} Y_m \tag{3}$$

Substituting, we find

$$E(T) = \sum_{m=1}^{M} P_m, \quad Var(T) = \sum_{m=1}^{M} P_m(1 - P_m) \tag{4}$$

We now apply a decision tree to estimate P_m. The current cohort of Marines is assigned to leaves $1 \ldots L$ that are mutually exclusive and exhaustive. For Marine m, we

estimate P_m as the proportion of successes for Marines classified in the same leaf during model training. Let n_l denote total number of Marines classified in leaf l and let $l(m)$ denote the leaf to which Marine m is assigned.. Let $p_1 \ldots p_L$ denote the probabilities that Marines assigned to their respective leaves will continue in the population. Using this notation, we now have

$$\mu_T = E(T) = \sum_{l=1}^{L} n_l p_l, \quad \sigma_T^2 = Var(T) = \sum_{l=1}^{L} n_l p_l (1 - p_l) \tag{5}$$

Implicit in Equation (5) is the assumption that the decision tree represents an accurate partitioning of the Marines into classes of individual that share a common probability of succession. We acknowledge that this assumption is not perfectly true, but assume that it provides a sufficiently close approximation to reality. We then would obtain a 95% prediction interval for T by taking $\mu_T \pm 1.96\sigma_T$ except for the fact that the expressions in Equation (5) depend on parameters $p_1 \ldots p_L$ that must be estimated from data. Our approach to estimation is to use a learning data set to formulate the decision tree and to estimate parameters; and a test data set that is independent of the learning data to form population forecasts.

For each leaf, the estimated proportion of successes is subject to sampling error, which increases as we split each leaf into smaller child leaves. This sampling error is present both when the model is trained ("Learn") and when the algorithm projects a cohort forward ("Test"). The sampling error causes the estimated mean value of the leaf to have sampling variability. The impact of this variability depends in the variance of the leaf as well as the number of samples on the leaf during both training ($n_{l,Learn}$) and forecasting ($n_{l,Test}$). We define $\hat{T}$ as an estimator for $E(T)$ obtained by replacing true probabilities p_l with estimates $\hat{p}_l$ in Equation (5) obtained from the training data. The variance of $\hat{T}$ can be expressed as

$$Var(\hat{T}) = \sum_{l=1}^{L} \frac{n_{l,Test}^2}{n_{l,Learn}} p_l (1 - p_l) \tag{6}$$

We acknowledge that the splitting rules used to define leaves on the training data are also subject to sampling error. Given the large sample sizes involved in this study and

the measures taken to avoid overfitting, we assume that the impact of this error is minor and we therefore do not attempt to account for its effect on variance estimates.

The prediction variance is the sum of the estimation variance induced by $\hat{T}$ and the innovation variance induced by a new realization of T:

$$\sigma^2_{\text{Pred},\hat{T}} = \sum_{l=1}^{L} n_{l,Test} p_l (1 - p_l) + \sum_{l=1}^{L} \frac{n^2_{l,Test}}{n_{l,Learn}} p_l (1 - p_l) \tag{7}$$

Upon substituting estimates for the unknown probabilities from the learning data we obtain the estimated prediction variance:

$$\hat{\sigma}^2_{\text{Pred},\hat{T}} = \sum_{l=1}^{L} n_{l,Test} \hat{p}_l (1 - \hat{p}_l) + \sum_{l=1}^{L} \frac{n^2_{l,Test}}{n_{l,Learn}} \hat{p}_l (1 - \hat{p}_l) \tag{8}$$

Finally, we calculate a $100(1-\alpha)\%$ prediction interval (PI) for T using

$$PI = \hat{T} \pm z_{\alpha/2} \cdot \hat{\sigma}_{\text{Pred},\hat{T}} \tag{9}$$

where $z_{\alpha/2}$ is a standard normal quantile for the desired prediction interval. For example, a 95% prediction interval uses $\alpha = .05$ and $z_{.025} = 1.96$.

The estimated prediction variance $\hat{\sigma}^2_{\text{Pred},\hat{T}}$ is known to be biased due to model-induced bias and the non-linearity of the binomial variance estimator. The magnitude of the model-induced bias is dependent on the level of fit of the component leaves of the decision tree. In a tree, each leaf consists of a sum of one or more binomial random variables. An underfit decision tree could result in each leaf representing multiple binomial distributions. Grouping distinct binomial distributions into one leaf would result in $\hat{\sigma}^2_{\text{Pred},\hat{T}}$ being positively biased, as illustrated in Table 5.

Table 5. Effect of Tree Fit on the Bias of Variance Estimates of a Binomial Distribution

ORIGINAL DISTRIBUTION	UNDERFIT ON SAME LEAF	OVERFIT ON MULTIPLE LEAVES
Binomial 1 $p = 0.2$ $\bar{\sigma}^2 = 0.16$	$\hat{p} = 0.3$ $\hat{\bar{\sigma}}^2 = 0.21$	$\hat{p} = 0.19$ $\hat{\bar{\sigma}}^2 = 0.1539$
		$\hat{p} = 0.21$ $\hat{\bar{\sigma}}^2 = 0.1659$
Binomial 2 $p = 0.4$ $\bar{\sigma}^2 = 0.24$		$\hat{p} = 0.39$ $\hat{\bar{\sigma}}^2 = 0.2379$
		$\hat{p} = 0.41$ $\hat{\bar{\sigma}}^2 = 0.2419$
Actual Total $p = 0.3$ $\bar{\sigma}^2 = 0.20$	Estimated Total $\hat{p} = 0.3$ $\hat{\bar{\sigma}}^2 = 0.21$	Estimated Total $\hat{p} = 0.3$ $\hat{\bar{\sigma}}^2 = 0.1999$

As Table 5 demonstrates, the variability estimate of underfit trees is positively biased. Conversely, overfit trees will underestimate variance. These effects occur even without consideration of sampling error.

In practical application, the situation is more complex. A decision tree is composed of combinations of underfit and overfit leaves. This makes estimation of confidence intervals from classification algorithm outputs challenging. The calculated $\hat{\sigma}^2_{\text{Pred},\hat{T}}$ will bias differently for all levels of model fit or even for different input distributions acting on identical model fits. Our approach is to measure the level of bias for each level of fit or *complexity* and use it as a guide for selecting the proper complexity for each ensemble of projection horizons.

Given our large sample sizes and that the process results in a sum, we can invoke the Central Limit Theorem (Ross, 2006) to assume approximate normality of errors. Therefore, we can express our model as

$$T = g(x) + \varepsilon \tag{10}$$

where $g(x)$ is the model prediction, and ε is the prediction error. We can interpret ε as the randomness remaining after the prediction, i.e., as a residual term. In this context, the

variance of ε, which we denote as σ_{ε}^2, is equivalent to $\sigma_{\text{Pred},\hat{T}}^2$. Given that we can estimate σ_{ε}^2 directly during training, it is commonly used as a replacement for $\sigma_{\text{Pred},\hat{T}}^2$. However, σ_{ε}^2 is only equivalent to $\sigma_{\text{Pred},\hat{T}}^2$ for sample distributions that are similar to the training set. In order to make use of our training estimate with other sample distributions, we a cross-validated estimate of the variance of ε, which we denote $\hat{\sigma}_{\varepsilon}^2$ as an unbiased target to compare to the calculated $\sigma_{\text{Pred},\hat{T}}^2$. The ratio of these two estimates, which we term the Variance Ratio (VR), performs several functions. During training, we can use the VR to judge the overall level of fit of the model using only the training data. During prediction, we can use the VR as a multiplier for our calculated $\sigma_{\text{Pred},\hat{T}}^2$ to calibrate the result to the proper range.

First, we calculate $\hat{\sigma}_{\varepsilon}^2$ using a variation of jackknife resampling. (Tukey, 1958) Jackknife resampling produces a calculation of an estimated model *k*-times, each formed by excluding a random sample of the data (a *fold*) that encompasses a fraction of about 1/*k* of the sample. Model fit is estimated by evaluating the estimated model on the held-out fold using squared error as a criterion. The average of all squared errors (across all folds) is taken as an overall estimate of goodness-of-fit. In our variation, we fold the training data in a deterministic manner, along the time axis (SEQ). In addition, instead of excluding the fold in each iteration, we calculate our parameter for the fold only. In this way, we can measure the standard error of the modelled count over time and calculate the underlying variance of the model error, including variance induced by autocorrelation and distributional shifts. We set the number of folds, *k*, equal to the number of unique SEQ values. Ultimately, our algorithm provides a prediction for specific points in time, T_f, using a set of input predictors, x_f, that are processed by a model (*g*). For each fold, from 1 to *k*;

$$\varepsilon_f = T_f - g(x_f), \quad f = 1, \ldots, F \tag{11}$$

or simply, that the error of the model is the actual population size minus the predicted population size at time *t*. With ε_f and its sample mean ($\overline{\varepsilon}$), we now obtain the estimated prediction variance $\hat{\sigma}_\varepsilon^2$ as follows:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{f=1}^{F}(\varepsilon_f - \overline{\varepsilon})^2}{F} \tag{12}$$

We express the Variance Ratio for the training set, derived from the prediction variance and the prediction standard error ($SE_{T_{Learn}}^2$) obtained via Equation (8), as

$$VR_{T_{Learn}} = \frac{\hat{\sigma}_\varepsilon^2}{SE_{T_{Learn}}^2} \tag{13}$$

In the final step, we multiply the estimated standard errors of the test set by $VR_{T_{Learn}}$ and a single scalar, β, used across all models, which is a manually estimated value that adjusts for the divergence between learning and test set variance.

$$\begin{aligned} \hat{\sigma}_{\varepsilon,Test}^2 &= SE_{T,Test}^2 VR_{T_{Learn}} \cdot \beta \\ &= \frac{SE_{T,Test}^2}{SE_{T,Learn}^2}\hat{\sigma}_\varepsilon^2 \cdot \beta \end{aligned} \tag{14}$$

A more detailed discussion on estimation of β is given in the following section.

3. Complexity Tuning

For complexity tuning, we separate the data into test and training sets, each comprised of the records of half the available Marines. Sensitivity testing demonstrates that the value of $\hat{\sigma}_\varepsilon^2$ of the training and the validation sets diverge as the Variance Ratio decreases toward one. The estimation error (and the error variances) on test and validation sets diverges as a model approaches optimal fit with the learning errors being lower than the test errors. This is a common feature of classification algorithms. In general practice, the divergence is ignored and the algorithm complexity tuned to minimize validation error. However, a key objective of this study is to obtain accurate prediction intervals, we must balance model accuracy with the quality of the variance estimate. The projections of the individual Marine model outputs are secondary to the projection of the aggregate total and

the estimate of its prediction intervals. Therefore, we utilize model complexities that have lower than the maximum achievable prediction accuracy, but provide small and predictable training/test variance divergence characteristics so that a prediction interval can be more reliably estimated.

Figure 2 provides an illustration of this phenomenon. In our models, we control model complexity with the MaxLeafs parameter, which determines how much the decision tree grows. As the number of tree nodes increases (horizontal axis), the validation error (black) tends to decrease. However, the difference in VR between training (red) and test (blue) sets generally increases with complexity, with training VR falling below the test VR.

