2016-06

Demand forecasting: an evaluation of DOD's accuracy metric and Navy's procedures

Rigoni, Michael P.; Correia de Souza, Wagner

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

MBA PROFESSIONAL REPORT

DEMAND FORECASTING: AN EVALUATION OF DOD'S ACCURACY METRIC AND NAVY'S PROCEDURES

June 2016

By: **Michael P. Rigoni**
Wagner Correia de Souza

Advisors: **Geraldo Ferrer**
Kenneth Doerr

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2016	**3. REPORT TYPE AND DATES COVERED** MBA professional report

4. TITLE AND SUBTITLE
DEMAND FORECASTING: AN EVALUATION OF DOD'S ACCURACY METRIC AND NAVY'S PROCEDURES

5. FUNDING NUMBERS

6. AUTHOR(S) Michael P. Rigoni and Wagner Correia de Souza

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)
Naval Supply Systems Command, Weapon Systems Support
5450 Carlisle Pike, PO Box 2020, Mechanicsburg, PA 17055-0788

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
2016-5

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

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

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

In 2013, the Department of Defense (DOD) implemented an accuracy metric to monitor how well the services and Defense Logistics Agency were forecasting demand for inventory items. After three years, results were still poor. DOD uses a metric derived from the Mean of Absolute Percentage Error, yet it differs in significant ways, such as including unit cost to enable the aggregation of data pertaining to all items.

In this study, we analyze how unit cost and other parameters affect the validity of DOD metric results. Our research included a review of academic literature on forecast accuracy measurement that uncovered an alternative metric, Mean of Absolute Scaled Errors (MASE), which we tested against the DOD metric.

We found the DOD metric produced non-intuitive results and was adversely affected by unit cost and demand volume, while MASE avoided these errors. We utilized MASE to compare six forecasting methods and found that flexibility in choice of forecasting method produced better results than the naïve method when coefficient of variation (CV) is below 2.0.

We recommend that the DOD and Navy adopt MASE for aggregation and item-level forecast accuracy evaluation. We recommend that Navy utilize flexibility in choice of forecast method for individual items with CV below 2.0.

14. SUBJECT TERMS comprehensive inventory management improvement plan, mean of absolute scaled error, lead time adjusted squared error, forecast accuracy, benchmarking, naïve method, coefficient of variation	**15. NUMBER OF PAGES** 121
	16. COST 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

DEMAND FORECASTING: AN EVALUATION OF DOD'S ACCURACY METRIC AND NAVY'S PROCEDURES

Michael P. Rigoni, Lieutenant Commander, United States Navy
Wagner Correia de Souza, Captain Lieutenant, Brazilian Navy

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF BUSINESS ADMINISTRATION

from the

NAVAL POSTGRADUATE SCHOOL
June 2016

Approved by: Geraldo Ferrer, Ph.D.

Kenneth H. Doerr, Ph.D.

Bryan Hudgens
Academic Associate
Graduate School of Business and Public Policy

THIS PAGE INTENTIONALLY LEFT BLANK

DEMAND FORECASTING: AN EVALUATION OF DOD'S ACCURACY METRIC AND NAVY'S PROCEDURES

ABSTRACT

In 2013, the Department of Defense (DOD) implemented an accuracy metric to monitor how well the services and Defense Logistics Agency were forecasting demand for inventory items. After three years, results were still poor. DOD uses a metric derived from the Mean of Absolute Percentage Error, yet it differs in significant ways, such as including unit cost to enable the aggregation of data pertaining to all items.

In this study, we analyze how unit cost and other parameters affect the validity of DOD metric results. Our research included a review of academic literature on forecast accuracy measurement that uncovered an alternative metric, Mean of Absolute Scaled Errors (MASE), which we tested against the DOD metric.

We found the DOD metric produced non-intuitive results and was adversely affected by unit cost and demand volume, while MASE avoided these errors. We utilized MASE to compare six forecasting methods and found that flexibility in choice of forecasting method produced better results than the naïve method when coefficient of variation (CV) is below 2.0.

We recommend that the DOD and Navy adopt MASE for aggregation and item-level forecast accuracy evaluation. We recommend that Navy utilize flexibility in choice of forecast method for individual items with CV below 2.0.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AR	Average Ranking
ARRSES	Adaptive Response Rate Simple Exponential Smoothing
ASD(L&MR)	Assistant Secretary of Defense for Logistics and Material Readiness
CG	Comptroller General
CIMIP	Comprehensive Inventory Management Improvement Plan
$CIMIP_f$	CIMIP's forecast accuracy equation
CV	Coefficient of Variation
DAM	Denominator-Adjusted MAPE
DLA	Defense Logistics Agency
DOD	Department of Defense
ERP	Enterprise Resource Planning
FMFIA	Federal Managers Financial Integrity Act
FY	Fiscal Year
GAO	General Accounting Office / Government Accountability Office
GMRAE	Geometric Mean Relative Absolute Error
IM	Item Manager
JASA	Journal of the American Statistical Association
LASE	Lead-time Adjusted Squared Error
LCI	Life Cycle Indicator
MA	Moving Average
MAE	Mean Absolute Error
MASE	Mean of Absolute Scaled Error
MAPE	Mean of Absolute Percentage Error
MdAE	Median Absolute Error
MdAPE	Medians of Absolute Percentage Errors
MdRAE	Median Relative Absolute Error
MdSAPE	Median Symmetric Absolute Percentage Error
MRAE	Mean Relative Absolute Error
MSE	Mean Squared Error

NAVSUP	Naval Supply Systems Command
NDAA	National Defense Authorization Act
NIIN	National Individual Identification Number
OMB	Office of Management and Budget
OSD	Office of the Secretary of Defense
PB	Percentage Better
RMAE	Relative Mean Absolute Error
RMAPE	Relative Mean Absolute Percentage Error
RMdAE	Relative Median Absolute Error
RMdSPE	Root Median Squared Percentage Error
RMSE	Root Mean Squared Error
RMSPE	Root Mean Squared Percentage Error
RRMSE	Relative Root of Mean Squared Error
SES	Simple Exponential Smoothing
sMAPE	Symmetric Mean Absolute Percentage Error
sMdAPE	Symmetric Median Absolute Percentage Error
TSP	Thrift Savings Plan
USD(AT&L)	Under Secretary of Defense for Acquisition, Technology, and Logistics
WSS	Weapon Systems Support

ACKNOWLEDGMENTS

We would like to thank all of the personnel at NAVSUP WSS who assisted with our research, especially Mr. Eric Liskow and Ms. Erin Groft. Their willingness to speak with us on multiple occasions and provide valuable data was critical to the success of our research.

We would like to thank the NPS Thesis Processing Office, especially Michele D'Ambrosio, for guiding us on this endeavor and providing helpful critiques and corrections.

Last, but not least, we would like to thank our two faculty advisors, Dr. Geraldo Ferrer and Dr. Kenneth Doerr. While they continually challenged us to produce ever more valuable and academically sound research, we would not have been able to complete this work without their guidance and mentorship.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

In an environment of increasing congressional pressure and decreasing defense funding, the Department of Defense (DOD) has been investing considerable effort and resources into increasing the efficiency and effectiveness of its supply chain management. To set a common understanding of that issue, the next section provides a summary of the pivotal events that led to the Comprehensive Inventory Management Improvement Plan (CIMIP), which aims to reduce excess DOD secondary inventory. "DOD defines secondary items as minor end items; replacement, spare, and repair components; personnel support and consumable items. Examples of secondary items include aircraft, tank, and ship components; construction, medical, and dental supplies; and food, clothing, and fuel" (General Accounting Office [GAO], 1988, p. 1). Principal inventory items consist of items such as aircraft, vehicles and ships. DOD stratifies secondary inventory into four categories: approved acquisition objective, economic retention stock, contingency retention stock, and potential reutilization stock. The approved acquisition objective stock is calculated in order to meet current requirements, while the other three categories are considered by GAO to be in excess of current requirements (Government Accountability Office [GAO], 2015b). While not directly stated, the DOD appears to only consider potential reutilization stock as excess and seems reluctant to dispose of economic and contingency retention stocks due to the potential that they will be needed in the future. Figure 1. shows how much of the Navy's secondary inventory was considered excess in fiscal years 2004 through 2007.

1. Pre CIMIP

On September 8, 1982, the U.S. Congress enacted the Federal Managers Financial Integrity Act (FMFIA). Primarily an amendment to the Accounting and Auditing Act of 1950, it required "ongoing evaluations and reports of the adequacy of the systems of internal accounting and administrative control of each executive agency" (Federal Managers Financial Integrity Act of 1982, 2012). While implementation of the act did not

immediately solve the issues that it intended to address (GAO, 1989), it became a driving force behind the ongoing efforts to improve the way that the federal government manages resources.

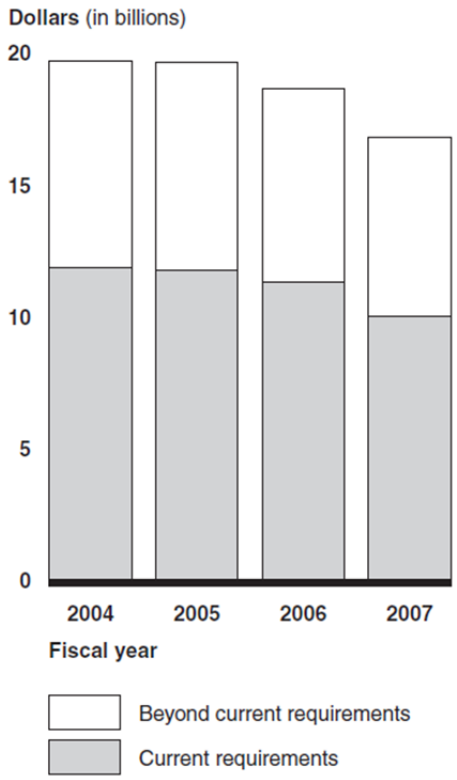
In July 1988, the General Accounting Office, as GAO was known at the time, published a report in response to Senate inquiries regarding the growth of secondary item inventories within the DOD (GAO, 1988). Between 1980 and 1987, according to the report, the value of the DOD's secondary items grew from $43 billion to $94 billion, about $19 billion of which was attributable to the Navy. Of this $51 billion dollar increase, $27 billion was due to the increasing size of the U.S. military, while $19 billion was considered to be in excess of requirements and $5 billion was "unstratified," which means that it was not allocated to a specific inventory purpose such as current requirement or economic retention. This report contributed to the growing number of GAO studies concluding that DOD needed to do a better job of managing its inventory.

On January 23, 1990, GAO released a letter from the comptroller general (CG) of the United States addressed to the chairman of the U.S. Senate committee on governmental affairs and the chairman of the U.S. House of Representatives committee on government operations (GAO, 1990). In the letter, the CG highlights the need to improve the internal controls and financial management systems of the federal government. In October 1989, in support of the Office of Management and Budget (OMB) identification of "high risk" areas, and after reviewing reports submitted under FMFIA, the CG identified 14 target areas that would receive special attention from the GAO. One of those areas singled out for special review was the DOD inventory management systems, due to growing excess inventory levels now valued at over $30 billion, and numerous other indicators of poor financial management. Since that time, the GAO has considered the DOD's inventory management a high-risk area, and although the name of the problem has changed to DOD supply chain management, it remains one of the 32 high-risk areas on the GAO's 2015 list (GAO, 2015a).

In December 2008, the GAO published a report that evaluated the cost efficiency of the Navy's spare parts inventory. In explaining why the Navy had accumulated excess secondary inventory, the report concluded, "much of the inventory that exceeded current

requirements or had inventory deficits resulted from inaccurate demand forecasts" (GAO, 2008, p. 34). The report also documented the results from surveys of the Navy's Item Managers (IM) who identified many additional factors that they felt were contributing to inventory excesses and deficits (GAO, 2008). From 2004 to 2007, GAO calculated that secondary inventory in excess of current requirements averaged about 40%, or $7.5 billion, of total Navy inventory. Figure 1. from the report shows this trend in 2007 dollars. This report was the second in a series of GAO reports that reviewed the secondary inventory management of the Air Force (GAO, 2007), Army (GAO, 2009) and Defense Logistics Agency (DLA) (GAO, 2010). To varying degrees, each of these reports commented on the need for improved demand forecasting. Subsequently, GAO concluded that "inaccurate demand forecasting is the leading reason for the accumulation of excess inventory" (GAO, 2011, p. 11) throughout the services and DLA.

Figure 1. Navy Secondary Inventory Meeting and Exceeding Requirements (FY 2004–2007). Source: GAO (2008).



After 20 years of effort with little improvement, Congress inserted language into the fiscal year (FY) 2010 National Defense Authorization Act (NDAA) that required the development of an extensive plan that would improve the inventory management practices within the DOD. When the NDAA was enacted on October 28, 2009, section 328 required that this plan be provided to Congress for review within 270 days. The plan was required to address eight separate elements intended to improve "the inventory management systems of the military departments and the Defense Logistics Agency with the objective of reducing the acquisition and storage of secondary inventory that is excess to requirements" (NDAA, 2009, para [a]). The most relevant aspect to this research is the second part of the first element, which required the "development of metrics to identify bias toward over-forecasting and adjust forecasting methods accordingly" (NDAA, 2009, para [b(1)]). This legal requirement would eventually result in the DOD developing a common metric for forecast accuracy and forecast bias that would measure the performance of each military service and DLA.

2. CIMIP

As required by the FY10 NDAA section 328, the Assistant Secretary of Defense for Logistics and Materiel Readiness published the DOD's Comprehensive Inventory Management Improvement Plan in October 2010. In addition to fulfilling the demands of Congress, the objective of the plan was to drive "a prudent reduction in current inventory excesses as well as a reduction in the potential for future excesses without degrading materiel support to the customer" (Assistant Secretary of Defense for Logistics and Material Readiness (ASD[L&MR]), 2010, p. iii). In that document, chapter one contains an overview of inventory management improvement, assigns responsibilities and highlights the implementation strategy. Chapters two through nine detail the eight sub-plans that have been developed to address the eight elements required by section 328, while chapter ten details four additional improvement actions that the DOD is developing on their own initiative. Although these department-wide actions were not specifically required by section 328, they were included in the plan because "these actions support the Department's intent to improve DOD inventory management and reduce excesses" (ASD[L&MR], 2010, p. 10–1). Appendix A lists 17 other DOD strategies, plans, or

efforts that are consistent with the CIMIP overall objective of reducing secondary item inventory levels. Most importantly, Appendix A highlights that the plan is consistent with the objectives of the DOD Logistics Strategic Plan, which "identifies high level goals, performance measures, and key initiatives that support the DOD priorities and drive the logistics enterprise improvements" (ASD[L&MR], 2010, p. A-1). Appendix B lists the 12 GAO reports published between March 2006 and May 2010 that are related to secondary item inventory, summarizes their findings, and briefly states how the plan will address each finding. Appendix C reprints the entirety of section 328 of the FY10 NDAA, while Appendix D provides a list of abbreviations.