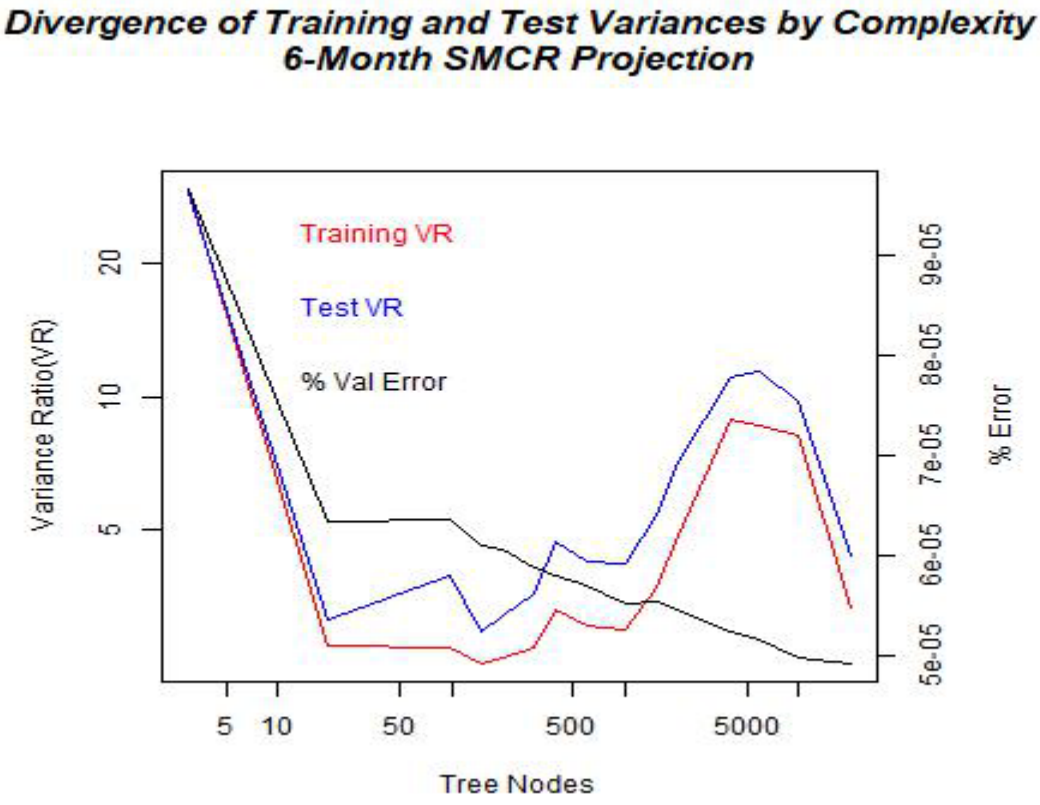

Figure 2. Example of the Divergence of Variance between Training and Test Sets

For each status group, the MaxLeafs parameter and a single variance scalar, β, are manually estimated as tuning parameters. Additionally, we limit the minimum samples

required in each leaf during training to 100 for IADT and 200 for all other status groups to ensure that $\hat{\sigma}^2_{\text{Pred},\hat{f}}$ remains small for likely forecast distributions.

D. METHODOLOGY

1. Import and Initial Formatting

The MCRISS and TFDW datasets consist of monthly snapshots of each their respective databases, each contained in separate CSV files. Each snapshot contains the SEQ number corresponding to the month the snapshot originated. We first replace each EDIPI in each file with a numeric key. For each dataset, we begin data processing by importing each snapshot file and merging them. We merge the two datasets by matching the EDIPI and the SEQ/RSEQ listed in each record, producing a single table.

2. Feature Selection

All date fields in the initial dataset are structurally related to each other. Older records tend to have older dates for every field. For the model to be effectively extract information from the dates, we must offset each date from a basis field, as we defined in Table 4. Mathematically, this aligns each record to others across time.

We screen explanatory variables in order to increase the computational efficiency of the model and minimize overfitting. We eliminate each explanatory variable in a step-backwards procedure if its relative significance is below 0.02 relative to all surviving explanatory variables. While some of the eliminated variables may have correlation with the outcome, many of them are collinear with each other and redundant when used together. Table 6 lists the explanatory variables used in the final model.

Table 6. Final Explanatory Variables

EXPLANATORY VARIABLE	ORIGIN	DESCRIPTION
DODTCPG	TFDW	Current service status
RCOMPCODE	TFDW	Sub-statuses to DODTCPG
Obligation_Remaining	Constructed	# months remaining in a Marines service obligation
EAS_Remaining	Constructed	# months remaining in a Marines current training status
TIS_Now	Constructed	# months since entry into SMCR
FY	Constructed	The fiscal year of the current SEQ

We group the DODTCPG field per sponsor specification and create a Boolean (TRUE/FALSE) field for each status group we defined in Section III.B.1 Table 3. We extract the Status Groups and its SEQ to use as predicted variables, but we retain the original DODTCPG in the explanatory variable vector. Finally, we convert all data to a numeric format (with missing values set to -9999) for compatibility with the Python SciKit-Learn package (Pedregosa et al., 2018).

3. Model Training

We split the input dataset by individual Marines into approximately equally sized training and validation datasets. In a production environment, the algorithm would use all data for training. Model training begins by recombining our explanatory variable data, representing all data available on a Marine at time τ and the predicted data, representing the status group of the Marine at a *future* time $\tau+\eta$, where η is the number of months in the future the model is trained to predict. To achieve this, we offset the Status Groups by η SEQ numbers. We then provide the respective predicted and explanatory variable fields to the SciKit-Learn Decision Tree function for training.

Upon completion of training, we fold the training dataset by the SEQ field to create a time-series of inputs and outcomes. For each fold, we use the newly-create model to predict results, errors, and subsequently calculate VR, according to Equation (13). We then store each model and the model's VR for future use. We repeat the entire process for every desired prediction horizon and status group.

4. Forecast Computation

We obtain the estimated probability for each forecast Marine from a stored model and sum the expected values, as explained in Equation (4). By repeating the process with individual models trained for each projection horizon, we can construct a forecasted time series of population levels.

We calculate $\hat{\sigma}^2_{\text{Pred},\hat{T}}$ using Equation (8). However, as we indicate in Section III.C.2, there is divergence between the predictions of the training set and any forecast set. We correct the divergence and generate $\hat{\sigma}^2_{\varepsilon}$ by using the VR parameter recorded for each model, as applied in Equation (13). We also apply the global variance scalar, β. Finally, we use $\hat{\sigma}^2_{\varepsilon}$ and the appropriate normal quantile to calculate the desired prediction interval.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS

In order to consider the algorithm results, we must look at three parts. First, we examine the IADT status group models, then the SMCR status group. Finally, we demonstrate the general impact of the Variance Ratio process. For testing purposes, we retrain the algorithm using all available data from FY06 through FY14 and catalogue the training errors. We then conduct hindcast projections for each SEQ from FY15 through FY17 and document the results.

A. IADT STATUS GROUP

We train the IADT status group models with a MaxLeafs parameter of 100. The ensemble of IADT models projects up to 12 months into the future. First, we will examine the variable relationships at select points in the ensemble, and then we analyze the output errors of the ensemble.

1. Explanatory and Predicted Variable—Relationships

According to the sklearn decision-tree algorithm documentation, the importance of an explanatory variable "is computed as the (normalized) total reduction of the criterion (Gini purity) brought by that feature" (Pedregosa et al., 2018). In our algorithm, we use Gini purity as our criterion. For each model, sklearn provides a function to extract the relative importance or the marginal fraction of Gini purity each variable provides. Figure 3 shows the importance measurements for the IADT ensembles.

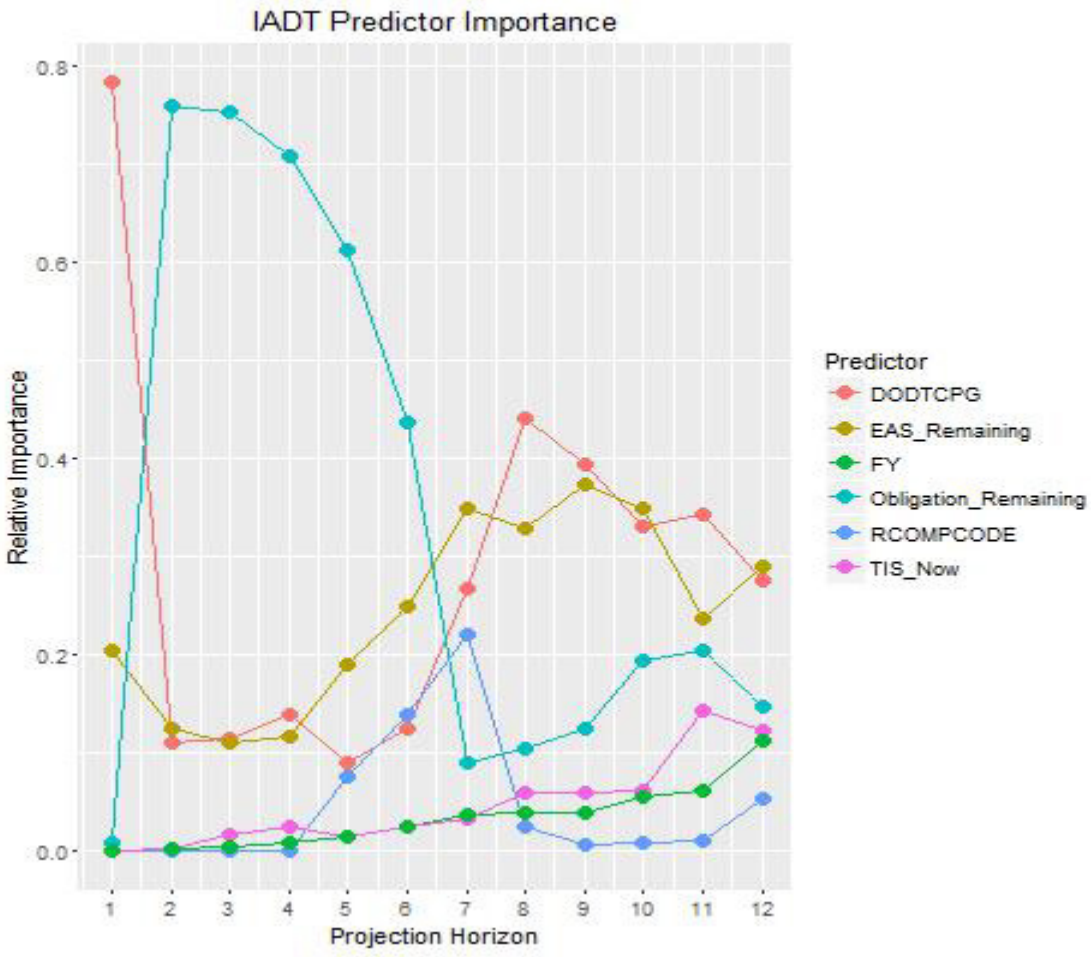

Figure 3. IADT Explanatory Variable Relative Importance

Explanatory variable significance varies substantially across the 13 prediction horizons of the IADT ensemble. While most of the prediction horizons smoothly transition into each other, there are several points of apparent discontinuity where the explanatory variable significance shifts abruptly. The remainder of this section illustrates the one, three, and seven-month models, as those are the points that are most dissimilar to each other in explanatory variable's significance.

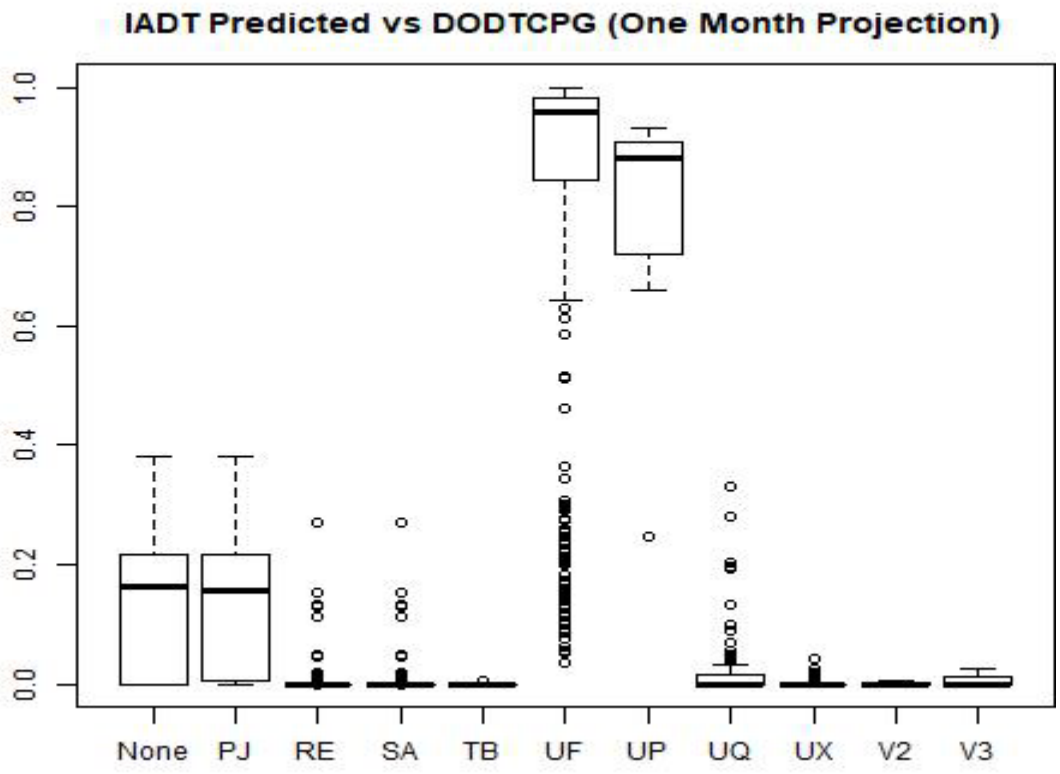

Figure 4. IADT Predictions versus DODTCPG (One-Month Projection)

On one-month projection horizons, DODTCPG is the dominant explanatory variable, owing to the high autocorrelation of a Marine's status. Its significance drops abruptly by the two-month horizon. While DODTCPG is still highly significant, the autocorrelation for IADT statuses is less important at that point, particularly in the presence of Obligation Remaining as both DODTCPG and Obligation Remaining are blank for pre-IADT Marines, so they encode much the same information. However, the lack of significance for Obligation Remaining in the one-month prediction indicates that once a Marine has begun IADT, the short term accuracy of the field is lower that the initial autocorrelation of the status.

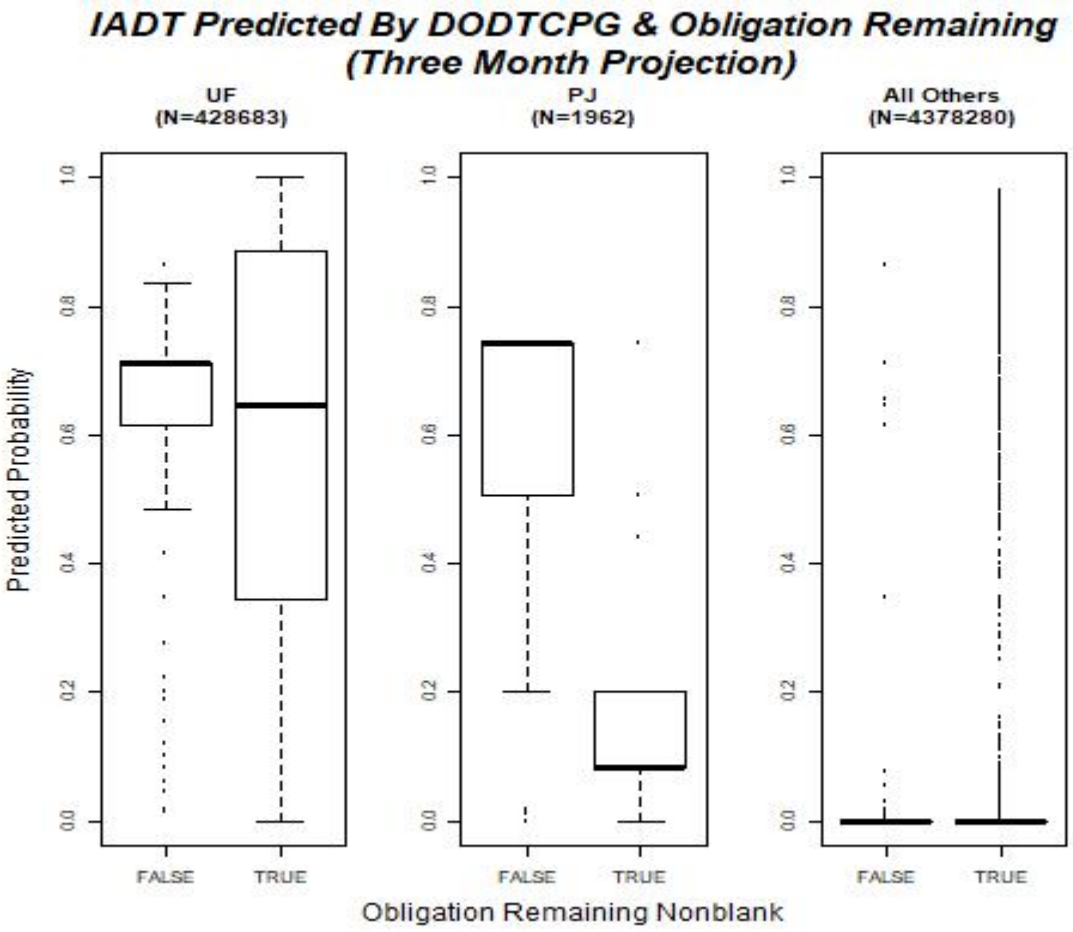

Figure 5. IADT Predicted by DODTCPG and Obligation Remaining (Three-Month Projection)

By the three-month projection, DODTCPG has far less effect on the output when used alone. The most significant indicator of outcome is whether Obligation Remaining is blank, but only if the Marine is in a "PJ" (Individual Ready Reserve) or blank (recruiting) DODTCPG. If a "UF" status (IADT) is used, the predicted probability is largely insensitive to the presence of Obligation Remaining, but the model predicts a wider range of values if Obligation Remaining is present, indicating the marginal effect of less significant explanatory variables.

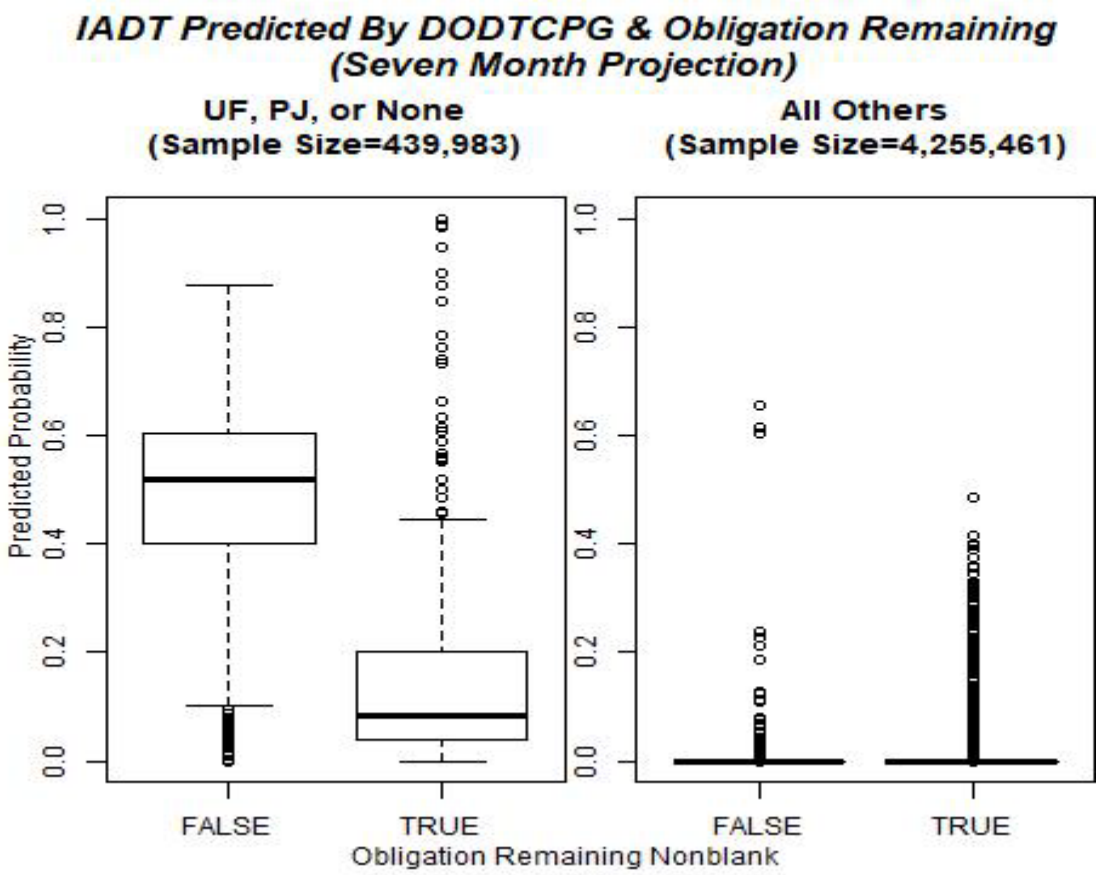

Figure 6. IADT Predicted by DODTCPG and Obligation Remaining

By the seven-month projection, "UF" and "PJ" statuses have merged into similar distributions, with non-blank obligations indicating reduced probabilities. The range of probabilities for blank obligations has also increased indicating interactions with other explanatory variables become significant in this case.

The significance of Fiscal Year (FY) peaks in the seven-month model. As Figure 7 illustrates, FY interacts with the "UQ" and "UF" statuses but is not statistically significant. Despite this, we choose to retain FY as an explanatory variable to ensure the models can adapt to future states in a production environment. If an explanatory variable relationship changes in the future, when retraining occurs the presence of FY allows the models to distinguish the break in relationships as soon as it is statistically significant and incorporate

it into the updated model. In the current dataset, this has minimal impact on the predicted values, but does have a noticeable impact on forecast uncertainty.

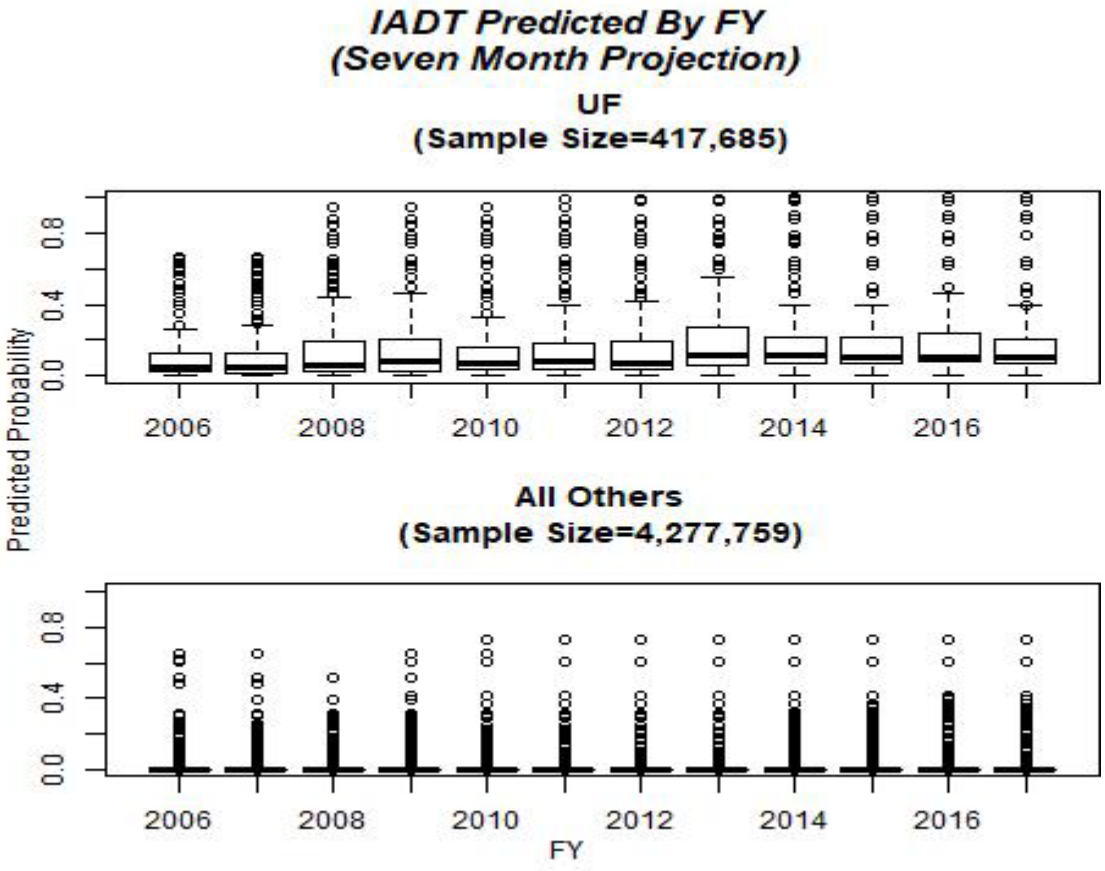

Figure 7. IADT Predicted by Fiscal Year (FY) (Seven-Month Projection)

2. Output Analysis

Because we determine the estimate of probability of leaves during training, this also fixes its estimate of variance. The range of prediction interval values estimated in any particular model corresponds to the distribution of leaves the data occupies. Because of this, we can infer that the distribution of leaves used in hindcast mode is smaller than that used in training because the range of estimated prediction intervals is narrower. We find that the primary interaction driving the range of prediction intervals in training is the FY explanatory variable. Although it does not interact strongly, when it is used, it does

influence the variability of the leaf, either positively or negatively. As discussed in the previous section, FY alters the model by creating breakpoints for changes over time. For the hindcast of FY 14–17, the model only uses the set of leaves that include the most recently trained FY. This has the advantage of incorporating temporal changes as soon as they are significant. A disadvantage is that FY-based leaves will have smaller sample sizes and higher sampling errors.