While the plan is a comprehensive approach to improving materiel management, only chapter II, *Sub-Plan A: Demand Forecasting*, is relevant to our research. The overall objective of sub-plan A "is to improve the prediction of future demands so that inventory requirements more accurately reflect actual needs" (ASD[L&MR], 2010, p. 2-3). In order to accomplish this objective, the DOD did a thorough review of current forecasting procedures and methodologies in search of ways to improve the process. As a result of this review, the DOD established five action items that required further work to address the issues with demand forecasting. Of these five action items, *Action A-2: Implement Standard Metrics to Assess Forecasting Accuracy and Bias* is the basis for this research project. DOD targeted the end of fiscal year 2011 to identify these two metrics and the end of fiscal year 2012 to establish the processes by which the DOD components could set targets and begin utilizing the common metrics. The accuracy metric intends to measure forecast performance while minimizing bias and generating results for various inventory segments. The bias metric intends to identify over- and under-forecasts in order to prevent inventory excesses and deficits.

3. Post CIMIP

In January 2011, GAO published its required 60-day assessment of the DOD's plan to meet the eight elements identified in section 328 (GAO, 2011). While GAO concluded that the plan did address all eight elements from section 328 of the FY10 NDAA, the report identified five general areas that could produce implementation

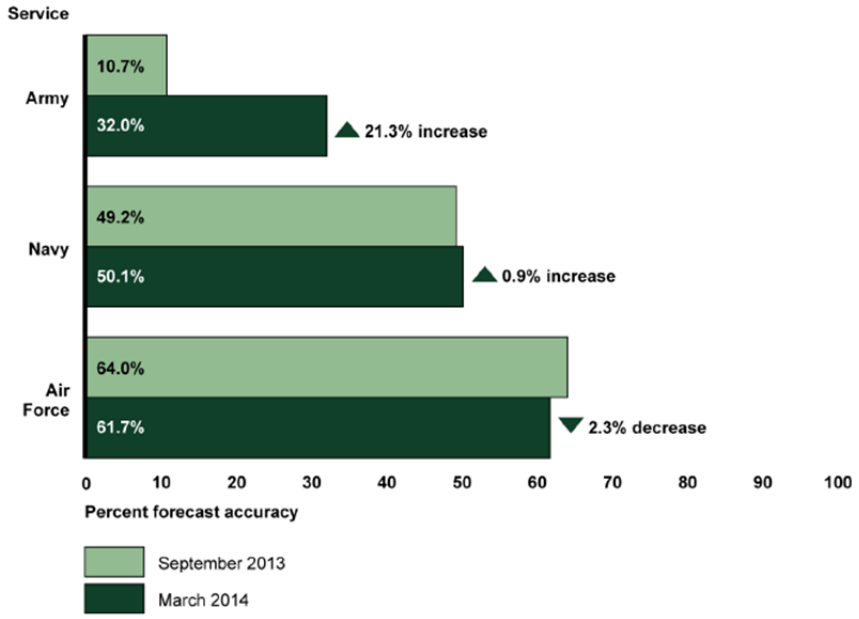
challenges if not managed properly. One of the examples the report used to highlight potential friction areas was the requirement to develop a standard accuracy metric and performance targets. GAO felt that this level of standardization could be difficult to reach given the fact that the services and DLA had different approaches to measuring demand forecast accuracy (GAO, 2011, p. 6). In May 2012, GAO fulfilled its final requirement from section 328 by publishing its 18-month assessment of the effectiveness in which the services and DLA have implemented the plan they developed. GAO concluded that while the DOD was "making progress towards…establishing a department-wide set of standardized metrics for inventory management. Moving forward, DOD's inventory management improvement efforts would benefit from challenging, but achievable targets for reducing its on-order and on-hand excess inventory" (GAO, 2012, p. 30). Within the demand-forecasting sub-plan, GAO determined that while DOD had successfully developed the forecast accuracy and bias metrics, the effective implementation of these metrics still required a sustained effort to meet the expected completion date of September 2012. The accuracy metric that was developed is an absolute error metric, while the bias metric is a signed error metric. The formulas for these two metrics are discussed further in Chapter II and are shown in Equations (2.24) and (2.25).

Reinforcing CIMIP efforts, the acting Under Secretary of Defense for Acquisition, Technology, and Logistics (USD[AT&L]) signed DOD Instruction 4140.01 in December 2011, establishing that DOD's supply chain materiel management "shall operate as a high-performing and agile supply chain responsive to customer requirements during peacetime and war while balancing risk and total cost" (Kendall, 2011). In addition to clearly defining policy and assigning responsibility for management of material across the DOD supply chain, this instruction laid out the framework for 11 *DOD Supply Chain Material Management Procedures* manuals. In February 2014, the 11 manuals were published as volumes 1 through 11 of DOD Manual 4140.01 with each covering specific supply chain procedures. Volume 2, *Demand and Supply Planning*, among other things provided guidance on how DOD components should forecast customer demand. Volume 10, *Metrics and Inventory Stratification Reporting*, required among other things that the DOD utilize metrics that were specific, measureable,

actionable, realistic, and timely, which included demand forecast accuracy as an example of such a metric.

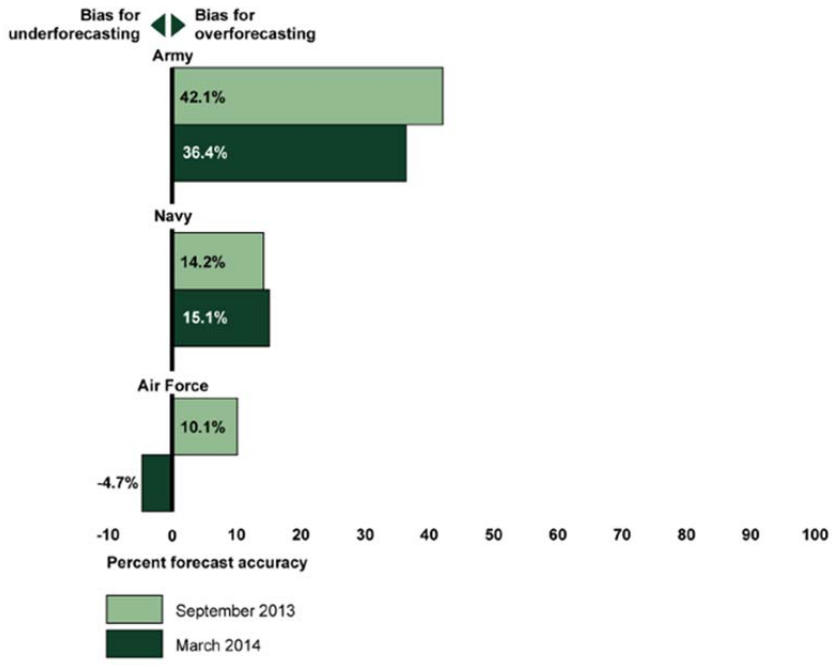
In April 2015, GAO released its most recent report related to defense inventory management, concluding that the services had generally been able to reduce their excess inventory, which was the primary objective of section 328 of the FY10 NDAA. Although this result was positive, GAO had seven recommendations to improve how DOD managed inventory. While GAO recommended that DOD establish goals for these metrics, DOD wanted to collect more data to establish a performance baseline before setting any department-wide goals (GAO, 2015b, p. 43). The report also reviewed results from the first and second metrics reporting periods. These metric results are reported semi-annually for the preceding 12-month period, so the first period covered all 12 months of FY13 ending in September 2013. The second period covered the last six months of FY13 and the first six months of FY14 ending in March 2014. Figure 2. and Figure 3. show the results reported by three services during these two 12-month reporting periods. The figures do not include the results for DLA or the non-aviation material for the Marine Corps. The Marine Corps aviation material is included in the Navy results.

Figure 2. Demand Forecast Accuracy Performance by Service. Source: GAO (2015).



The Air Force reported the highest forecast accuracy for these periods. The Army showed the greatest improvement over the two reporting periods.

Figure 3. Demand Forecast Bias by Service. Source: GAO (2015).



The Army had the largest bias for over-forecasting demand, followed by the Navy and the Air Force. In the second reporting period, the Air Force reported a negative bias, which indicates that they were under-forecasting their demand.

In response to the Navy's relatively poor performance in both forecast accuracy and bias, Naval Supply Systems Command (NAVSUP) reported that they were "reviewing and analyzing their demand forecasting processes and planning factors to improve performance on DOD's forecast accuracy and bias metrics tracked across the department" (GAO, 2015b, p. 46).

B. DATA DESCRIPTION AND RECENT RESULTS

The business rules for calculating the DOD's demand forecasting accuracy and bias metrics that were provided to DLA and each of the services specify eight forecast data elements and two demand history data elements that should be included in their data captures (DOD, 2013). These elements were

Forecast Data Elements

- NIIN / family head / subgroup master
- Demand forecast (monthly/quarterly/semi-annually)
- Latest acquisition cost or moving average cost
- Reparable/consumable indicator
- Unit of issue
- Unit of measure
- Time frame of the forecast (start date)
- Date the forecast was made (forecast date)

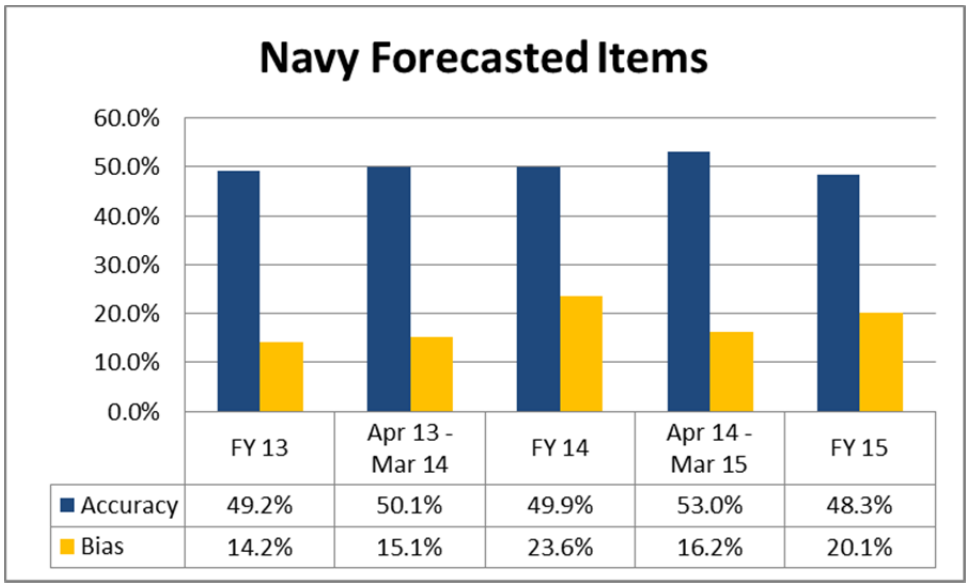
Demand History Data Elements

- Actual demand
- Timeframe of demand

NAVSUP Weapon Systems Support (WSS) provided CIMIP compliant data for fiscal years 2013, 2014 and 2015. The raw data elements provided the national individual identification number (NIIN), quarterly demand forecast, repair indicator, stock routing code, replacement cost, acquisition advice code, performance based logistics indicator, family group code, unit of measure, life cycle indicator (LCI), cognizance code, actual

annual demand, and annual naïve forecast. The FY14 and FY15 data calculated additional elements such as annual demand forecast, total dollar calculations, absolute and signed errors, and line item forecast accuracy and bias metrics. The FY15 data also included a bar graph of the Navy's overall CIMIP results reported to DOD for the five previous 12-month evaluation periods (Figure 4).

Figure 4. Navy CIMIP Forecast Metric Results FY13-FY15. Source: NAVSUP (2015).



	FY 13	Apr 13 - Mar 14	FY 14	Apr 14 - Mar 15	FY 15
Accuracy	49.2%	50.1%	49.9%	53.0%	48.3%
Bias	14.2%	15.1%	23.6%	16.2%	20.1%

Accuracy and bias results are reported to DOD semi-annually for the preceding 12-month period, which creates a six-month overlap in the data. The accuracy result is an absolute error metric that summarizes the Navy's forecasting performance. The bias result is a signed error metric that represents the degree of over-forecasting

C. PURPOSE AND BENEFITS OF STUDY

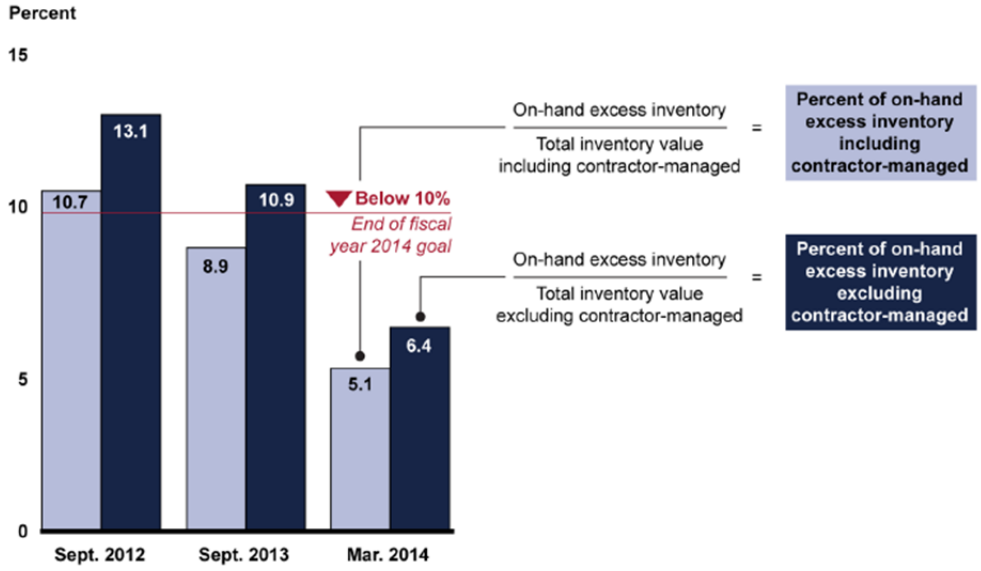
This research effort intends to review the validity of the DOD's newly implemented CIMIP forecasting metrics and identify weaknesses that may not be apparent to the casual observer of forecast accuracy metrics. We also intend to provide recommendations that will improve the DOD's efforts to increase forecast accuracy, which should result in better forecasts in the future, decreasing levels of excess inventory and ultimately, substantial cost savings to the DOD. While we certainly appreciate the complexity of forecasting future demand and accurately measuring those forecasts, and

recognize the amount of effort that has already been devoted to this issue, we will demonstrate that our research can provide value to these DOD efforts. Even if the DOD disregards our recommendations, there are still opportunities for the Navy, or the other services, to implement our recommendations and improve their demand forecasting efforts.

D. RESEARCH QUESTIONS

In 2011, GAO declared that "inaccurate demand forecasting is the leading reason for the accumulation of excess inventory" (p. 11), and as Figure 5 shows, the Navy has been making steady progress in reducing its on-hand excess inventory; however, despite this good news, the CIMIP forecast results have not significantly changed (Figure 4).

Figure 5. Navy On-Hand Excess Inventory, Sept. 2012 to Mar. 2014. Source: GAO (2015).



	Sept. 2012	Sept. 2013	Mar. 2014
Total value excluding contractor-managed	$22.6	$22.6	$22.1
Total value including contractor-managed	27.7	27.9	27.6
Value of on-hand excess inventory	3.0	2.5	1.4

The vertical bars represent excess inventory as a percentage of total inventory. The bottom table shows inventory dollar values in billions. While total inventory value has remained constant, whether you exclude contractor-managed inventory or not, excess inventory has been decreasing in real dollar values and as a percentage of total inventory.

Although many factors contribute to excess inventory levels, if the "leading reason for the accumulation of excess inventory" (GAO, 2011, p. 11)—forecast accuracy—is not improving or getting worse while excess inventory is declining, then this raises the question of whether demand forecasting is actually the largest contributor; or alternatively, if forecasting performance is not being accurately measured. Our intuition is that the answer lies in the second justification, and we intend to show it by addressing the following questions:

- Does the CIMIP forecasting metric capture forecast error in a way that is actionable?
- Are the CIMIP forecasting accuracy results impacted by variables or data set characteristics that are not directly related to forecast error?
- Does the CIMIP forecasting metric provide a useful product to the forecasters that enables them to prioritize their forecast improvement efforts?
- Is there an alternative forecast accuracy equation that both enables the aggregation of accuracy results for multiple line items, with various units-of-measure, while also providing actionable results at the item level?

Finally, it is also important to investigate how the forecast accuracy can generate valuable information to the Navy's managers.

E. SCOPE, ORGANIZATION AND METHODOLOGY

While inventory management improvement efforts span a large range of topics detailed in the 2010 CIMIP, this research will focus primarily on one line of effort to improve demand forecasting: the measurement of forecast accuracy. Chapter II is a summary of the traditional academic accuracy metrics and a compilation of the most valuable findings in the existing literature. Chapter III aims to present an in-depth analysis of the CIMIP equation and a comparison to an alternative accuracy metric, Mean Absolute Scaled Error (MASE). Those analyses are composed of specific tests to uncover the existence of inherent flaws or undesirable characteristics in the current metric. In order to compare the accuracy metrics, we assess them utilizing four desirable characteristics. The tests we will conduct utilize three different methods, according to specific purposes. The first method uses fictional numbers, the second uses real numbers

extracted from available data, and the third generates Monte Carlo Simulations using the Crystal Ball program.

Although we did not intend to make this a discussion on forecasting methods, the interrelatedness of forecasting methods and results measurement make it unavoidable. Therefore, in Chapter IV, we analyze alternative ways to generate more accurate demand forecasts. In Chapter V, we summarize the most important findings, make recommendations for DOD and Navy, and propose future areas of research to continue to advance the effectiveness of DOD and Navy forecasting efforts.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW

A. INTRODUCTION

This chapter presents a review of the evolutionary path of knowledge in the field of forecast accuracy, while also providing an overview of the most popular forecast accuracy measures.

B. FORECAST ACCURACY

Demand forecasts are a key component to effective inventory management. Delivery of inputs to production takes time and, even considering a deterministic scenario, managers need to be precise in determining the correct time to transmit their orders to suppliers, in order to avoid costs from shortages or by holding excess inventory.

Reinforcing that idea, Makridakis and Hibon (2000) claim that "forecasting accuracy is a critical factor for, among other things, reducing costs and providing better customer service" (p. 451). The effects of an inaccurate prediction are intensified when variability takes place, making the importance of forecast accuracy even more important.

1. History of Forecast Accuracy Measurement

Over the last 50 years, researchers have invested considerable time and effort to increase the understanding of forecast accuracy. While there is not a consensus about the first academic article on forecast accuracy, Ferber (1956) and Schupack (1962) are considered pioneers in this field. They tested multiple forecasting methods, using correlation index and various accuracy metrics to determine whether forecast models that demonstrated a good fit to past data could then generate good forecasts. The results did not support this hypothesis and they concluded that best fit on past data is not a good measure of forecast accuracy. Moreover, forecast method rankings do not change much by using different forecast accuracy metrics and there is no absolute best forecast method.

As computer processing capabilities grew, researchers could proceed with broader studies to measure the accuracy of different forecast methods. Fildes and Makridakis (1995) found that in the 20 years from 1971–1991, approximately 130 articles per year

were published in the *Journal of the American Statistical Association* (JASA) related to time series analysis. For example, Newbold and Granger (1974) used average squared forecast errors to assess the accuracy of three forecast methods, each one applied to 106 time series. A few years later, Nelson and Granger (1979) were able to analyze five forecast methods through twenty-one time series, utilizing ten forecast horizons and two different accuracy metrics.

Newbold and Granger (1974) were able to make insightful conclusions regarding the use of non-automated forecast methods. They found that the Box-Jenkins forecast method was capable of making up for its significantly longer calculating time by producing more accurate estimations. Moreover, results from that forecasting method could be further improved by combining them with other fully automated procedures, like Holt-Winters or a stepwise autoregressive forecast. They also provided guidelines to optimize the choice of forecast methods according to the length of the time series. The idea of combining forecast methods in order to increase accuracy is one of the most valuable contributions in the field of forecasting and was first investigated by Reid (1968) and was further discussed by Nelson (1972), Cooper and Nelson (1975), Makridakis and Winkler (1983), Nelson (1984), Clemen (1989), Fildes (1989), among many others.

By the late 1970s, the question of what is the best forecast method seemed to be far from a solid answer. Utilizing the increasing power of computing capabilities and availability of new knowledge in the field of time series, Makridakis et al. (1979) and (Makridakis et al., 1982) conducted accuracy analysis on a much greater number of forecast methods.

Moreover, Makridakis et al. (1982) was the first empirical study of what became known as the M-1 Competition, which began the M-series Competitions. Makridakis et al. (1993) and Makridakis and Hibon (2000) published the M-2 and M-3 Competitions, respectively, which attempted to uncover situations in which one forecast method is expected to outperform others.

M-1 Competition in 1982 was based on the consensus of nine authors and made important contributions to the literature. It analyses 24 different forecast methods using 1,001 time series and five accuracy metrics: Mean Average Percentage Error (MAPE), Mean Squared Error (MSE), Average Ranking (AR), Medians of Absolute Percentage Errors (MdAPE), and Percentage Better (PB). The major findings of M-1 Competition are that there is no forecast method capable of minimize forecast errors in all kinds of demand patterns; more complex forecast methods do not always outperform rudimentary ones; and the best technique changes from one forecast horizon to the next and when different measures of accuracy are used.

The M-1 Competition also developed the categorization of time series in order to allow for the possibility of one technique to perform better when specific circumstances are present. That method is in accordance to Gilchrist (1979), which affirmed that averaging accuracy measures for several time series might hide the ability of a forecast method to deal with one specific type of time series better than others. However, one may infer that the way the time series were then grouped may have influenced the results.

Those findings were criticized by Armstrong and Lusk (1983) who identified the lack of interpretation or discussion about the results as an opportunity to open a discussion among experts aiming to clarify important aspects of forecast accuracy.

In order to address critics related to organization of results, M-2 Competition in 1993 made a simpler analysis, evaluating 16 forecast methods, each one applied to 29 time series, just using one accuracy metric, MAPE. It concludes in favor of both the exponential smoothing and the Dampen and Single smoothing methods, considered as being among the simplest. It also found that relatively sophisticated forecast methods are expected to perform better when randomness of series is small.

The M-3 Competition in 2000 moved back to extensive analysis, while as many as 3000 time series were used to generate forecasts, using 24 different methods, which accuracy were measured by five metrics: MAPE, AR, Median Symmetric Absolute Percentage Error (MdSAPE), PB and Median of Relative Absolute Error (MdRAE). It rejected the argument that more complex methods outperform simpler ones. It found that

the best method varies according to the accuracy metric used and that a combination of forecast methods is able to increase forecast accuracy.

Armstrong and Collopy (1992) presented a different approach on the use of forecast accuracy, as it evaluated measures of forecast accuracy, instead of forecast methods themselves, by using 191 economic time series. They provided a new approach to judge accuracy metrics, by using a framework composed by reliability, construct validity, sensitivity to small changes, protection against outliers, and relationship to decision making. Final conclusions were favorable to the use of MdRAE as an accuracy metric.

Following that discussion, Hyndman and Koehler (2006) provide a comprehensive critical survey of accuracy measures to uncover significant inadequacy in all of them. They sort the accuracy metrics into five categories: scale-dependent measures, measures based on percentage errors, measures based on relative errors, relative measures and scaled errors; describe each category and provide critical analysis of their weaknesses. Acknowledging inherent flaws of the existing accuracy metrics, they propose MASE. The metric was retroactively applied to the M-3 Competition data to test its potential.

The most important findings were that MASE can be used in all patterns of demand, that it produced results in accordance to what was found by Makridakis and Hibon (2000) about best-performing methods, and that MASE represented a more powerful test than any other metrics, since its results show more significant differences between forecast methods.

Finally, after considering the existing literature, Fildes et al. (2008) claim that "establishing an appropriate measure of forecast error remains an important practical problem for company forecasting".

2. Traditional Academic Measures of Forecast Accuracy

A starting point to discuss forecast accuracy measurement is that it is based on observation of errors. Those errors are comparisons between the demand that what was

forecasted for a given period of time and actual observation during that same time period. Therefore, the most basic idea about forecast accuracy is that a better forecast method is expected to produce smaller errors.

Furthermore, forecast accuracy can be considered a two-dimensional problem. One can think in terms of measuring accuracy over many periods of time for one item, while others may need a number that represents the goodness of forecast method for many items in the same time period. Table 1. exemplifies the generation of forecast accuracy values in both dimensions mentioned.

Table 1. The Two Dimensions of Forecast Accuracy

	Time									
	1			2			3			Mean of Absolute Errors
Items	f	a	Abs Error	f	a	Abs Error	f	a	Abs Error	per item
1	7	9	2	1	2	1	0	5	5	**2.67**
2	0	9	9	5	7	2	0	3	3	**4.67**
3	4	1	3	5	8	3	4	2	2	**2.67**
4	2	0	2	3	8	5	0	3	3	**3.33**
	Mean of Absolute Errors in time 1		**4**	**Mean of Absolute Errors in time 2**		**2.75**	**Mean of Absolute Errors in time 3**		**3.25**	

Mean of Absolute Errors is one of the existing forecast accuracy metrics. It can be calculated either at the line item level or at the aggregated level, for each period. In this case, the forecast method used performed better for items 1 and 4, while period 2 was the time in which the overall forecast accuracy was considered the best. Considering the scale dependency of that metric, discussed in the Chapter II, this hypothetic data set assumes that all line items have the same unit.

First, it is possible to isolate one time series, for example, the repeated demand for one item, and compute the accuracy along the time, which is called by Hyndman and Athanasopoulos (2014) as a type of time series cross-validation. Fildes et al. (2008) reinforce the importance of this process by claiming that forecasters should measure accuracy as a result of sequential errors.

One particular way to conduct such analysis is to calculate errors, for specific times, by comparing one period forecast and actual values. Afterwards, there is a variety of ways to combine errors and produce significant information about accuracy of forecast for that specific item.

However, Fildes et al. (2008) points out that "a common requirement, within an organization, is to provide a one-figure summary error measure, for many different time series" (p. 1158). That procedure is also known in literature as aggregation, which is both criticized and defended by many studies, like Jenkins (1982), Fildes and Makridakis (1995) and Hyndman and Koehler (2006).

In order to enable aggregation, Fildes and Makridakis (1995) affirm that errors must be standardized. In fact, Hyndman and Koehler (2006) applied scaled errors as a form of standardization, thus enabling aggregation by simple average.

Therefore, we infer that an effective measure of accuracy should be able to produce results for both dimensions. However, as we could not find any further discussion about the best way to aggregate accuracy values, hereafter, we are going to discuss a variety of metrics used to calculate forecast accuracy across time, which are exhaustively discussed in literature and often used by organizations.

To do so, we are going to present the most common accuracy metrics using the same taxonomy found in Hyndman and Koehler (2006). Basically, we review the many possibilities of handling the error, which is calculated as:

$$e_t = f_t - a_t \tag{2.1}$$

where:

e_t = forecast error at a given time

f_t = forecast value at a given time

a_t = actual value at a given time

a. Scale-Dependent Metrics

Metrics that fall in this category generate values accompanied by their respective units. Their use has to be restricted to series cross-validation in order to avoid the problem of mixing units of different items. That is the main source of criticism to M-1 Competition, in Makridakis et al. (1982), since it inappropriately uses the MSE across time series.

The most common scale-dependent measures are:

Mean Squared Error

$$MSE = Mean(e^2_t) \tag{2.2}$$

Root Mean Squared Error

$$RMSE = \sqrt{mean(e^2_t)} \tag{2.3}$$

Mean Absolute Error

$$MAE = mean|e_t| \tag{2.4}$$

Median Absolute Error

$$MdAE = median|e_t| \tag{2.5}$$

All equations in this category use central tendency measures. It is worth noting that *means* and *medians* are the extreme opposites in terms of sensitiveness to outliers. Hence, large errors will dominate the results in formulas based on means and cause almost no change in results of formulas based on medians. Therefore, in both cases the quality of the results are harmed.