The impact of the phenomenon described in the previous paragraph is evident in Figure 8. The range of hindcast prediction widths is significantly narrower than in the training set. Most often, they follow the mean of the training set. This indicates that FY13 is using a similar distribution of leaves as the bulk of the training set—leaves that either have multiple fiscal years included to minimize sampling errors or are FY insensitive.

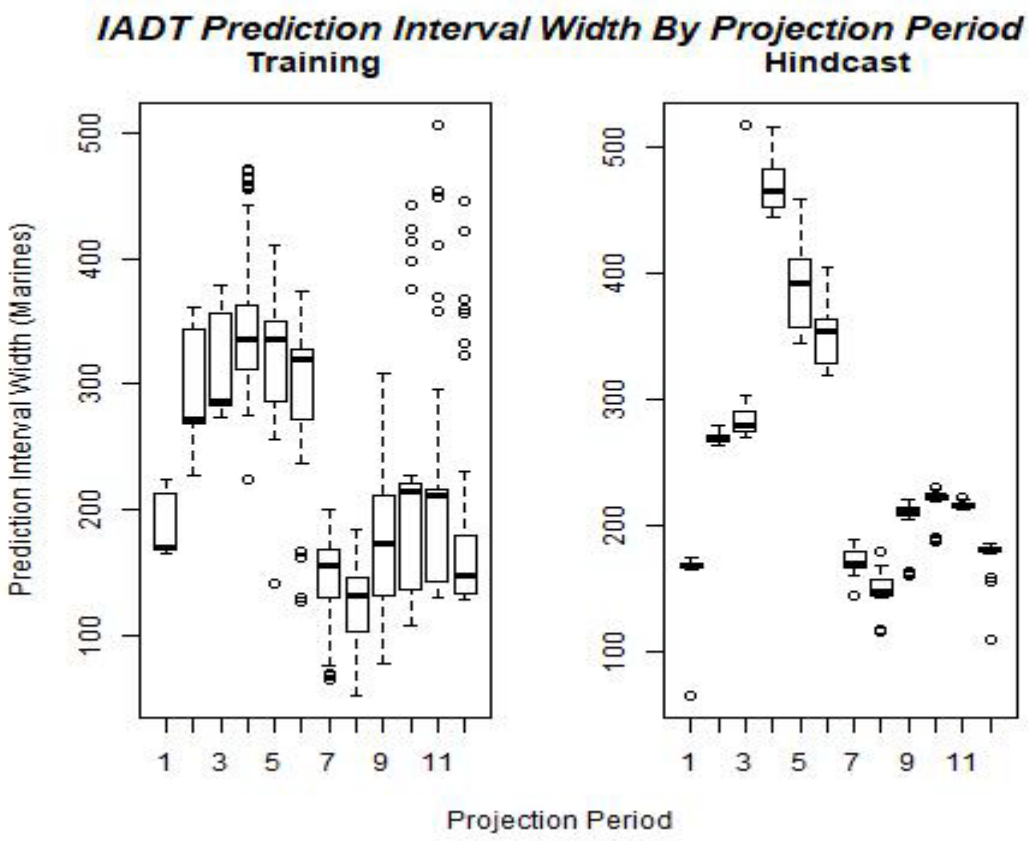

Figure 8. IADT Prediction Interval Width by Projection Period

There are several notable exceptions. Projection models 4–6 use leaves that are in the top quantile or are even outliers in variance compared to the training set. Only two explanations are possible. Either the FY breakpoints occur very recently, likely including only the most recent FYs or the hindcast period has a very different distribution of data than the training period.

We can see some indications of these impacts in Figure 9. All errors are scaled by the estimated prediction interval of the forecast. The yellow line indicates the model-estimated 95% prediction interval while the red line represents the β adjustment across all models. While the errors fall into a normal range for most of the training and hindcast errors, we can see higher error variance in the 18 month period from FY06-07 and both bias and error variance in the most recent 12 months (FY17). These outliers occur despite the larger estimated prediction intervals in projection models 4–6, which would reduce this kind of error. Given the prediction intervals calculation already accounts for the lower sampling error, we must conclude the FY17 data includes significant shifts in population or administrative processes that our algorithm is unable to anticipate.

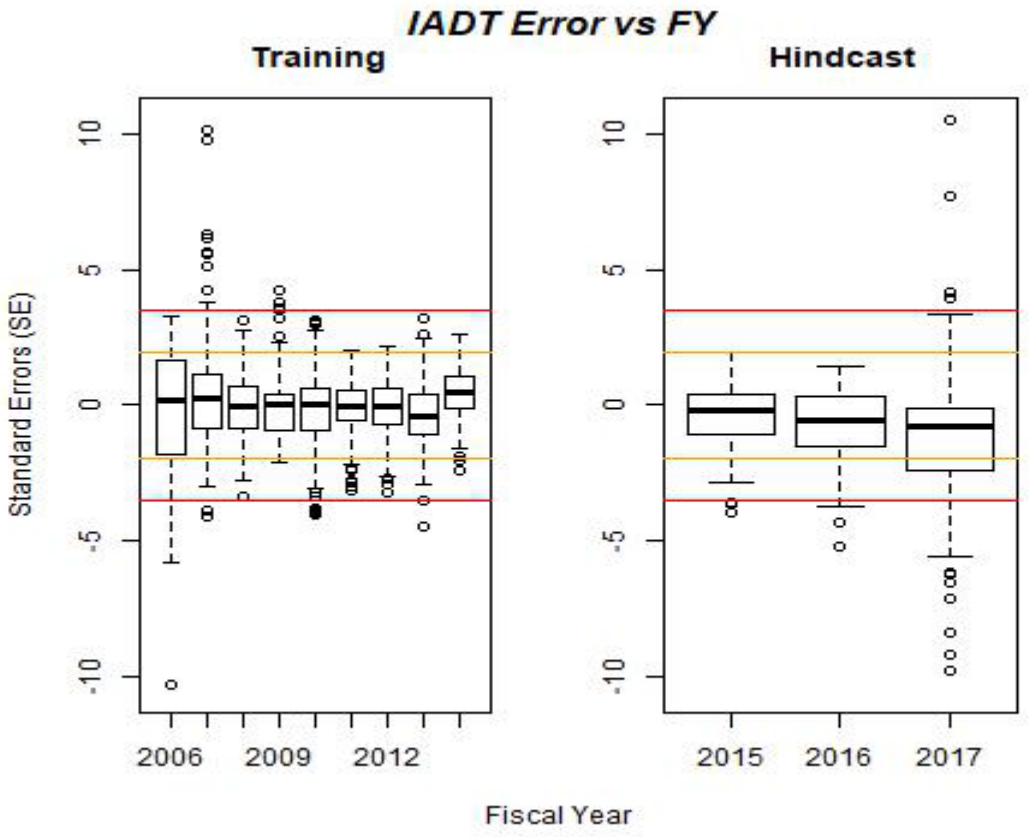

Figure 9. Training and Hindcast Errors by FY

This type of occurrence is exactly the reason we retained FY as an explanatory variable. While the algorithm could never anticipate future changes in underlying data relationships, frequent retraining allows the algorithm to incorporate any autocorrelated errors into future predictions. In the test set illustrated in Figure 14, the model performs within β-adjusted limits for almost to 24 months with no retraining before experiencing consistent deviations. Indications of bias did begin to show during that period; however, retraining after FY 15 or even FY 16 would likely have muted or eliminated the error trend and kept FY 17 within tolerance. We recommend retraining the model at the end of each FY. While more frequent training may detect trends earlier, it may also introduce spurious sampling errors (and associated increase in prediction intervals widths) if the data for the current FY is incomplete.

Despite the distributional drift inherent in the hindcast period, the algorithm tends to produce low biases across the models overall. Figure 10 demonstrates some of the

algorithms shortcomings. Without the β adjustment, prediction intervals are too small for shorter projection periods and too large for longer-term models. While the β adjustment improves the fit for the shorter period, it exacerbates it for the longer periods.

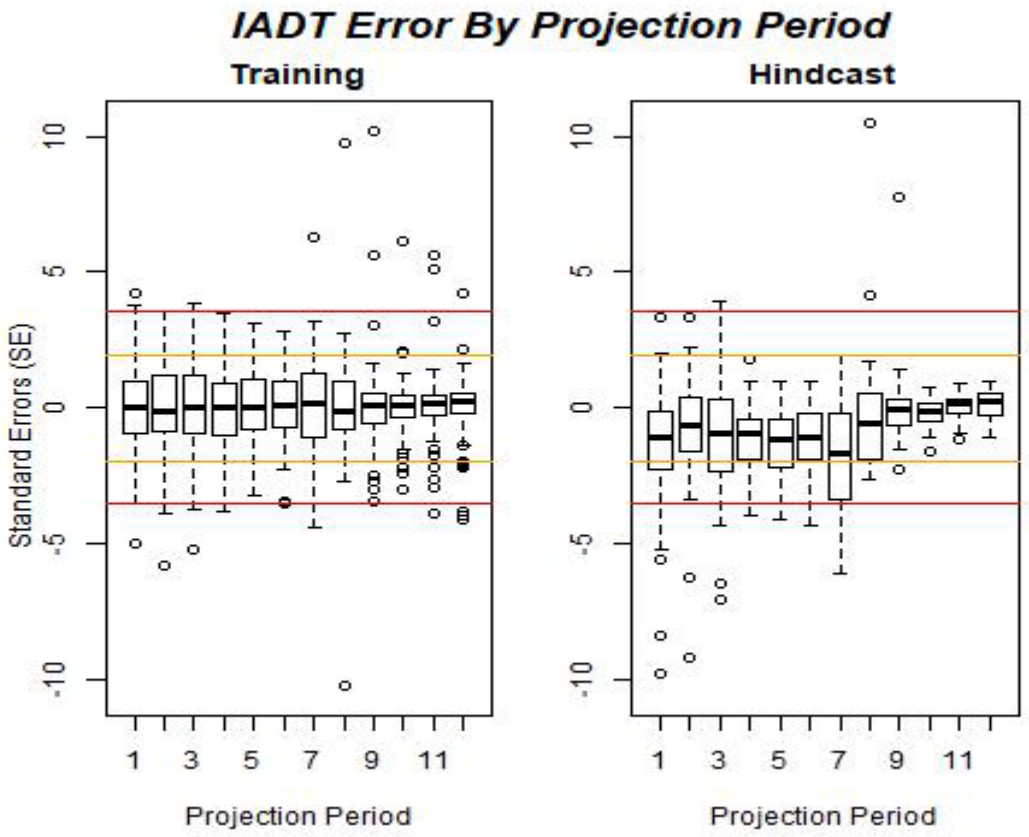

Figure 10. IADT Error by Projection Period

Because our algorithm only forecasts the specific cohort provided as input, the cohort population declines as it is projected forward and attrition takes its toll. This is especially noticeable in the IADT status group as it is rare for Marines to remain in IADT for more than 6 months. It is also consistent with the effects of model fit across the range of projection models having markedly different underlying variances. We choose a single complexity for all projection models. Because the population level changes markedly over the range of the ensemble, it is impossible to select an optimal complexity parameter for the overall ensemble. Using the median of ideal model complexities tends to underfit on

shorter periods and overfit on longer periods. As explained in Section III.C.2, this results in biased estimates of variance from each model. Further research could substantially improve the prediction interval estimates of this algorithm by developing fitting complexity parameters for each projection period or by a training algorithm that is capable of maintaining the correct variance in each leaf individually.

Figure 11 illustrates the results of the algorithm projected from the beginning, middle, and end of FY 15. The points and error bars represent the hindcasted value and 95% prediction interval. For comparison, the line represents the actual values for that period. The initial hindcast from SEQ 319 (SEP 2014) falls within the 95% prediction intervals throughout the entire period, though the 3–6 month periods appear to be positively biased. The SEQ 325 (JUN 2015) hindcast appears more accurate, with only periods 2 and 3 deviating noticeably from the median prediction. The SEQ 331 (SEP 2015) hindcast includes is well outside the prediction intervals for periods 4 and 5 and is negatively biased for most of the forecast. While significant in this example, such errors would likely be reduced if the algorithm were retrained prior to every FY.