Additionally, measures that use squared errors have the potential to penalize large deviations, in comparison to small ones, which make them appear attractive to some managers. However, their use was tested and not recommended by Armstrong and Collopy (1992) and Armstrong (2001), due to the disproportional harm caused by outliers.

b. Percentage Errors Metrics

Hyndman and Koehler (2006) define percentage error (p_t) by the following equation:

$$p_t = 100e_t / a_t \tag{2.6}$$

Means, medians and squares are applied to p_t to derive new forecast accuracy metrics. The most common percentage error measures found in literature are:

Mean of Absolute Percentage Error

$$MAPE = mean|p_t| \tag{2.7}$$

Median of Absolute Percentage Error

$$MdAPE = median\left|p_t\right| \tag{2.8}$$

Root Mean Square Percentage Error

$$RMSPE = \sqrt{mean(p^2{}_t)} \tag{2.9}$$

Root Median Square Percentage Error

$$RMdSPE = \sqrt{median(p^2{}_t)} \tag{2.10}$$

An inherent flaw with percentage error (p_t) is that it produces an infinite result when $a_t = 0$. Therefore, none of these metrics are recommended in data sets that contain actual demand values equal to zero.

Additionally, Tayman and Swanson (1999) state that "*MAPE* does not meet the criterion of validity, as it systematically overstates the average error of estimates, therefore, harming the degree of correspondence between its measures and actual values" (p. 299).

Furthermore, Makridakis et al., (1993) noticed that these metrics also penalize positive and negative errors differently because negative errors $(e_t < 0)$, in terms of inventory, are limited to the amount of the actual value (a_t), while positive errors $(e_t > 0)$ are unbounded. In order to deal with that, he defined symmetric measures:

Symmetric Mean Absolute Percentage Error

$$sMAPE = mean(200\left|a_t - f_t\right| / \left(a_t + f_t\right)) \tag{2.11}$$

Symmetric Median Absolute Percentage Error

$$sMdAPE = median(200\left|a_t - f_t\right| / \left(a_t + f_t\right)) \tag{2.12}$$

However, while Hyndman and Koehler (2006) found that these metrics reduced the unwanted effects caused by small actual demand values, it did not completely solve the problem. Moreover, some studies proved that these metrics are not as symmetric as they were supposed to be, Goodwin and Lawton (1999) and Koehler (2001).

c. Relative Error Metrics

These metrics are based on the division of an error produced by one forecast method, by the error of another forecast method, which serves as a benchmark method. Often, the benchmark forecast method consists of just a replication of previous period values, which Hyndman and Koehler (2006) define as random walk. That procedure is also known in literature as the naïve method Makridakis et al. (1993). Hence, relative error (r_t) is expressed by the following equation:

$$r_t = e_t / e^*_t \tag{2.13}$$

where, e^*_t is the error produced by the benchmark method, at time *t*.

The most common relative error measures are:

Mean Relative Absolute Error

$$MRAE = mean|r_t| \tag{2.14}$$

Median Relative Absolute Error

$$MdRAE = median|r_t| \tag{2.15}$$

Geometric Mean Relative Absolute Error

$$GMRAE = gmean|r_t| \tag{2.16}$$

Scrutinizing the relative error equation, we found that it is inherently flawed when the error produced by the benchmark method is zero and relative error goes infinite, or very small benchmark errors induce extremely high relative errors.

Regarding that issue, Armstrong and Collopy (1992) proposed a particular way to soften the mentioned effect by trimming results, the so-called *Winsorizing*. Basically, they attributed fixed values when benchmark errors are under or above certain thresholds. According to Hyndman and Koehler (2006), this procedure increases complexity and inserts arbitrariness.

d. Relative Metrics

Instead of simply dividing errors, these metrics are based on dividing results of one accuracy metric, regarding errors produced by different forecast methods. Therefore, Relative Mean Absolute Error is the division of MAE generated by one forecast method by MAE generated by a second method. Following are some of the possible metrics:

Relative Mean Absolute Error

$$RMAE = MAE_a / MAE_b \tag{2.17}$$

Relative Root of Mean Squared Error

$$RRMSE = RMSE_a / RMSE_b \tag{2.18}$$

Relative Median Absolute Error

$$RMdAE = MdAE_a / MdAE_b \tag{2.19}$$

Relative Mean Absolute Percentage Error

$$RMAPE = MAPE_a / MAPE_b \tag{2.20}$$

As the name of this group of metrics suggest, the results are given in relation to another forecast method. Hence, values from zero to one mean better forecast, compared to forecast method. When result is one, there is no significant difference among the considered forecast methods. Results bigger than one mean that forecast method used performed worse than the benchmark. Hyndman and Koehler (2006) consider the characteristic of easy interpretability as an advantage of these metrics.

The only limitation found is that it is impossible to use these metrics across items, regarding just one period in time, since they use scale dependent measures in numerator and denominator that do not allow aggregation of different time series.

Wheelwright et al. (1998) mentions a specific relative metric, called Theil's U Statistic and its variation, Theil's U-2 Statistic. Theil developed the first of those metrics in 1966, and it was modified into the second one in 1978. The article claims that Theil's U-2 statistic is just a particular case of RMAE, when the benchmark method is the naïve and forecasts are generated to one period ahead.

Another metric that uses the same principle of relative measures is PB. It is the percentage of times that one measure performs better than another, using any kind of the mentioned accuracy measures. Hyndman and Koehler (2006) mention two disadvantages of this metric. First, it is not sensible to the size of errors and second, it does not provide a clear idea of how much improvement is possible.

e. Scaled Error Metric

Hyndman and Koehler (2006) developed a new metric based on the principles of Relative Error Metrics and Relative Metrics. The rationale is to solve existing problems in the mentioned metrics by dealing with scaled errors *(q_t)*. The scaling factor, denominator of the scaled error, is the MAE of in-sample values of a benchmark forecast method.

The scaled error is defined by the following equation:

$$q_t = \frac{e_t}{\frac{1}{k}\sum_{j=1}^{k}\left|a_j - f_j\right|} \tag{2.21}$$

where,

j = sample time index

k = time index of the last in-sample observation

Hence, the error measured in a given time (e_t) is divided by the MAE of a benchmark forecast method, only considering the in-sample time period.

Hyndman and Koehler (2006) propose a particular type of scaled error, in which the benchmark is the naïve method. Because of that, the identity $f_j = a_{j-1}$ can be applied to adjust the equation. Moreover, they assume that the in-sample data comprehends periods from 1 to k. That makes the difference $a_j - f_j$ applicable from period 2 to k, as the first f_j value possible uses a_1 value. As result of that, there are *k-1* observations to be considered in the denominator of q_t .

Applying the mentioned adjustments, the following equation results:

$$q_t = \frac{e_t}{\frac{1}{k-1}\sum_{j=2}^{k}\left|a_j - a_{j-1}\right|} \tag{2.22}$$

After that, the Mean of Absolute Scaled Error is just given by:

$$MASE = mean\left|q_t\right| \tag{2.23}$$

The interpretation of results has to follow the same instructions as exposed for relative measures. The only case that scaled error equations do not work, is when all in-sample errors equal zero. We were also not able to find any negative critiques of this metric in literature, so because of these factors we choose *MASE* to be our metric of choice to compare against the accuracy metric proposed in the CIMIP. Additionally, in Chapter III, we present a further discussion on the importance of using benchmarks when measuring forecast accuracy.

3. Forecast Accuracy Metrics Currently Used in the Defense Environment

As part of CIMIP implementation, Office of the Secretary of Defense (OSD) established two metrics to measure forecast accuracy and forecast bias, while components already had their own ways to keep track of the goodness of their forecasts. This section aims to introduce the equations used by DOD and Navy, presenting brief comments about their main features.

a. DOD's Forecast Accuracy Metrics

The challenge with a common metric that is self-reported is to ensure that each group is calculating the metric correctly. To address this issue, the DOD published internal business rules to standardize the reporting effort among the components (DOD, 2013). As mentioned in Chapter I of our research, the CIMIP metric required specific data elements of the forecast and demand history, yet these business rules also detail what data should not be included. As stated in the introduction to the business rules document, the CIMIP "forecasting metrics are not the mechanism to reduce error; however the metrics will create a common baseline from which to measure the impact of other initiatives" (DOD, 2013, p. 2). The results of these forecast accuracy and bias metrics are

to be reported semi-annually at the DOD's inventory management reviews, as well as monitored by the CIMIP forecasting, total asset visibility, multi-echelon modeling working group and the supply chain metrics group (DOD, 2013).

The components are responsible for collecting all of the data necessary to compute the metrics, which should include all items for which the components use some type of forecast algorithm. This excludes items whose requirements determination is impacted by component business rules, performance-based contracts and foreign military sales. The metric also excludes unforecastable items, which either do not have a demand forecast rate, or whose forecast and actual demand during the reporting period is equal to zero. Although the components are free to generate forecasts with the method and time horizon of their choosing, they are required to insert 12-month forecasts and actual demands in the calculations.

The implementation of standard metrics to assess forecast accuracy and bias is one of the required actions, contained in CIMIP, to address the DOD need for better forecasts. From this point on, we are going to refer to those metrics as being $CIMIP_f$, aggregated forecast accuracy obtained at a given period of time, and $CIMIP_b$, forecast bias, as follows:

$$CIMIP_f = \left[1 - \frac{\sum_{i=1}^{n} c_i \left| f_i - a_i \right|}{\sum_{i=1}^{n} c_i a_i} \right] * 100\% \tag{2.24}$$

$$CIMIP_b = \frac{\sum_{i=1}^{n} c_i (f_i - a_i)}{\sum_{i=1}^{n} c_i a_i} * 100\% \tag{2.25}$$

where,

n = number of items in the forecast dataset

c_i = unit cost for item i

f_i = demand forecast for item i

a_i = actual demand for item i

A close look at $CIMIP_f$ metric reveals a certain similarity to *MAPE*, Equation (2.7). The first notable difference is the one minus before the fraction. It implies the rationale that accuracy is better when error is small and does not represent any harm to the interpretation of results. Another important difference is that $CIMIP_f$ is a division of summations, instead of a summation of divisions. Additionally, we assume that $CIMIP_f$, as an inventory forecast accuracy metric, uses unit costs to weight the importance of expensive items within the dataset and not as an evaluation of budget impacts.

As mentioned in the introduction, the accuracy metrics contained in CIMIP are the central issue of this research. Therefore, careful discussion and evaluation about those characteristics are presented in Chapter III.

b. Navy's Forecast Accuracy Metric

GAO criticized NAVSUP's secondary inventory management and recommended that it "evaluate and improve demand forecasting procedures," (GAO, 2008, p. 5). Then, a NAVSUP team developed the Lead-time Adjusted Symmetric Error (*LASE),* as their demand forecast accuracy metric, more than a year prior to the release of the CIMIP forecast accuracy metrics (Bencomo, 2010).

After determining that traditional accuracy measurements, such as *MSE* and *MAPE*, were insufficient, they combined two proposed solutions for calculating percentage-error for intermittent demand: *sMAPE* and Denominator-Adjusted MAPE (*DAM*) (Hoover, 2006).

The advertised benefits of the *LASE* metric were that it is capable to provide results with demand data that is highly intermittent, it does not generate a division-by-zero error, and it returns a symmetrical assessment of over and under forecasting. The equation follows:

$$LASE = \frac{\left| f_t - a_t \right|}{\left\lfloor \left(f_t + a_i \right) / 2 \right\rfloor + 1} \tag{2.26}$$

Actually, *LASE* equation is a combination of two aspects present in Hoover (2006). The first is that *sMAPE*, Equation (2.11), is a good way to measure forecast accuracy, when forecast or actual demand is different from zero. The second is that when forecast and demand are zero, managers should adjust the denominator by applying the addition of one. However, instead of applying the denominator adjustment only in cases that forecast and actual demand are both zero, the *LASE* metric applies the adjustment as a general rule. This characteristic aims to ensure consistency, as opposite to the use of different criteria for different items.

The following equation is a more consistent version of the *LASE* equation to the one proposed in Hoover (2006):

$$LASE' = I * \frac{\left| f_t - a_t \right|}{\left(f_t + a_t \right) / 2} + J * \frac{\left| f_t - a_t \right|}{\left\lfloor \left(f_t + a_t \right) / 2 \right\rfloor + 1} \tag{2.27}$$

being,

if $f + a = 0$, then $I = 0$ and $J = 1$;

if $f + a \neq 0$, then $I = 1$ and $J = 0$.

However, we consider the complexity of LASE' as a drawback, as well as its lack of criteria consistency, as different items are subjected to different rules.

One year after the metric was released, Jackson (2011) demonstrated that the *LASE* metric had an inherent smoothing effect that hampers the identification of large divergences between the forecast methods. By the end of the study, he concluded against of its use. Despite that, NAVSUP continues to utilize the *LASE* metric as an internal managerial tool to measure forecast accuracy.

C. CHAPTER SUMMARY

The forecast accuracy field of research has significantly evolved during the last sixty years, following the evolution of computing capabilities. Massive analyses and deep

considerations, in literature, provide relevant findings. From those, we highlight the following as the key learning points of this Chapter:

- There is no absolute best forecast method.
- More complex forecast methods do not always improve accuracy.
- Combining forecast methods will likely result in more accurate forecasts.
- Forecast accuracy can be measured across two dimensions: the first is time and the second is line items.
- Scale dependent metrics do not allow aggregation of results.
- Percentage error metrics are vulnerable to zero actual demand.
- Relative error metrics and relative metrics are vulnerable to the occurrence of any zero error.
- MASE avoids the flaws of many traditional metrics and remains in good standing among academic literature reviews.

Separate from the evolutionary process of academic literature on forecast accuracy, the DOD and Navy developed their own forecast accuracy metrics, respectively $CIMIP_f$ and *LASE,* in an attempt to quantify and improve their forecasting efforts.

III. ANALYSES ON CIMIP FORECAST ACCURACY METRIC

A. INTRODUCTION

This chapter will examine whether the current DOD forecast accuracy metric has any inherent flaws and if there are any alternative forecast accuracy metrics that avoid these flaws and produce higher quality, more relevant results.

B. EVALUATION OF CURRENT METRIC

At first glance, the $CIMIP_f$ metric, Equation (2.24), appears to be similar to *MAPE*, Equation (2.7), which as we discussed in Chapter II is a traditional forecast accuracy metric. The main difference between the two metrics is that *MAPE* is a *summation of divisions*, while $CIMIP_f$ is a *division of summations* that includes unit costs as a way to convert values to a common unit of measurement and prioritize the forecast performance of expensive items.

While *MAPE* is a broadly studied, traditional metric, it contains specific flaws that limit the scope of it applicability. In this section, we will investigate whether those differences, along with other specific characteristics, make $CIMIP_f$ a recommendable managerial tool to assess forecast accuracy.

1. Division of Summations

One of the main objectives of any forecast accuracy metric that utilizes division of a numerator by a denominator is to avoid unit-of-measure dependence in order to enable aggregation of results across a range of products. $CIMIP_f$, on the other hand, aggregates the results into dollars, by including unit costs, in both the numerator and denominator before the division occurs. This division of the total forecast error in dollars by the total actual demand in dollars produces a scale-free, dollar-weighted result.

To illustrate the methodologic difference, we compare $CIMIP_f$ metric to a cross-sectional extension of *MAPE*, in the manner that they determine their results. The equation for that variation of *MAPE* is:

$$MAPE_f = mean|p_i| \quad (3.1)$$

where;

$p_i = \frac{e_i}{a_i}$; and

$e_i = f_i - a_i$

$MAPE_f$ calculation first obtains the absolute percentage errors $|p_t|$ at the item level, then the individual results are averaged. $CIMIP_f$ first converts the numerator and denominator for each item into dollars, proceeds the summations the numerators and denominators separately, and then divides one by the other to generate a forecast accuracy result that represents the entire population. In this example we have adjusted *MAPE* to the aggregated level to enable comparison, yet we could have adjusted $CIMIP_f$ to the individual level to accomplish the same. Later, Equation (3.2) will present this extension of $CIMIP_f$. Table 2. and Table 3. provide an example of the methodologic distinction.

Table 2. *MAPE* Calculation

Items	$\mathbf{f_i}$	$\mathbf{a_i}$	$\mathbf{e_i}$	$\mathbf{\lvert p_i \rvert}$
1	23.84	32	-8.16	25.5%
2	21.26	20	1.26	6.3%
3	0	2	-2	100%
4	235.42	151	84.42	55.9%
			$MAPE_f$	**46.93%**

The far right column shows how *MAPE* first calculates individual absolute percentage errors and then averages them to get the final value.

Recalling $CIMIP_f$ metric:

Equation (2.24): $$CIMIP_f = \left[1 - \frac{\sum_{i=1}^{n} c_i |f_i - a_i|}{\sum_{i=1}^{n} c_i a_i}\right] * 100\%$$

Table 3. *CIMIP$_f$* Calculation

Items	f_i	a_i	c_i	$\lvert f_i - a_i \rvert$	$c_i * a_i$	$c_i * \lvert f_i - a_i \rvert$
1	23.84	32	$ 1,354,173.00	8.16	$ 43,333,536.00	$ 11,050,051.68
2	21.26	20	$ 43,125.00	1.26	$ 862,500.00	$ 54,337.50
3	0	2	$ 32,815.00	2	$ 65,630.00	$ 65,630.00
4	235.42	151	$ 260,000.00	84.42	$ 39,260,000.00	$ 21,949,200.00
				Sum	$ 83,521,666.00	$ 33,119,219.18
				CIMIP$_f$	**60%**	

The two far right columns of Table 2 demonstrate how *CIMIP$_f$* sums the numerator (total dollar error) and denominator (total dollar demand) separately before dividing them, subtracting from one and then multiplying by 100 to generate the final *CIMIP$_f$* value.

Moreover, as mentioned in Chapter II, *MAPE*'s results at the item level do not generate a solution when actual demand is zero. This division by zero error negates the ability to generate an average result, unless those non-solutions are ignored, which then degrades the entire accuracy measurement.

Meanwhile, *CIMIP$_f$* metric avoids that effect by applying a summation in the denominator to account for the fact that the data can include items with zero demand. Thus, *CIMIP$_f$* metric is able to produce valid results even when the data set contains values of zero for either the actual demand or forecast of individual line items.

Therefore, we claim that *CIMIP$_f$* metric is more robust than *MAPE.* The only case which *CIMIP$_f$* equation does not produce a valid result is when actual demands of all items considered are zero. Table 4. aims to provide evidence of the superiority of *CIMIP$_f$*, in terms of robustness, when compared to *MAPE$_f$*.

Table 4. Test of Relative Robustness of $CIMIP_f$ Compared to $MAPE_f$

Items	f_i	a_i	p_i	$\lvert f_i-a_i \rvert$	c_i*a_i	$c_i*\lvert f_i-a_i \rvert$	$\lvert f_i-a_i \rvert / a_i$
1	23.84	32	$ 1,354,173.00	8.16	$ 43,333,536.00	$ 11,050,051.68	25.5%
2	21.26	0	$ 43,125.00	21.26	$ -	$ 916,837.50	∞
3	0	2	$ 32,815.00	2	$ 65,630.00	$ 65,630.00	100%
4	235.42	151	$ 260,000.00	84.42	$ 39,260,000.00	$ 21,949,200.00	55.9%
				Sum	$ 82,659,166.00	$ 33,981,719.18	
				$CIMIP_f$	**59%**		**$MAPE_f$ ∞**

In this case, the actual demand of item 2 is zero, what harms the entire calculation of *MAPE*, while $CIMIP_f$ still produces a valid result. This supports the Hyndman & Koheler (2006) recommendation that MAPE should not be used in data sets that contain actual demands of zero.

2. The Role of Unit Costs

As mentioned, $CIMIP_f$ is calculated differently than the most traditional forecast accuracy metrics, as it implies that summations of forecast errors and actual demand values have to be made before the division, thus requiring the input data to be in the same unit-of-measure. In that context, unit costs are used as a means to standardize the units-of-measure of an items' demand, allowing the summations to occur in both the numerator and denominator.

In addition, the inclusion of unit cost also provides a weighting mechanism that prioritizes the accuracy of more expensive items over less expensive items. In the literature we reviewed, there is no mention of the use of weightings by the forecast accuracy metrics. All traditional equations are calculated around the forecast error, Equation (2.1), considering just two independent variables, forecast values and actual demands. The introduction of another independent variable such as unit cost, in the case of $CIMIP_f$, may affect the results. While measuring forecast demand error in dollars is a workable metric, the stated goal of $CIMIP_f$ is to produce a percentage measure of forecast accuracy.

Another point against the use of unit costs is that a secondary objective of $CIMIP_f$ metric is to avoid excess inventory and the related costs. One can think that organizations must avoid excess inventory of high unit cost items to reduce unwanted financial impacts. However, total inventory cost is composed of holding, transportation, handling,

acquisition and shortage costs. Of these five costs, only holding cost is directly affected by unit costs, and although positive correlations between unit costs and transportation, handling and shortage costs are possible, they are not certain. While cost is important to prioritize forecasting efforts, other factors such as criticality and interchangeability could also be considered. Acknowledging that unit cost is not the main driver for the total inventory cost or prioritization, we infer that forecast accuracy should be measured as a function of forecast and actual demand values.

To determine the positive and negative of using unit cost in the equation, we need to test to what extent it can significantly affect the interpretation of forecast accuracy. To do this, we built a test composed of four data sets, Table 5. through Table 8. , that keep forecast and demand values constant, while allowing the unit costs to vary:

Table 5. Test of Cost Impact on $CIMIP_f$ – Data Set 1

Items	$\mathbf{f_i}$	$\mathbf{a_i}$	$\mathbf{c_i}$	$\mathbf{\lvert f_i-a_i \rvert}$	$\mathbf{c_i*a_i}$	$\mathbf{c_i* \lvert f_i-a_i \rvert}$
1	90	100	$1,000.00	10	$100,000.00	$10,000.00
2	30	100	$50.00	70	$5,000.00	$3,500.00
3	50	100	$20.00	50	$2,000.00	$1,000.00
4	80	100	$250.00	20	$25,000.00	$5,000.00
				Sum	$132,000.00	$19,500.00
				$\mathbf{CIMIP_f}$	**85%**	

Table 6. Test of Cost Impact on $CIMIP_f$ – Data Set 2

Items	$\mathbf{f_i}$	$\mathbf{a_i}$	$\mathbf{c_i}$	$\mathbf{\lvert f_i-a_i \rvert}$	$\mathbf{c_i*a_i}$	$\mathbf{c_i* \lvert f_i-a_i \rvert}$
1	90	100	$50.00	10	$5,000.00	$500.00
2	30	100	$1,000.00	70	$100,000.00	$70,000.00
3	50	100	$20.00	50	$2,000.00	$1,000.00
4	80	100	$250.00	20	$25,000.00	$5,000.00
				Sum	$132,000.00	$76,500.00
				$\mathbf{CIMIP_f}$	**42%**	

Table 7. Test of Cost Impact on $CIMIP_f$ – Data Set 3

Items	f_i	a_i	c_i	$\lvert f_i-a_i \rvert$	c_i*a_i	$c_i*\lvert f_i-a_i \rvert$
1	90	100	$20.00	10	$2,000.00	$200.00
2	30	100	$50.00	70	$5,000.00	$3,500.00
3	50	100	$1,000.00	50	$100,000.00	$50,000.00
4	80	100	$250.00	20	$25,000.00	$5,000.00
				Sum	$132,000.00	$58,700.00
				$CIMIP_f$	**56%**	

Table 8. Test of Cost Impact on $CIMIP_f$ – Data Set 4

Items	f_i	a_i	c_i	$\lvert f_i-a_i \rvert$	c_i*a_i	$c_i*\lvert f_i-a_i \rvert$
1	90	100	$250.00	10	$25,000.00	$2,500.00
2	30	100	$50.00	70	$5,000.00	$3,500.00
3	50	100	$20.00	50	$2,000.00	$1,000.00
4	80	100	$1,000.00	20	$100,000.00	$20,000.00
				Sum	$132,000.00	$27,000.00
				$CIMIP_f$	**80%**	

$CIMIP_f$ results ranged from 42% to 85%, what may lead to diverse interpretations of forecast accuracy.

The results of this test demonstrate that the presence of unit cost in $CIMIP_f$ metric harms the quality of the item demand forecast accuracy measurement.

3. Production of Intuitive Results

$CIMIP_f$ uses two features commonly found in percentage equations. It first applies the complementary concept of "one minus the fraction", then it multiplies that fractional value by 100 to produce a percentage result.

However, percentage equations are expected to produce values between zero and one, which does not occur in $CIMIP_f$. The summation of errors, $CIMIP_f$'s numerator, can be higher than summation of actual demands, $CIMIP_f$'s denominator. That condition causes the fraction to be bigger than one and the final number to be negative and unbounded, which we consider counter-intuitive.

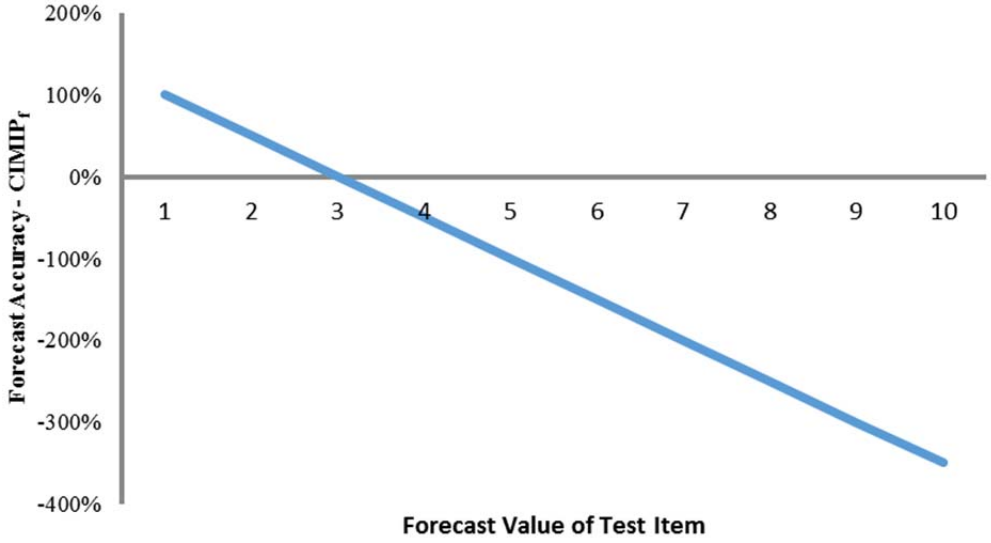
To demonstrate that, we built a test comprised of two hypothetical items, as follows:

Table 9. Generation of Counter-Intuitive Results - Initial Data Set

	f_i	a_i	p_i	$\lvert f_i - a_i \rvert$	$c_i * a_i$	$c_i * \lvert f_i - a_i \rvert$
Test item	1	1	1	0	1	0
Fixed item	1	1	1	0	1	0
				Sum	2	0
				$CIMIP_f$	**100%**	

By allowing the forecast value of the test item to vary from one to 10, we obtained:

Figure 6. Generation of Counter-Intuitive Results by $CIMIP_f$



Counter-intuitive, negative results are generated by $CIMIP_f$ in cases where the summation of errors is larger than the summation of actual demands. Considering the results at the item level, we infer that products with errors larger than actual demand may exert significant negative pressure on the aggregated $CIMIP_f$ result.

Furthermore, under-estimations are bounded by zero and all cases of forecast errors larger than actual demand only occur with over-estimations. That inherent characteristic of forecast errors helps all accuracy metrics to penalize the occurrence of

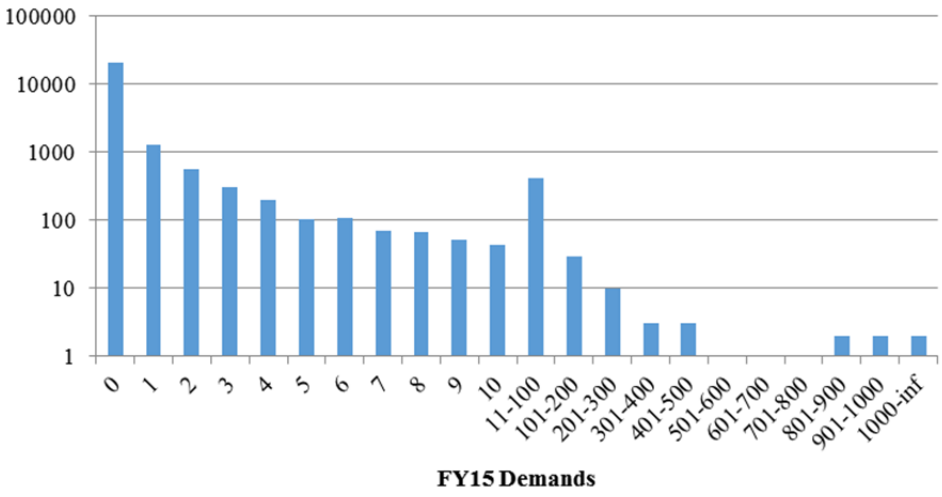
extremely large over-estimations, which are closely related to the formation of excess inventory.

4. Composition of Data Matters

If the probability of occurrence of errors, that are bigger than actual demand, is assumed to change along with demand size, then the composition of data may affect $CIMIP_f$ results. One can intuitively assume that low-demand items are more likely to have errors bigger than their actual demands. Considering that, if a data set is primarily comprised of low-demand items, a poor, or even negative, $CIMIP_f$ result is to be expected.