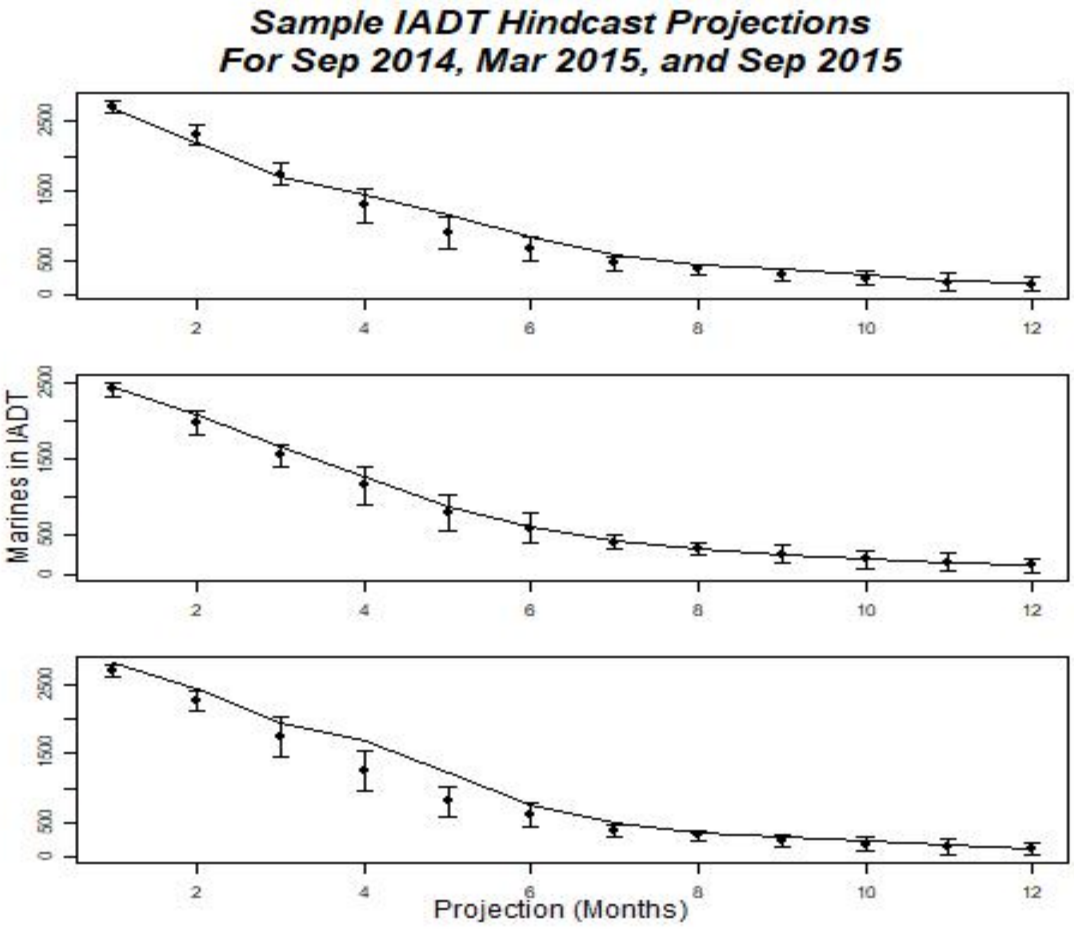

Figure 11. Sample IADT Hindcast Projections

As Figure 12 illustrates, seasonality in the training set is largely under control and well within adjusted prediction intervals. The hindcast does show the same seasonality pattern indicating that the variations seen in the training period are not random. While the projections are mostly within the adjusted prediction intervals, a seasonal bias is apparent in both training and hindcast. While the prediction intervals largely encompass it, additional countermeasures against seasonality in IADT errors would result smaller prediction intervals and a more certain forecast. In the absence of another explanatory variable, the obvious solution is to use calendar month as an explanatory variable. We consider it unwise to use both FY and Month as explanatory variables simultaneously. The use of both may result in "memorization" (i.e., overfitting) of the specific month-years with the highest deviations rather than a simple adjustment of the calendar month. We

recommend a mean adjustment for seasonality of errors prior to variance estimation, but we did not implement such an approach in this algorithm.

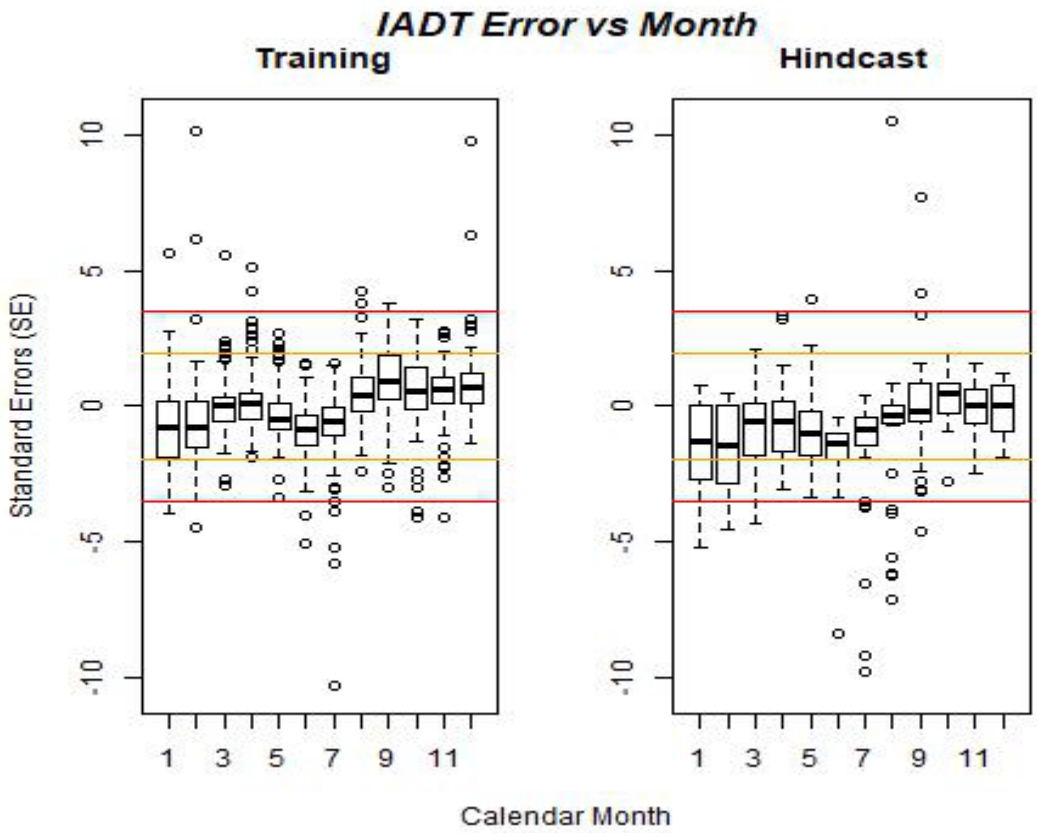

Figure 12. IADT Errors versus Calendar Month

B. SMCR STATUS GROUP

We train the IADT status group models with a MaxLeafs parameter of 200. The ensemble of IADT models projects up to 12 months into the future. First, we will examine the variable relationships at select points in the ensemble, then we analyze the output errors of the ensemble.

1. Explanatory and Predicted Variable—Relationships

The importance measurements for the IADT ensemble are shown in Figure 13.

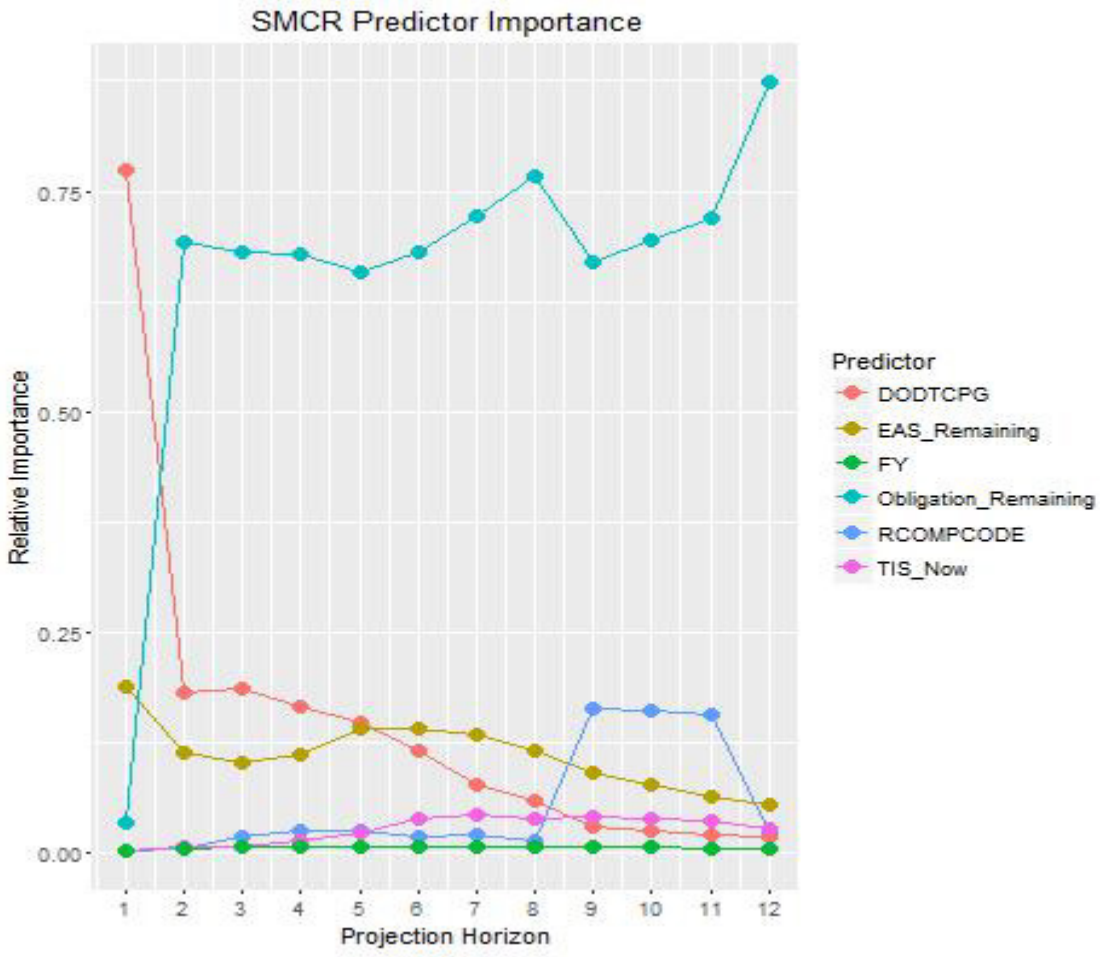

Figure 13. SMCR Explanatory Variable Relative Significance

Explanatory variable significance varies substantially across the 12 prediction horizons of the SMCR model set. While most of the prediction horizons smoothly transition into each other, there are several points of apparent discontinuity where the explanatory variable significance shifts abruptly. For the remainder of this section, we illustrate the one-, three-, and seven-month models as those are most dissimilar to each other in explanatory variable significance.

In the one-month SMCR model, DODTCPG has an 82% variable importance. The second-most important explanatory variable is EAS Remaining. The relationship between EAS Remaining and SMCR status group is an almost perfect complement of the relationship between EAS Remaining and the IADT status group. By design, it maps the

phase of service that each Marine goes through and is a good explanatory variable for statuses that occur while in the Marine Corps Reserve.