To validate the rationale that composition of data matters, first, we need to test the assumption that errors bigger than demand are more frequent in low-demand items. According to FY15 data, among 44,675 NIINs, 24,309 (54.41%) had errors bigger than demand and they were distributed according to the following histogram:

Figure 7. Histogram of Items with Errors Bigger than Demand in FY15



Vertical axle in exponential scale helps to picture the extreme skewness of the data.

Second, we divided the data into low-demand and high-demand items, according to a quantile approach, to compare $CIMIP_f$ results.

Table 10. *CIMIP$_f$* Results on Low-Demand Versus High-Demand Items (FY15)

	Dmd size	Qty of items	Dollar error	Dollar dmd	CIMIP$_f$
Low-demand	0-1	28,235	$555,585,938.00	$125,993,049.00	-341%
High-demand	2-inf	15,690	$2,022,307,097.00	$4,863,140,807.00	58%
Aggregate	0-inf	43,925	$2,577,893,035.00	$4,989,133,856.00	48%

There is clear evidence that low-demand items can exert a negative pressure on the overall *CIMIP$_f$* result.

Additionally, Table 11. shows that *CIMIP$_f$* results tend to be better as we only consider items with higher demand. The aggregate *CIMIP$_f$*, 48%, disguises the fact that for high-demand items the dollar-error is relatively small, while for low-intermittent demand items, the dollar-error relative to the actual dollar-demand is very large.

Table 11. Data Composition and *CIMIP$_f$* Variation (FY15)

Dmd size	Qty of items	Dollar error	Dollar dmd	CIMIP$_f$
0-inf	43925	$ 2,577,893,034.49	$ 4,989,133,856.14	48.33%
100-inf	546	$ 294,122,953.00	$ 941,957,920.00	68.78%
500-inf	90	$ 22,186,297.00	$ 79,316,220.00	72.03%
1000-inf	49	$ 13,291,976.00	$ 48,555,715.00	72.63%

We partially attribute those increasing *CIMIP$_f$* results to the fact that the errors bigger than demand are more unlikely as demand increases. But, on top of that, there is the fact that items with higher demand usually display a pattern that facilitates the generation of accurate forecasts.

Therefore, combining results of the three tests conducted in this section, we infer that the composition of the data set, expressed as a ratio of high and low-demand items, can significantly affect *CIMIP$_f$* results. The higher the ratio of low to high-demand items, the more likely the result will be a lower forecast accuracy measurement.

C. COMPARATIVE ANALYSIS

Considering the potential flaws of *CIMIP$_f$*, mentioned above, a comparative analysis is necessary to allow a judgment about the existence of a better metric. After

reviewing the existing literature, we selected an alternative metric and developed a framework to allow a fair comparison between the two metrics.

1. Alternative Metric Selection

As discussed in the literature review, *MASE* is intuitively expected to gather most of the desirable characteristics of a forecast accuracy measure, thus justifying its use as an alternate metric for comparison. Specifically, one of the main characteristics of *MASE* is the capacity to produce accuracy results at the item level, even when actual demand is zero, as well as at the aggregate level. Another important characteristic is that it enables a fair comparison among the services and DLA through its use of a benchmark method instead of generating absolute values.

a. Further Discussion on Performance Benchmarking

According to Dictionary.com, the word *benchmark* is "any standard or reference by which others can be judged" and the practice of using a benchmark to measure performance is widely practiced. An additional definition of the word is "a standard of excellence, achievement, etc., against which similar things must be measured or judged" (Ditcionary.com) and this idea of comparing similar things is key. Most people have heard a version of the phrase *comparing apples and oranges* and it applies to many areas where comparisons are made between two or more things. In our research we have discussed how DOD intends to measure the forecasting performance of the military services and DLA by calculating how well each of them generated forecasts for the material that they manage. While this exercise in measurement and comparison is intended to complement the goals of the overall CIMIP, it does not mean that we are making a true "apples to apples" comparison.

$CIMIP_f$ is simply computed by inserting forecasted demand, actual demand and unit cost for each item into the equation, which then produces one number. Although each service and DLA is engaged in managing secondary inventory, the material, quantity and demand patterns of this inventory are not the same. While they may appear similar and in some ways are, the fact is that they each face unique challenges in forecasting their demand and it is potentially misleading to directly compare their

performance. To illustrate this point with something that all federal employees are familiar with, we will examine the use of benchmarks by the Thrift Savings Plan (TSP).

On April 1, 1987, the TSP began operations with a single fund, known as the G Fund, which invested solely in government securities that were not available to the public. By 2001, the number of investment funds available in the TSP had grown to five with the inclusion of the fixed income F Fund, the common stock C Fund, the small capitalization stock S Fund and the international stock I Fund. Following common industry practice, since each of these four new funds were invested in securities available to the public, each funds' performance is compared against a commercial index made up of similar assets. These commercial indexes act as performance benchmarks for the funds. Since the TSP funds are modeled after these commercial indexes, a strategy known as passive-management, their performance does not vary much from the index. This common industry practice becomes more important with actively managed funds, where managers are attempting to outperform these commercial indexes. Table 12. shows the TSP fund with its respective index or benchmark and Table 13. compares the performance of the TSP funds against their benchmark index.

Table 12. TSP Fund and Benchmark Index. Adapted from Thrift Savings Plan (n.d.b).

TSP Fund	Commercial Benchmark
G Fund	N/A
F Fund	Barclays Capital U.S. Aggregate Bond Index
C Fund	Standard & Poor's 500 Stock Index
S Fund	Dow Jones U.S. Completion TSM Index
I Fund	MSCI EAFE Stock index

Table 13. TSP and Index Annual Returns 2011–2015. Source: Thrift Savings Plan (n.d.a).

Year	G Fund	F Fund	U.S. Agg. Bond Index	C Fund	S&P 500 Index	S Fund	DJ U.S. Completion TSM Index	I Fund	EAFE Index
2011	2.45%	7.89%	7.84%	2.11%	2.11%	-3.38%	-3.76%	-11.81%	-12.14%
2012	1.47%	4.29%	4.22%	16.07%	16.00%	18.57%	17.89%	18.62%	17.32%
2013	1.89%	-1.68%	-2.03%	32.45%	32.39%	38.35%	38.05%	22.13%	22.78%
2014	2.31%	6.73%	5.97%	13.78%	13.69%	7.80%	7.63%	-5.27%	-4.90%
2015	2.04%	0.91%	0.55%	1.46%	1.38%	-2.92%	-3.42%	-0.51%	-0.81%

This table demonstrates how an individual TSP funds' performance compares to a benchmark index, rather than a simple comparison to the other TSP funds.

The comparison to these benchmark index funds enables managers and potential investors to better judge the effectiveness of the TSP fund managers to meet their intended objective. For example, an S Fund investor should be satisfied with the management of his fund for all five years even though the C Fund had better returns in three of the five years. An apples-to-oranges comparison of the S and C Funds over these five years would conclude that the S Fund manager performed better in only two of the five years, while the C Fund manager performed better in three of the five years. An apples-to-apples comparison of these two fund managers would conclude that both of them matched or exceeded the performance of their benchmark index in all five years.

b. DOD Forecasting Benchmarks

The same principle of comparing investment fund performance to a relevant benchmark applies to the comparison of the services and DLA in their year-to-year forecasting performance. Concluding that one service forecasted better than another, based on a single $CIMIP_f$ metric result, ignores the fact that the lower-performing service may be managing material that is much more challenging to forecast than the higher-performing service. To date the DOD has resisted GAO recommendations to set standard forecasting performance goals, which could potentially result in apples-to-oranges comparisons. The DOD has stated that it wanted "to establish a baseline of performance on the metrics prior to setting any department-wide goals" (GAO, 2015b, p. 43), yet a department-wide goal, while simple, may not be as effective in measuring true forecast performance. An alternate method would be for each service to generate forecast

accuracy metrics for a naïve method forecast of their material and compare that with their actual performance. In keeping with our investment fund analogies, this method of evaluation is similar to how actively managed investment portfolios are compared against an index of similar assets.

The calculation of a naïve method simply requires the user to determine the level of demand for the preceding period and then assume that the demand will remain the same in the future period.

In order to exemplify the function of naïve method as a benchmark, Table 14. presents a set of three hypothetic items with different levels of demand variability, what is visualized in Figure 8. , along with their accuracy results, measured by four different metrics.

Table 14. Naïve Method as a Benchmark

Item 1

Time	Demand	naïve	Err	Abs err	Sq err	APE
1	10					0%
2	9	10	-1	1	1	11%
3	11	9	2	2	4	18%
4	10	11	-1	1	1	10%
5	9	10	-1	1	1	11%
6	11	9	2	2	4	18%
7	11	11	0	0	0	0%
8	10	11	-1	1	1	10%
9	9	10	-1	1	1	11%
10	10	9	1	1	1	10%

Stdev	0.8164966	MAE	1.11
Avg	10	MSE	1.56
CV	0.0816497	CIMIP	90%
		MAPE	10%

Item 2

Time	Demand	naïve	Err	Abs err	Sq err	APE
1	10					0%
2	6	10	-4	4	16	67%
3	14	6	8	8	64	57%
4	10	14	-4	4	16	40%
5	6	10	-4	4	16	67%
6	14	6	8	8	64	57%
7	14	14	0	0	0	0%
8	10	14	-4	4	16	40%
9	6	10	-4	4	16	67%
10	10	6	4	4	16	40%

Stdev	3.2659863	MAE	4.44
Avg	10	MSE	24.89
CV	0.3265986	CIMIP	60%
		MAPE	43%

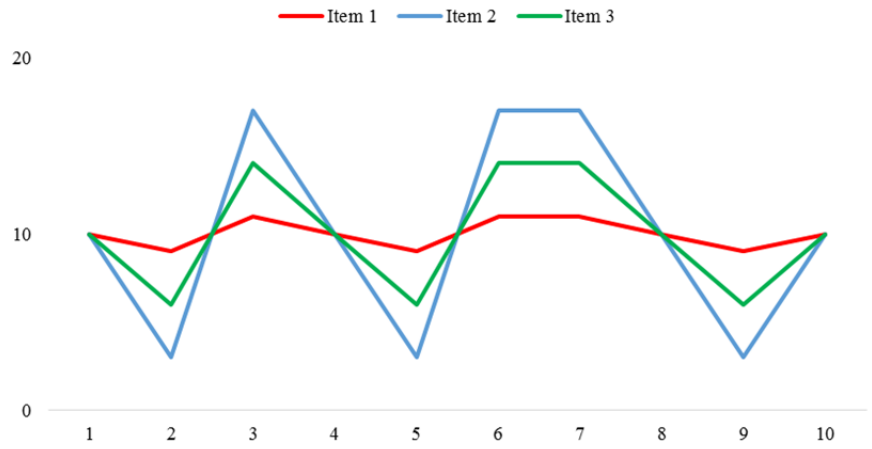
Item 3

Time	Demand	naïve	Err	Abs err	Sq err	APE
1	10					0%
2	3	10	-7	7	49	233%
3	17	3	14	14	196	82%
4	10	17	-7	7	49	70%
5	3	10	-7	7	49	233%
6	17	3	14	14	196	82%
7	17	17	0	0	0	0%
8	10	17	-7	7	49	70%
9	3	10	-7	7	49	233%
10	10	3	7	7	49	70%

Stdev	5.7154761	MAE	7.78
Avg	10	MSE	76.22
CV	0.5715476	CIMIP	30%
		MAPE	107%

All four accuracies of naïve forecasts are higher in item 1, which has the smallest coefficient of variability in the dataset. The opposite also holds as the worst accuracy results in all metrics were obtained in the item that has the highest coefficient of variability. Since this analysis is at the item level, we applied $CIMIP_i$*, Equation (3.3).

Figure 8. Different Levels of Variability



Items' demands were designed to provide clear understanding of existing different levels of variability.

According to the example, with naïve method, material with lower level of variability generates relatively accurate forecast, while material with higher level of variability generates relatively poor forecasts.

The summing of all of individual accuracy results, in a big set of items, should provide the user with a general idea of how difficult the population of material is to forecast. A large error signifies a difficult population, while a small error signifies a simple population.

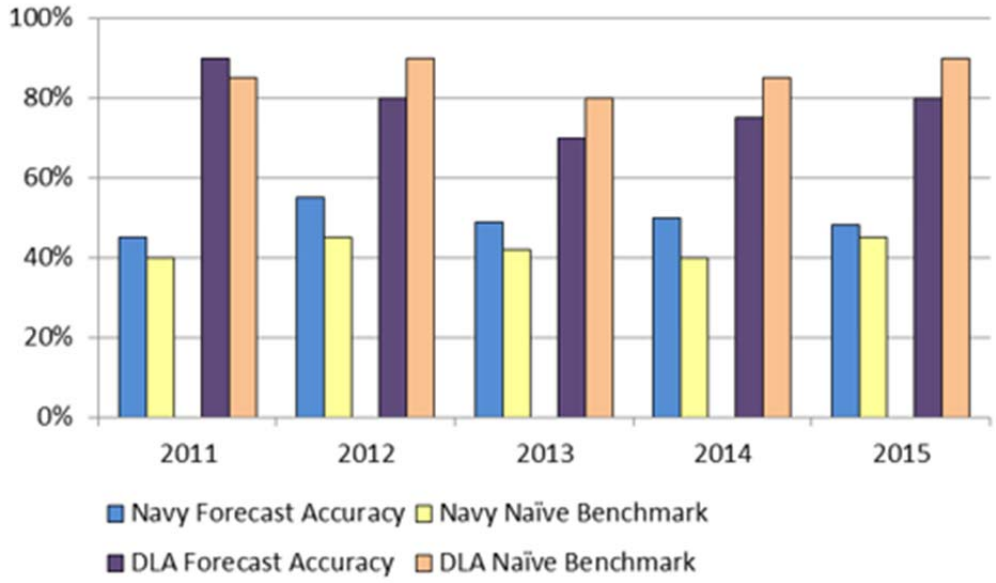
In the same manner that investors expect their asset managers to provide value greater than a passively-managed investment, so too should the DOD expect its material managers to generate forecasts that generally perform better than a naïve method benchmark. Table 14. and Figure 9. demonstrate how utilizing a naïve benchmark like this would give DOD leadership a better understanding of how well its components were actually forecasting. While the Navy is more interested in improving its own forecasting efforts, the DOD needs to be able to accurately assess the performance of all five reporting agencies.

Table 15. Theoretical Forecast and Benchmark Performance

Year	Army Forecast Accuracy	Army Naïve Benchmark	Navy Forecast Accuracy	Navy Naïve Benchmark	Air Force Forecast Accuracy	Air Force Naïve Benchmark	DLA Forecast Accuracy	DLA Naïve Benchmark
2011	30%	40%	45%	40%	55%	60%	90%	85%
2012	40%	35%	55%	45%	60%	75%	80%	90%
2013	10%	30%	49%	42%	64%	55%	70%	80%
2014	32%	25%	50%	40%	62%	65%	75%	85%
2015	40%	30%	48%	45%	70%	70%	80%	90%
Average	**30%**	**32%**	**49%**	**42%**	**62%**	**65%**	**79%**	**86%**

Numbers are fictional. This table demonstrates how naïve method benchmarks can bring forecast accuracy results into perspective, in a similar way that TSP fund performance is compared to a benchmark index.

Figure 9. Theoretical Chart Comparing Navy Versus DLA Forecasting Efforts (Numbers are Fictional)



This figure intends to demonstrate that if a manager considered forecast accuracy in isolation then they would conclude that DLA was outperforming the Navy, but if the manager was provided with benchmarks then they may reach the opposite conclusion.

2. Tests of Desirable Characteristics

We selected four characteristics regarded as relevant to any reliable forecast accuracy metric, as follows: sensitivity to volume heterogeneity, symmetry on error

treatment, robustness at individual and aggregated levels and allowance for a fair comparison.

In order to provide a means to a comparison between accuracy metrics, we designed particular tests to each one of the desirable characteristics. In the end of this section, we gathered results in a judgment table to point the best metric.

a. Sensitivity to Volume Heterogeneity

Assuming all items are of equal value, pure forecast accuracy aggregated metric must give equal importance to each item. Otherwise, if any kind of weight is applied to specific items, results can be seriously harmed. Since the impact of unit cost variation in $CIMIP_f$ has already been tested in this research, we still need to test whether its results are potentially dominated by large forecasts and actual demands. It is obvious that different items contribute different amounts to the overall $CIMIP_f$. But, since the item weight is composed of the demand volume and the unit cost, the degree to which high-volume items contribute disproportionately in any given dataset is an empirical question (again, assuming equal proportionality is what is desired). In this section, we test the relative sensitivity of $CIMIP_f$ and *MASE* to volume heterogeneity across inventory items.

We built a test, comprised of two fictional datasets per accuracy metric, to check the possibility of the generation of type I errors, saying the forecast is accurate when it is actually inaccurate, and type II errors, saying the forecast is inaccurate when it is actually accurate.

The first data set was designed to reflect a situation in which the forecast value is very close to the actual demand in one high-volume item, but the forecast model performs poorly in nine other low-volume items. In that situation, we should expect $CIMIP_f$ result to tell that the aggregated accuracy is low, thus the forecast method is performing poorly. Otherwise, type I error arises.

Table 16. High Volume and Type I Errors – $CIMIP_f$

Items	f_i	a_i	c_i	$\lvert f_i-a_i \rvert$	c_i*a_i	$c_i*\lvert f_i-a_i \rvert$
1	9000	10000	$ 1,000.00	1000	$ 10,000,000.00	$ 1,000,000.00
2	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
3	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
4	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
5	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
6	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
7	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
8	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
9	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
10	5	10	$ 1,000.00	5	$ 10,000.00	$ 5,000.00
				Sum	$ 10,090,000.00	$ 1,045,000.00
				$CIMIP_f$	**89.64%**	

Since there is not currently a DOD threshold for what constitutes an accurate forecast, we assume $CIMIP_f > 80\%$, to classify the forecast as accurate. The result of this data set is not aligned to the initial expectation of poor performance. Therefore, we state that the result led to a type I error.

The second data set aims to represent the opposite situation. A high-volume item has a poor forecast, while nine low-volume items have good quality on forecasts. In that situation, we should expect that $CIMIP_f$ result indicate a good forecast accuracy. Otherwise, a type II error is considered to occur.

Table 17. High Volume and Type II Errors – $CIMIP_f$

Items	f_i	a_i	c_i	$\lvert f_i-a_i \rvert$	c_i*a_i	$c_i*\lvert f_i-a_i \rvert$
1	5000	10000	$ 1,000.00	5000	$ 10,000,000.00	$ 5,000,000.00
2	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
3	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
4	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
5	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
6	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
7	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
8	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
9	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
10	9	10	$ 1,000.00	1	$ 10,000.00	$ 1,000.00
				Sum	$ 10,090,000.00	$ 5,009,000.00
				$CIMIP_f$	**50.36%**	

Using the same threshold of $CIMIP_f > 80\%$ to classify an accurate forecast, the result of this data set is also not aligned to the initial expectation of good performance. Therefore, we state that the result led to a type II error.

On the other hand, as *MASE* metric requires a slightly different type of data to be calculated. Hence, we created a very similar test, comprised of two other data sets that reflect the same situation as used to uncover the dominance of high-volume items in $CIMIP_f$ equation. Likewise, the same error-type definitions held true.

The first test, again, is the case in which nine low-volume items have relatively high forecast errors, while one high-volume item has a relatively low forecast error. In that arrangement, we should expect the result to tell a poor performance. Otherwise, we will consider the existence of type I error.

Table 18. High Volume and Type I Errors - *MASE*

Items	f_i	a_i	MAE of in-sample naïve	e_t	q_t
1	9000	10000	2500.00	1000	0.40
2	5	10	2.50	5	2.00
3	5	10	2.50	5	2.00
4	5	10	2.50	5	2.00
5	5	10	2.50	5	2.00
6	5	10	2.50	5	2.00
7	5	10	2.50	5	2.00
8	5	10	2.50	5	2.00
9	5	10	2.50	5	2.00
10	5	10	2.50	5	2.00
				MASE	**1.84**

Assuming a threshold of *MASE* < 0.8 to classify an accurate forecast, which is undoubtedly better than a naïve forecast, the result aligns with the initial expectation. Therefore, there is no evidence of type I error.

The second test is about the opposite situation, as nine low-volume items have good quality on their forecasts and one high-volume item has a poor forecast. We should expect a good accuracy result.

Table 19. Large Numbers and Type I Errors in *MASE*

Items	f_i	a_i	MAE of in-sample naïve	e_t	q_t
1	5000	10000	2500.00	5000	2.00
2	9	10	2.50	1	0.40
3	9	10	2.50	1	0.40
4	9	10	2.50	1	0.40
5	9	10	2.50	1	0.40
6	9	10	2.50	1	0.40
7	9	10	2.50	1	0.40
8	9	10	2.50	1	0.40
9	9	10	2.50	1	0.40
10	9	10	2.50	1	0.40
				MASE	**0.56**

Assuming the same threshold of $MASE < 0.8$ to classify an accurate forecast, the result aligns with the initial expectation. Therefore, we find no evidence of a type II error.

Considering the results of all four tests, it appears $CIMIP_f$ is less sensitive to volume heterogeneity than *MASE*, and hence, more likely to produce misleading results because of volume heterogeneity.

b. Symmetry on Error Treatment

As mentioned before, forecast errors in inventory demand data are bounded to the negative side, as result of underestimations, and unbounded to the positive side, as result of overestimations. However, forecast methods are expected to generate reasonable errors for the majority of items. Hence, we designed this test to verify whether equivalent variations of actual demand values, within a moderate range, to positive and negative sides, can result in different impacts for $CIMIP_f$ than *MASE*. Table 20. shows the initial arrangement of the test.

Table 20. Initial Dataset to Test Error Side Equality - $CIMIP_f$

Items	f_i	a_i	c_i	$\lvert f_i - a_i \rvert$	$c_i * a_i$	$c_i * \lvert f_i - a_i \rvert$
1	100	100	100	0	$ 10,000.00	$ -
2	100	100	100	0	$ 10,000.00	$ -
3	100	100	100	0	$ 10,000.00	$ -
4	100	100	100	0	$ 10,000.00	$ -
				Sum	$ 40,000.00	$ -
				$CIMIP_f$	**100%**	

Decision Variable: A1

Uniform distribution with parameters:

Minimum	0.00
Maximum	50.00

Decision Variable: A2

Uniform distribution with parameters:

Minimum	50.00
Maximum	100.00

Decision Variable: A3

Uniform distribution with parameters:

Minimum	100.00
Maximum	150.00

Decision Variable: A4

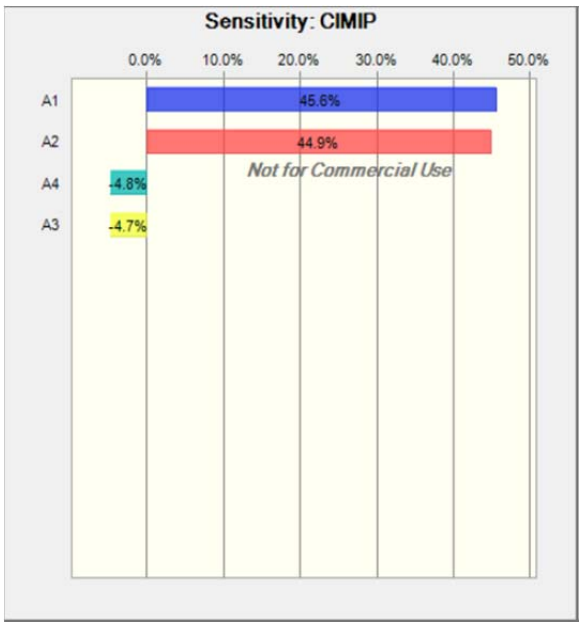
Uniform distribution with parameters:

Minimum	150.00
Maximum	200.00

Considering that the ranges of variation was designed to cause an equal proportion of positive and negative errors, an intuitive result should be that items with bigger errors on both sides would mostly contribute to $CIMIP_f$ variations.

However, according to Figure 10. , overestimations seem to impose a heavier pressure on $CIMIP_f$ results, compared to what underestimations do.

Figure 10. Sensitivity Chart of $CIMIP_f$ Equal Treatment Test



The equivalent test applied on *MASE* is shown in Table 21. .

Table 21. Initial Dataset to Test Error Side Equality - *MASE*

Item 1			
	FY13	**FY14**	**FY15**
ft	100	100	100
at	100	100	100
n	50	100	100
f_i-f_{i-1}	50	0	0
et	-	-	0
qt	0		

Item 3			
	FY13	**FY14**	**FY15**
ft	100	100	100
at	100	100	100
n	50	100	100
f_i-f_{i-1}	50	0	0
et	-	-	0
qt	0		

Item 2			
	FY13	**FY14**	**FY15**
ft	100	100	100
at	100	100	100
n	50	100	100
f_i-f_{i-1}	50	0	0
et	-	-	0
qt	0		

Item 4			
	FY13	**FY14**	**FY15**
f_t	100	100	100
a_t	100	100	100
n	50	100	100
f_i-f_{i-1}	50	0	0
e_t	-	-	0
qt	0		

MASE	**0**

Decision Variable: A1

Uniform distribution with parameters:

Minimum	0.00
Maximum	50.00

Decision Variable: A2

Uniform distribution with parameters:

Minimum	50.00
Maximum	100.00

Decision Variable: A3

Uniform distribution with parameters:

Minimum	100.00
Maximum	150.00

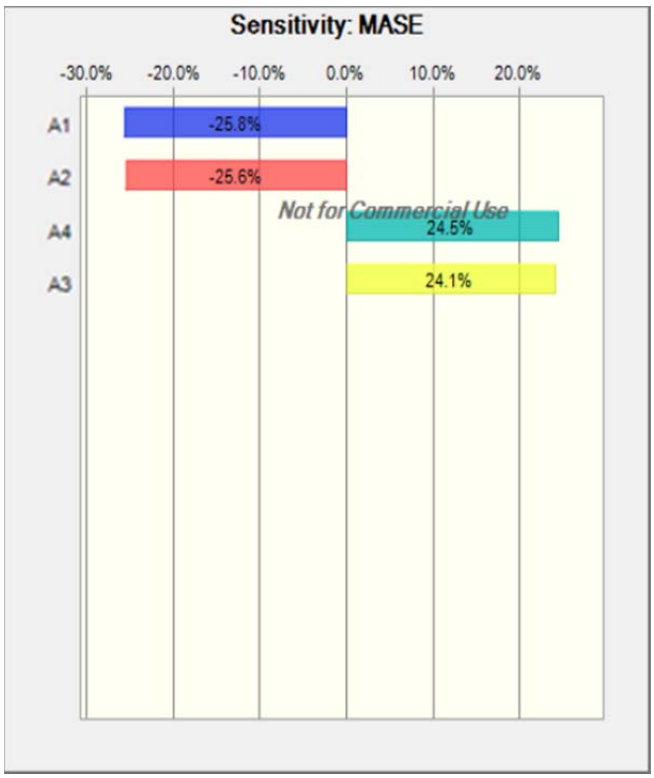
Decision Variable: A4

Uniform distribution with parameters:

Minimum	150.00
Maximum	200.00

Different than what happened to $CIMIP_f$, the sensitivity chart in Figure 11. shows that *MASE* gives balanced importance to errors in both sides.

Figure 11. Sensitivity Chart of MASE



c. Robustness at Individual and Aggregate Levels

Acknowledging the fact that no forecast method is expected to perform well in all situations, we agree with Fildes (1989) by stating that individual level analysis is more powerful for managers, as it enables to locate the origins of inaccuracy.

$CIMIP_f$ was initially designed and has been used to calculate an aggregate number that represent the overall forecast performance of each service. To do so, WSS has used twelve-month windows of data to allow calculations of total dollar-errors and total dollar-demands, the two key components of $CIMIP_f$ equation. As mentioned before in this chapter, $CIMIP_f$ is considered a robust metric at the aggregated level.

However, when the intent is to produce accuracy measures at the item level, WSS managers take out the summation signs and the unit cost from the $CIMIP_f$ original formula (E. Liskow, personal communication, April 4, 2016), resulting in the following:

$$CIMIP_i = 1 - \frac{|f_t - a_t|}{a_t} \tag{3.2}$$

We infer that this equation suffers from the same vulnerability as *MAPE*, Equation (2.7), which is returning an infinite value when actual demand is zero. In the specific case of Navy's demand data, the occurrence of zero demands are highly likely, as mentioned before.

In this research, we consider robustness as the ability to produce valid results, not undefined, in majority of situations, which is in accordance to Baker et al. (2006). Therefore, as $CIMIP_i$ returns invalid values in a significant amount of items in the Navy's dataset, the metric is classified as not robust.