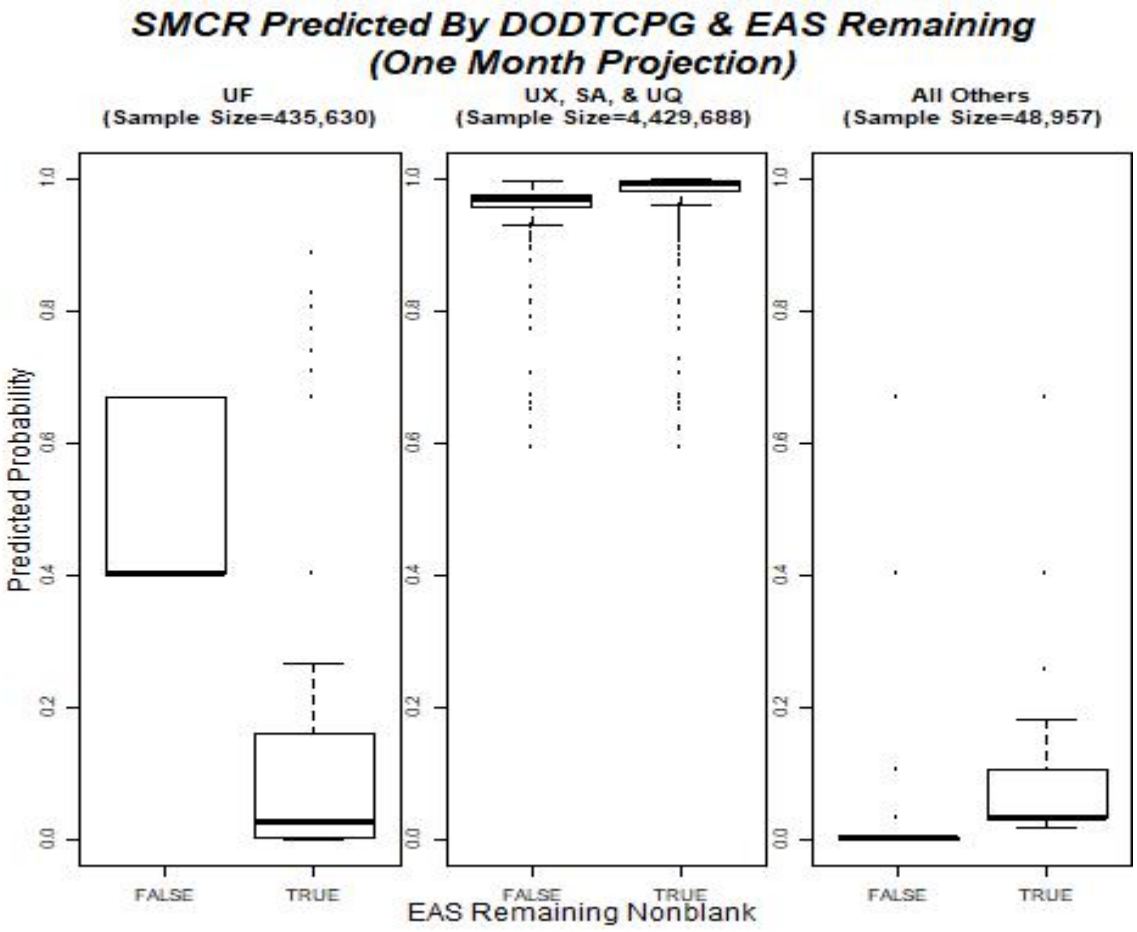

Figure 14. SMCR Predicted by DODTCPG & EAS Remaining (One-Month Projection)

The eight-month SMCR model exhibits similar relationship with the exception of RCOMPCODE. Figure 15 provides the results of output analysis of this explanatory variable. All Obligation Remaining values are insignificant with the exception of RCOMPCODEs "KA" and "K4," which have significantly higher residual mean and range than the other codes. A "KA" RCOMPCODE indicates that a Marine has completed his or her initial contract obligation but remains in the SMCR while a "K4" is a post-IADT Marine serving his or her first contract term. The first term contract obligation is not

necessarily synonymous with the Obligation Remaining field as a Marine may incur other obligations for service by exercising certain benefits such as tuition assistance. On the other hand, the negative values of Obligation for a "K4" RCOMPCODE are logically invalid, but even these administrative errors seem to convey predictive information. While Obligation Remaining is the most significant explanatory variable within each RCOMPCODE, the remaining explanatory variables have a minor impact and cause clusters or multimodal structures in Figure 15.

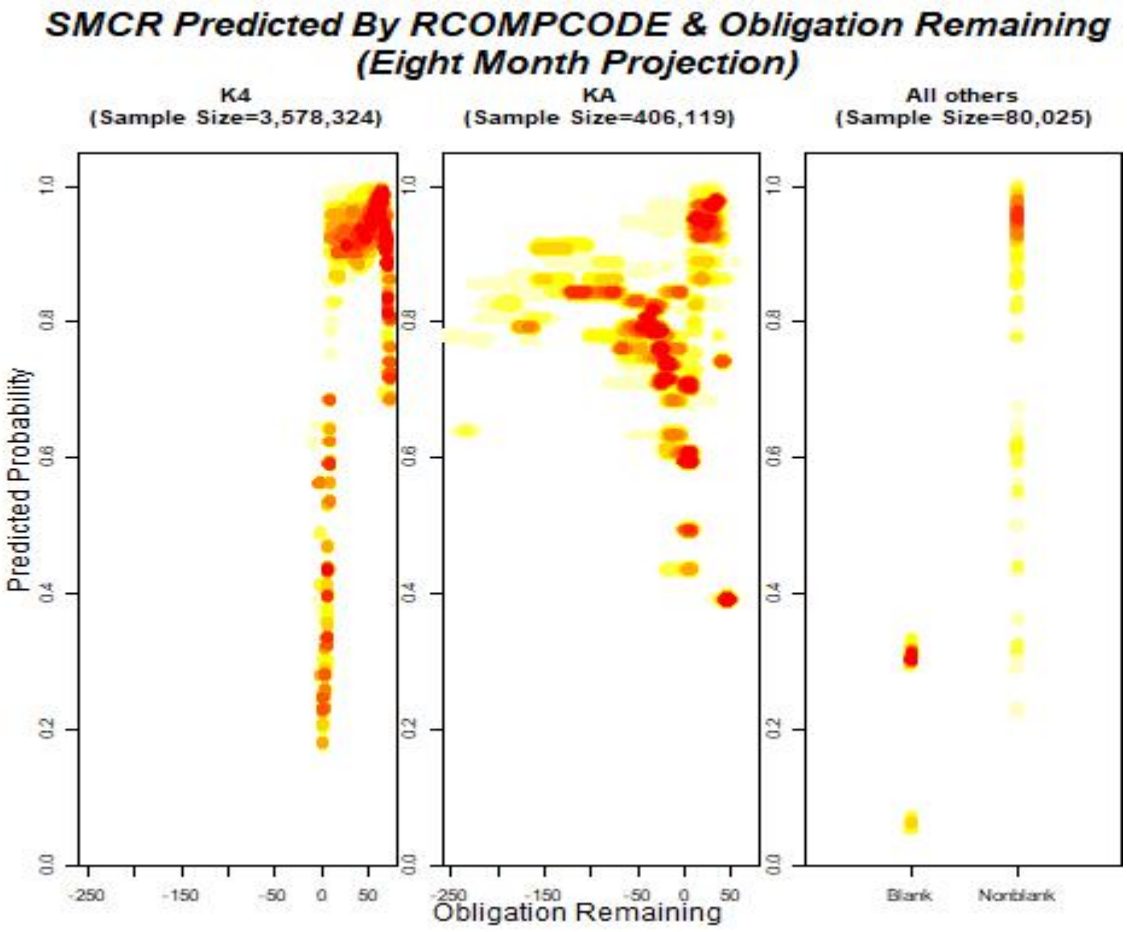

Figure 15. SMCR Predicted by RCOMPCODE & Obligation Remaining (Eight-Month Projection)

2. Output Analysis

Although we model the SMCR status group in the same way that we model the IADT status group, the differing characteristics of the populations yield significant differences in the modeled results. Specifically, the SMCR status group represents nearly 90% of the population in the dataset and has markedly different transition rates than the IADT status group. Over a typical twelve-month period, approximately 13% of the SMCR status group transitions to another group compared more than 95% of the IADT status group. As a result, the SMCR modeling ensemble increases and then plateaus in variance over time where the IADT ensemble plateaus and then declines.

As shown in Figure 16, the range of SMCR hindcast prediction interval widths are rarely narrower than or vary from the means of the training set. The exception is the four- through seven-month models that have consistently higher means. The four-month model also displays a narrowing of the ranges of prediction interval widths in hindcast. For more detail on the implications of this behavior, we refer the reader to the discussion of Figure 7 in Section IV.A.2.

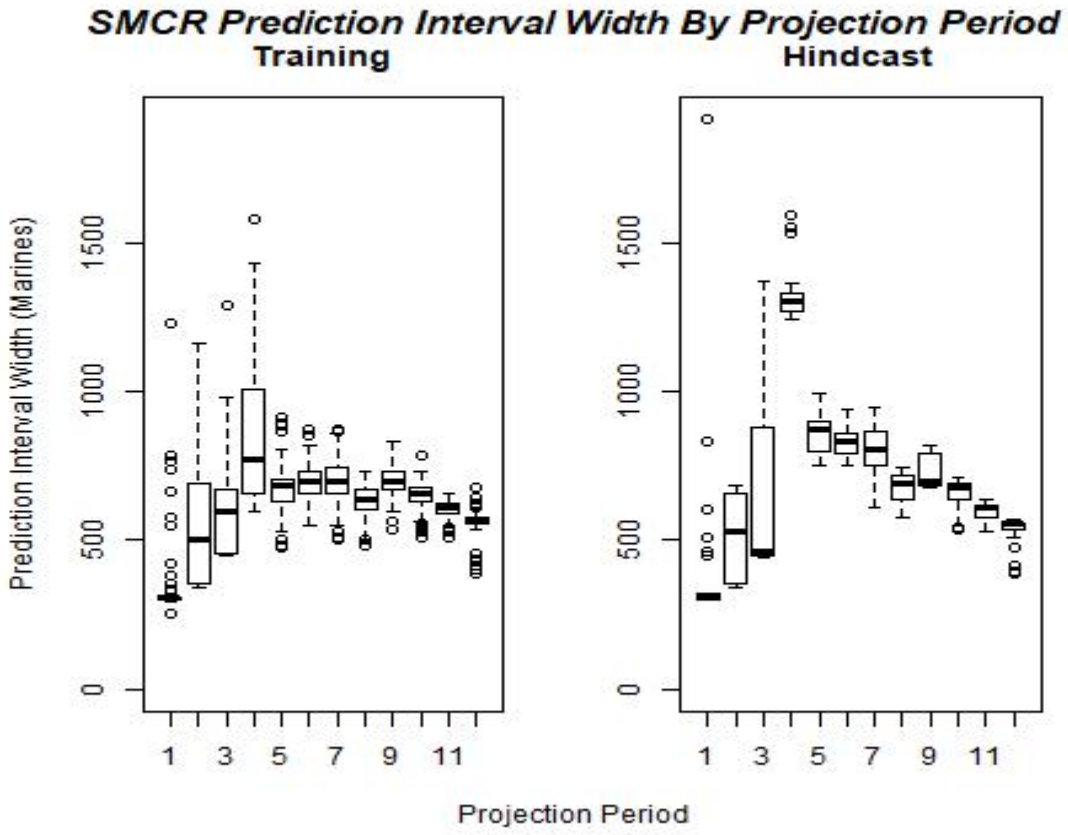

Figure 16. SMCR Prediction Interval Width by Projection Period

In Figure 17, we observe that FY 06 and FY 17 exhibit considerable error bias in SMCR in a similar fashion to the IADT results. In addition, FY 11 displays significant uncorrected bias within the training set. The SMCR Reserve Manpower Office has indicated that large policy changes influenced FY 11, and the office considers it to be an outlier in any analysis they conduct (S. Norton, Maj USMC-M&RA, personal communication, Apr 10, 2018). Also, unlike the IADT hindcast, FY 15 shows significant differences from the training period both in terms of bias and variance.

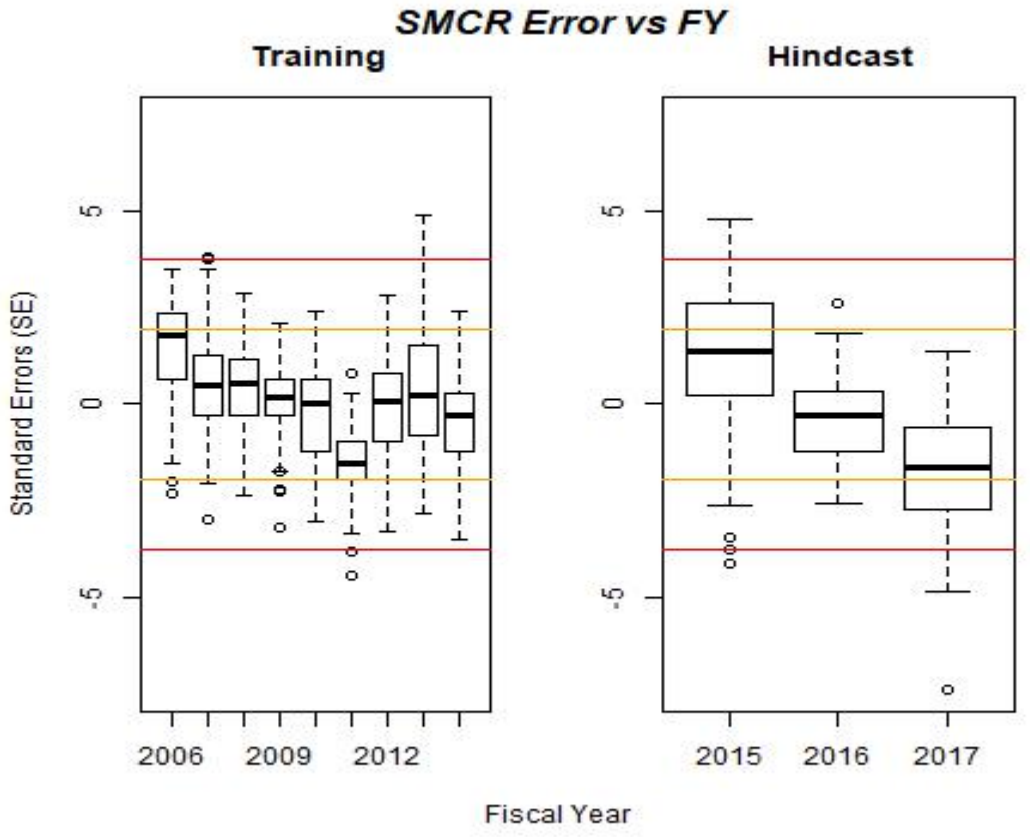

Figure 17. SMCR Training and Hindcast Errors by SEQ

When viewed by projection period, as in Figure 18, SMCR models show much more consistency. Unlike IADT, there is a tendency for near-term forecasts to have smaller deviations from predicted intervals. We expect this as the population size does not decay significantly in a 12-month period, so the underlying uncertainty drives most of the deviation. Although the underlying variance is relatively constant across the projection intervals, uncertainties increase beyond month 12. Based on sensitivity tests of model complexity, models beyond the nine-month horizon significantly underfit at any level of complexity. The explanatory variables necessary to project accurately that far into the future are not present in the data provided.

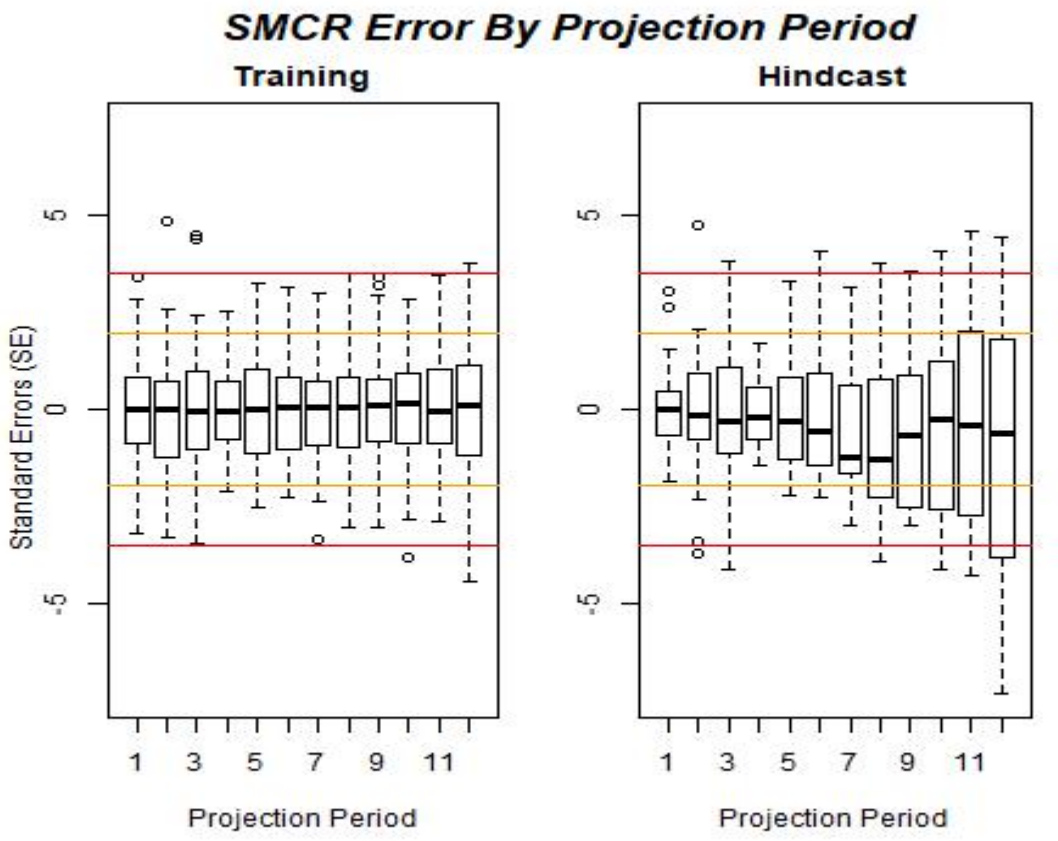

Figure 18. SMCR Error by Projection Period

Figure 19 illustrates the final results of the algorithm projected from the beginning, middle, and end of FY 15. The points and error bars represent the hindcasted value and 95% prediction interval. For comparison, the line represents the actual values for that period. The initial hindcast from SEQ 319 (SEP 2014) falls within the 95% prediction intervals throughout the entire period, with the exception of the 12-month period. The SEQ 325 (JUN 2015) hindcast is less accurate, with a significant upward trend that exceeds the prediction interval for the 6–12-month projections. The SEQ 331 (SEP 2015) hindcast exhibits similar errors in the 6–12-month projections. Clearly, the FY 16 data distribution atypically in ways that bias longer-term hindcasts. Retraining the algorithm at the end of FY 15 would reduce these errors.

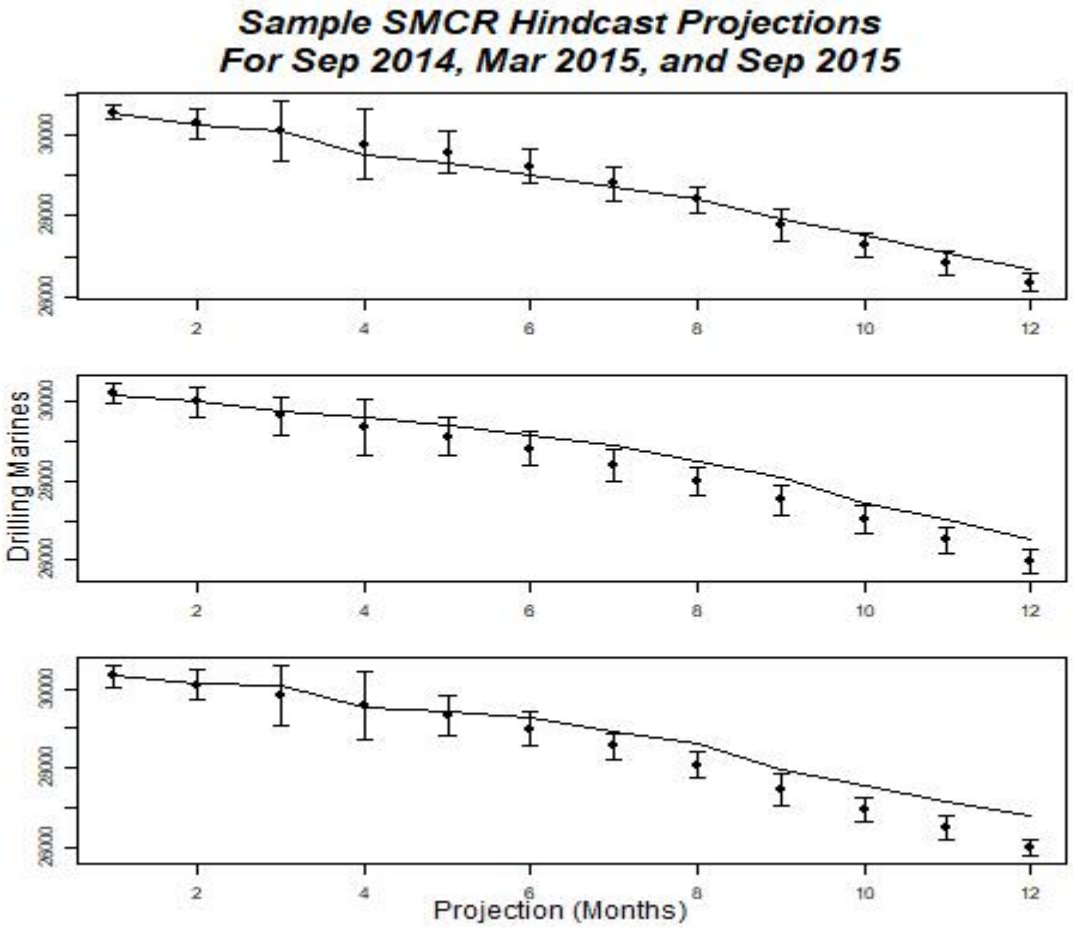

Figure 19. Sample SMCR Hindcast Projections

As Figure 20 illustrates, seasonality in the training set is largely under control and well within adjusted prediction intervals. While the hindcast does show seasonal variation, it does not resemble any pattern found in the training set and generally stays within adjusted prediction intervals. Given the small sample size of the hindcast, the high level of autocorrelation of the time series, and the lack of similar pattern in training, we attribute the monthly variations in SMCR errors to randomness.

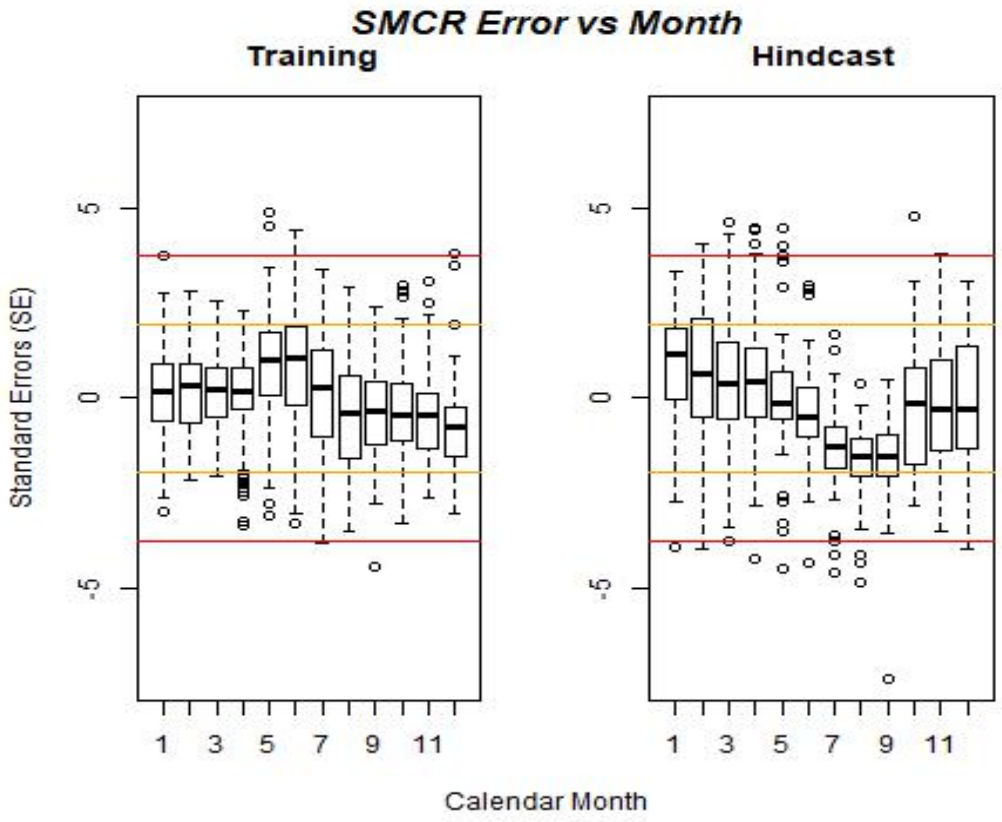

Figure 20. SMCR Error versus Calendar Month

C. EFFECT OF APPLYING VARIANCE RATIOS

A key aspect of this algorithm is the application of the Variance Ratio as a calibration for the variance estimate. Figure 21 illustrates the Variance Ratios calculated during training and used in prediction. The large values of VR for SMCR 1–4 month projections suggest that short-term projections of SMCR may not follow a binomial variance pattern or are significantly over-dispersed. Most of the remaining points are significantly above 1.0, indicating that large portions of these models are underfit. We believe that much of the underfitting it not due to the selected model complexities, but due to insufficient explanatory power. The exceptions are the IADT seven-month and eight-month projections with VRs of approximately .3 indicating that these two models are overfit.

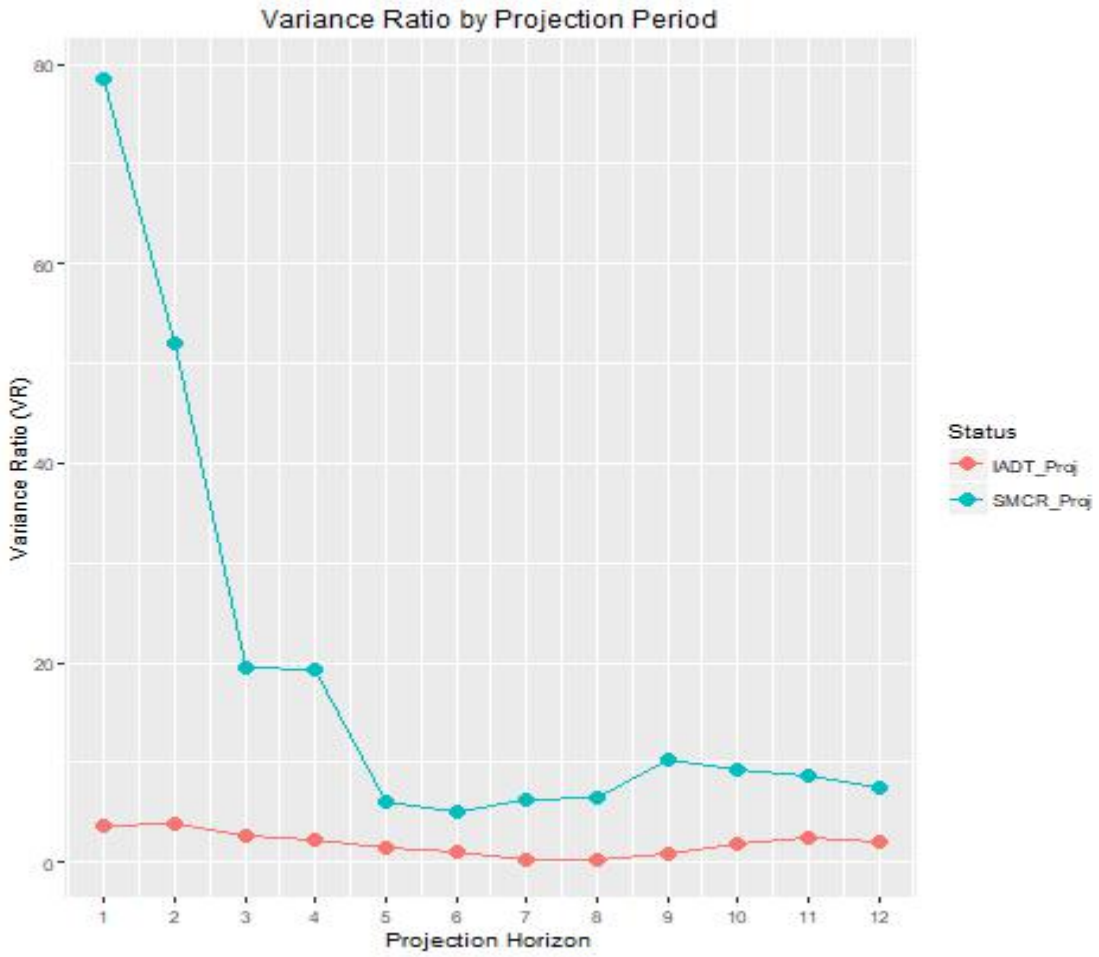

Figure 21. Variance Ratio (VR) by Projection Horizon

As seen in Figure 22, the calibration of estimated variance with VR has a significant effect. While the actual values tend to stay within the VR-adjusted prediction intervals, they are outside 0.95 envelope for nearly half of the raw prediction intervals. The VR adjustment represents a substantial improvement over the prediction intervals calculated solely from the leaf variance.

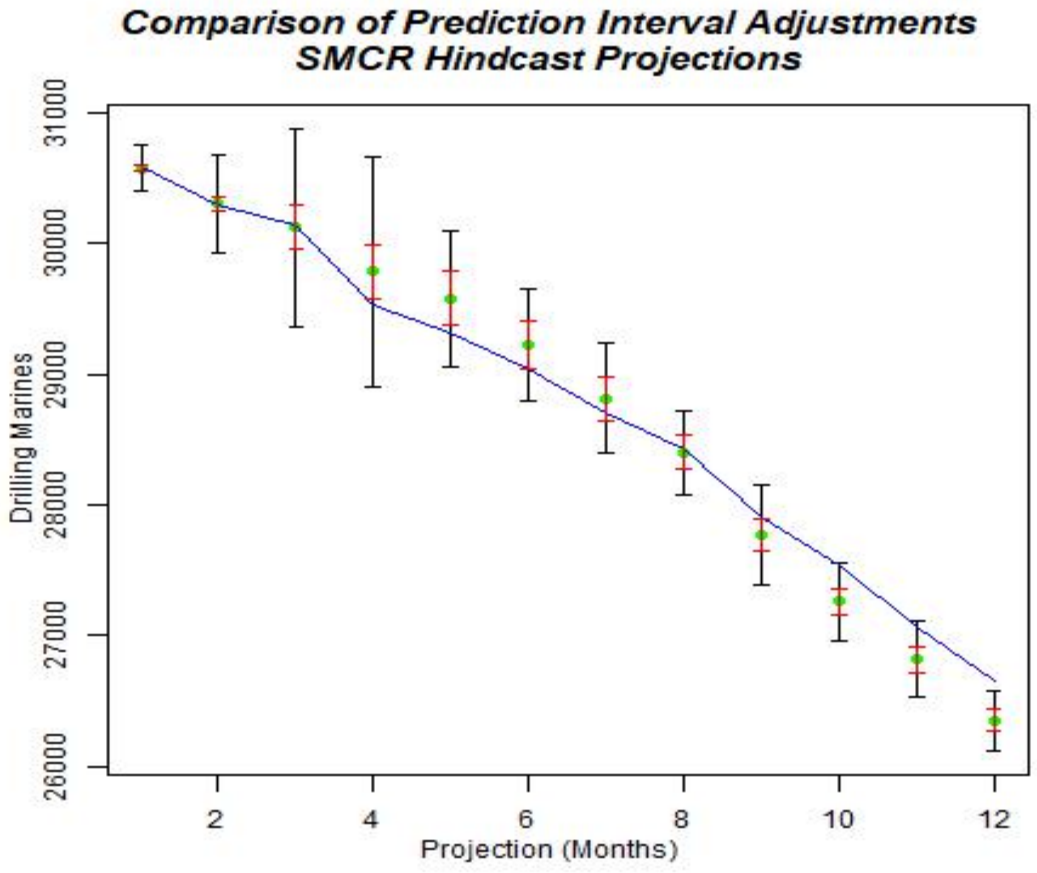

Figure 22. Comparison of Prediction Interval Adjustments

Figure 22 also makes apparent an unusual aspect of this dataset. The uncertainty of the projection in month one is relatively small, but grows increasingly uncertain through month four, as we expect. Beyond month four, the uncertainty declines, and by month 12, reaching the same level as month one. While the cohort population does decline during the forecast periods, the reduction in variance is far greater than a 20% population decline would produce. Based on the variable importance from Figure 13, we can infer that the low initial variability is due to the high autocorrelation of DODTCPG in the one-month projection. The decline in variability beyond the four-month projection is likely the result of explanatory variables that are predictive of a status change, but are uncertain in the timing. For example, if a given EAS date is only accurate to within a few months, it increases the uncertainty in the first few projections. For longer-term projections, the

uncertainty of the timing decreases as a proportion of the projection timing. Put another way, the explanatory variables may not have the skill to predict the specific month of a transition with high accuracy, but can predict the quarter of the transition much better.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

This study successfully demonstrates an algorithm to model future aggregate statuses for the SMCR population up to 12 months in the future. Key to this is the estimation of accurate prediction intervals from the underlying distribution of the data. While there are noticeable autocorrelated trends in the algorithm residuals, they are correctable with periodic retraining. This study also finds a clear gap in techniques to fit decision trees to predicted variables that have a defined variance. This gap results in difficulties maintaining proper model fit across the algorithms modeling ensemble.

We also find that many of the demographic explanatory variables significant to previous IADT attrition studies of the active duty population are found not to be significant contributors to the models in this study. This may be due to several factors. First, active duty populations may not be directly comparable to Reservist or SMCR populations. Second, by aggregating the population transitioning from IADT into the same model that forecasts continuation within SMCR, the IADT-specific explanatory variables do not achieve sufficient explanatory power relative to the SMCR explanatory variables to justify retention. Finally, it is possible that the demographic explanatory variables are not significant in the presence of the administrative explanatory variables. The administrative explanatory variables may encode the same information as the demographics through personnel selection processes.

Of the administrative data that we find to be predictive, the status codes within DODTCPG and RCOMPCODE are useful for short-term forecasts. For medium to long-term forecasts, the presence of an obligation date or an EAS date are usually more predictive than the dates themselves, indicating that the timing of data entry encodes some information on future status or continuation.

B. POTENTIAL APPLICATIONS

Although this study explores theoretical aspects of applying decision trees to variables with a defined variance, most of its practical findings are specific to the SMCR.

Further refinement is necessary before the techniques developed in this study could be applied generally and consistently.

C. TOPICS FOR FURTHER RESEARCH

1. Projection of the Unknown Population

As discussed in Section I.A.1, this study was limited to forecasting the known population of Marines in the SMCR and MCRISS databases. Full support of SMCR decision making also requires a forecast of the rates of arrival of the unknown population given a set of recruiting policies. As any gaps are identified in the existing population in future months, such a model would allow M&RA to predict the impact of policies and mission scenarios as they work to achieve their policy objectives.

2. Post-model Seasonality Adjustment

As indicated in Section IV.A.2, seasonality continues to limit the accuracy of the algorithm. Adding month as an explanatory variable in the models would likely cause interaction with FY that would overfit parts of the model. One alternative is to extend the model by post-processing the results with a simple seasonality adjustment.

APPENDIX. STATUS CODES

DODTCPG	
CODE	DESCRIPTION
PJ	Individual Ready Reserve
RE	Individual Ready Reserve
SA	Selected Marine Corps Reserve (SMCR)
SG	Active Reserve (Full-Time Support)
TB	Individual Mobilization Augmentee
UF	Initial Active Duty Training (IADT)
UP	Initial Active Duty Training (IADT)
UQ	SMCR Incremental Initial Active Duty Training (IIADT)
UX	SMCR

RCOMPCODE	
CODE	DESCRIPTION
K1	ENLISTED RES ON IADT AND/OR ELST
K2	ENLISTED RES 2ND INCREMENT IADT
K3	RES ON TEM ACDU FOR ETT OR RCT
K4	ENLISTED RES NPS OBLIGOR 6 YR ACDU & IDT
KA	SMCR IDT
KF	IMA IDT
K1	ENLISTED RES ON IADT AND/OR ELST
K2	ENLISTED RES 2ND INCREMENT IADT
K3	RES ON TEM ACDU FOR ETT OR RCT

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Baykiz, M. S. (2007). An analysis of Marine Corps Delayed Entry Program (DEP) attrition by high school graduates and high school seniors. Naval Postgraduate School (Master's thesis). Retrieved from https://calhoun.nps.edu/handle/10945/3655

Buddin, R. (1984). Analysis of early military attrition behavior. Santa Monica, CA: RAND Corporation.

Buttrey, S. E., & Whitaker, L. R. (2016). A scale-independent, noise-resistant dissimilarity for tree-based clustering of mixed data (Report No. NPS-OR-16-003). Retrieved from http://hdl.handle.net/10945/48615

Dausman, A. D. (2016). Determining a retention model for the Selected Marine Corps Reserve. Naval Postgraduate School (Master's thesis). Retrieved from http://hdl.handle.net/10945/48510

Emery, N. N. (2010). Forecasting United States Marine Corps Select Reserve end strength. Naval Postgraduate School (Master's thesis). Retrieved from http://hdl.handle.net/10945/5449

Erhardt, B. J. (2012). Development of a Markov model for forecasting continuation rates for enlisted prior service and non-prior service personnel in the Selective Marine Corps Reserve (SMCR) (Master's thesis). Retrieved from http://hdl.handle.net/10945/6791

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

Pedregosa, F., Varoquaux, G.,Gramfort, A., Michel, V.,Thirion, B.,Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2018, Jul 29). sklearn.tree.ExtraTreeClassifier. Retrieved from Scikit-learn: Machine Learning in Python: http://scikit-l.

Ross, S. (2006). The central limit theorem. In S. Ross (Ed.), *A First Course in Probability* (pp. 391–393). Upper Saddle River, NJ: Prentice Hall.

scikit-learn developers. (2017). sklearn.tree.DecisionTreeClassifier. Retrieved from scikit-learn.org: http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics 29*(2), 614–623.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California