However, a different approach is possible to improve the robustness of $CIMIP_i$ equation. Rather than taking the summation sign out, the Navy could sum forecast errors of one item, through the time. Unit cost is constant at the item level and is present in both summations of the fraction. Hence, they can be put in evidence and cancels out. After applying those adjustments, the proposed equation should be:

$$CIMIP_{i*} = 1 - \frac{\sum_{t=1}^{T} |f_t - a_t|}{\sum_{t=1}^{T} a_t} \tag{3.3}$$

That equation is only vulnerable to the specific case of all actual demands being zero, during the time considered. Therefore, as the time window increases, the probability of a zero value in the denominator is expected to reduce. Just as an example of the gain in robustness that this variation of the metric represents, when applied to a five year, quarterly demand dataset, $CIMIP_{i*}$ was able to return 100% of valid results, in contrast to only 52% of valid results of $CIMIP_i$ when applied to the FY15 demand dataset.

On the other hand, *MASE* metric was originally designed to be used in both dimensions of measurement, through the time and across the items, as used by

Hyundman and Koehler (2006). Moreover, the denominator vulnerability is related to the occurrence of all zero forecast errors, instead of all zero actual demands in $CIMIP_{i*}$, which is yet more unlikely to happen.

Therefore, we can state that *MASE* metric is potentially more robust than $CIMIP_{i*}$, although its gains are not perceived in the data considered, as the second could generate 100% of valid results.

d. Allowance for Fair Comparison

Forecast accuracy values are often used as a means of performance comparison. In that context, it is very important to set the ground for a fair comparison to occur. Non-relative metrics do not account for the fact that different datasets may comprise diverse amounts of variability that create different levels of predictability and makes the comparison in absolute numbers unfair. Therefore, comparisons of $CIMIP_f$ results at the aggregated and individual levels tend to be harmed by different levels of demand predictability in each dataset. *MASE*, conversely, uses naïve method as a benchmark to account for the level of demand predictability.

Table 22. helps to explain the difference in the interpretation of results.

Table 22. Difficulty to Forecast Test

	CV	$CIMIP_{i*}$	*MASE*
More Predictable	0.125494	92.23%	0.49
Less Predictable	1.937644	-0.001%	0.57

Values were calculated using data from two real items, picked as representatives of high and low coefficients of variation.

Considering a threshold of $CIMIP_{i*} > 80\%$ to classify an accurate forecast, only the forecasts of the "more predictable" item qualifies. To keep consistent, we applied a threshold of $MASE < 0.80$ to classify as an accurate forecast. By doing so, forecasts of both items surpass the requirement.

Based on this example, we see that if the forecast metric is to be used to compare accuracy of item forecasts (to compare IM's for example) *MASE* may do a better job

controlling for the underlying variability of the data, and present a better picture of the *relative* performance on each item (or by each IM). Of course, this is a simplification. *MASE* controls for only one source of variation: single period autocorrelation. Still, the point is that not all datasets are equally predictable, and caution should be used when comparing the accuracy of organizations managing different populations of material.

Extrapolating this result to the aggregated level, we can assume a hypothetical scenario of two datasets where one is mostly comprised of more predictable items and the other is mostly comprised of less predictable items. When measuring accuracy in absolute numbers, the results of the second dataset will more likely be worse than the first. Alternatively, *MASE* benchmarks performance against the naïve method, which enables the less predictable dataset to generate a relatively better result than the more predictable dataset.

D. CHAPTER SUMMARY

The main objectives of this chapter were to uncover evidences of inherent flaws of $CIMIP_f$ metric, through the application of specific tests, as well as to draw a comparison to an alternative metric, found in the literature.

The key lessons of the $CIMIP_f$ metric evaluation were:

- Type I and Type II errors are expected to occur;
- It can generate counter intuitive (e.g., negative) results;
- The composition of the data set (e.g., level of variability) influences its results.

Additionally, Table 23. aims to summarize the results of the tests contained on the comparative analysis.

Table 23. Ranked Comparison of *MASE* and $CIMIP_f$

Desirable Characteristics	*MASE*	$CIMIP_f$
Dominance of high-volume	1	2
Error side equality	1	2
Robustness at aggregate and individual levels	1	1*

Allow for comparability between items	1	2

This table ranks each desirable characteristic. * Grade attributed in case $CIMIP_{i*}$ is used.

In addition to demonstrating the theoretical problems with $CIMIP_f$, we compared it to another metric that has been highly recommended in the literature. Our comparison was based on the numerical analysis of a set of generated examples, which are not representative, so the generalization of the findings is problematic. Based on our test set, it appears that $CIMIP_f$ performs poorly relative to *MASE*.

IV. ANALYSES ON FORECAST PROCEDURES

A. INTRODUCTION

This chapter presents the calculations involved in the generation of a flexible forecast model rather than applying a fixed forecast method as a solution that fits all the items. The model uses a pool of forecast methods and forecast accuracy metrics, applied at the item level, as a means to optimize the selection of the forecast method to mitigate the expected error in forecasting.

B. BACKGROUND ON CURRENT NAVY'S FORECASTING PROCESS

NAVSUP is tasked with managing over 350,000 lines items (E. Liskow, personal communication, April 4, 2016) as they progress through six LCI categories. LCI's 1 and 2 cover the period from initial operational capability to the material support date when there is little to no historical demand data, while LCI 3 occurs during the demand development interval. LCI's 4 and 5 cover the periods when the weapon system program is mature and has been identified for retirement, while LCI 6 covers the period after the official retirement. The way the Navy forecasts demand is different throughout each of these LCI's, yet in this paper we will only focus on the forecasting procedures for LCI's 4 and 5. Currently, LCI 4 consists of approximately 284,000 lines items and LCI 5 consists of approximately 23,000 lines items (E. Liskow, personal communication, April 4, 2016); yet only about 40,000 of these lines items generate actual demand in a given year and meet the CIMIP definition of a forecastable item. The Navy utilizes a customized Enterprise Resource Planning (ERP) program to generate forecasts for all LCI 4 and 5 line items, yet not all of these forecasts will factor into the CIMIP forecast accuracy metrics.

In a broad sense, the forecasting process begins by segregating the global wholesale demand for the previous five years in to 20 quarterly buckets. It is important to note that this wholesale demand is not the retail, or end unit, demand, but rather the replenishment purchases made by the purchasing agents at the wholesale level. With

these 20 quarters of historical demand calculated for all LCI 4 and 5 line items, ERP runs an exponential smoothing with backcasting algorithm, utilizing a smoothing factor, or alpha (α), equal to 0.2. From these calculations ERP generates a constant quarterly forecast for the next five years. Since the forecasted demand is constant, it is sufficient to multiply one quarter by four to generate the annual forecasts for the next five years. This forecasting process is repeated every quarter in an attempt to capture demand changes in the items with higher variability. The forecasts generated by ERP are also subject to review by their IM who has the option to modify them as they deem appropriate. Upon completion of the IM review, the demand forecast is finalized and published for use in purchasing and other material management decisions.

The Office of the Secretary of Defense for Supply Chain Integration requires that each component report their forecast metrics semi-annually at the inventory management review. In April and October NAVSUP generates the Navy's official CIMIP accuracy and bias metrics by comparing the original forecast for the preceding 12 months with the actual demand during that period. Since the beginning of CIMIP metric reporting in FY13, NAVSUP has made attempts to improve their forecasting results by correcting erroneous data and identifying the specific line items with the most significant forecasting errors (E. Liskow, personal communication, April 4, 2016). While current capabilities have made it necessary to utilize a one-size-fits-all forecasting model, in the future they plan to enhance their ability to generate tailored forecasts for those items which the one-size-fits-all forecasting method produces inferior results (E. Liskow, personal communication, April 4, 2016).

C. OBJECTIVE OF THE MODEL

The mathematical model applied in this chapter aims to fill the existing gap between the current forecast process that uses a fixed method with fixed parameters and the desired stage of a tailored solution. The limitation of the model is that we arbitrarily chose the parameters to initiate the calculations, instead of using computational tools to optimize the choice.

As mentioned before in this research, DOD requests the generation of an accuracy number that is capable to represent the overall performance of the components in forecasting the items' demand in a given fiscal year. Those measures, combined with a certain threshold, aims to induce improvements in the components' processes of forecasting, what is expected to help in the effort of reducing the excess inventory.

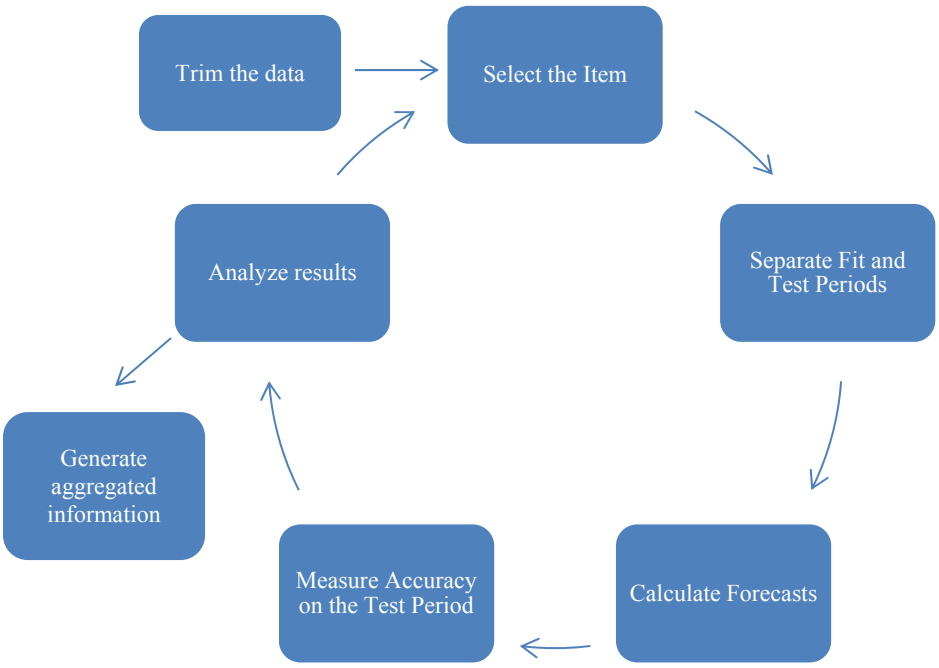
Additionally, we consider that the Navy's forecasters can benefit from the accuracy measures to improve their works. The idea is to use those measures as a means to identify relevant deviations and to help in deciding about the most effective way to generate the forecasts. Hence, from the perspective of forecasters, the information needed is slightly different. Rather than generating a number that represents the overall ability to produce accurate forecasts in a given period, a new approach should be the measurement of an item's accuracy, along the time.

We also acknowledge the fact that there is no absolute best forecast method, capable to generate the most accurate values for each one of the line items. Therefore, we designed a test that aims to test whether there are particular patterns of demand in which specific forecast methods tend to outperform the others. Moreover, we intend to present an aid for decision making, when a forecaster is dealing with an extensive and heterogeneous set of items' demands.

D. MODEL DESIGN

In order to generate the required information, we built a flexible forecast model, which selects each individual item, generates forecasts values using a pool of forecast methods and measures accuracy in a particular way to identify the forecast method that mitigates the forecast error. Once the whole data is trimmed, a cycle of events takes place in order to generate the intentioned information. Figure 12. shows the sequence of tasks involved in the model.

Figure 12. Model's Flow Chart



The following sections will describe the relevant tasks of the model.

1. Trim the Data

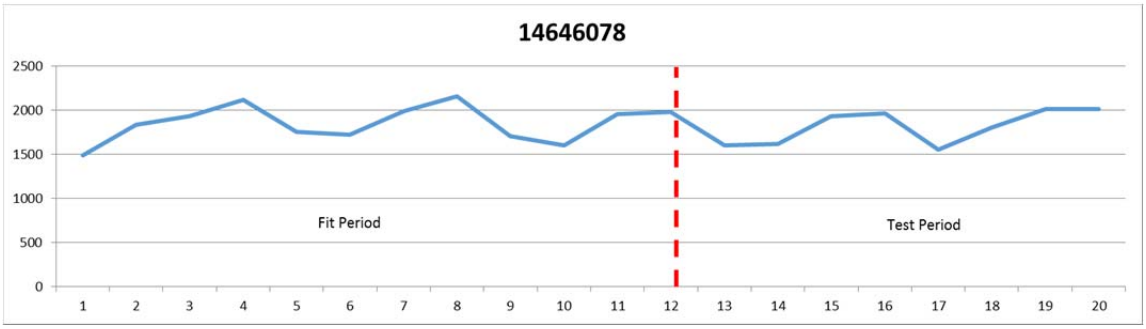
The original data set used to initiate the model comprehends five years of past demand of 80,427 NIINs. Demand data is grouped into 20 bins, each one representing a quarter of fiscal year. In order to allow the calculations of six different forecast methods and four different accuracy metrics, the items that did not meet the minimum requirements were withdrawn.

One limitation of the model used in this analysis is that one of the forecast methods and one of the accuracy metric are not able to generate valid results in all situations. In order to avoid invalid results, considered as infinite, the data set has to be trimmed to comprise only items that fulfill two conditions: variable demand in the first four periods and at least one demand of size bigger than zero in the last eight periods. Applying those conditions, 30,472 items remained out of a total dataset of 80,427 items.

2. Separate Fit and Test Periods

Following the procedure existing in Makridakis et al (1998), each item's demand is broken into two pieces. The first is called fit period and the second is called test period. The first set of data corresponds to the first 12 periods and is basically used to initiate the forecast methods. The second is formed by the demand on the subsequent eight periods and is used to test the difference between the forecast generated and the actual demand. Figure 13. present the demand and the two periods of a sample item in a visual form.

Figure 13. Fit and Test Periods



The blue curve shows the demand of item NIIN 01-464-6078. The dashed line in red is the break point of fit and test periods. Forecasts are generated from period 13 to 20 in order to allow comparisons to the actual demand.

3. Calculate Forecasts

Makridakis et al. (1998) define three categories of forecasts: quantitative, qualitative and unpredictable. All quantitative methods assume that the identified pattern of past demand is expected to hold in the future. Additionally, time series is the name of a family of forecast methods existing in the quantitative category.

Considering the fact that no item in LCI 4 and 5 is expected to generate demand shifts, trends or seasonality, we assumed that the demand pattern is stationary. Hence, our forecasting model comprises six of the simplest time series forecast methods found in literature.

We selected two averaging methods, two exponential smoothing methods, a combination of methods and the one that the Navy is currently using. We used the same taxonomy of (Makridakis et al., 1998) to present the methods, as follows:

a. Simple Average (SA)

This method averages all available demand data, according to the following equation:

$$f_{t+1} = \frac{1}{t}\sum_{i=1}^{t} a_i \tag{3.4}$$

where:

t = amount of available demand data at the moment that the forecast is generated.

Hence, as the variable *i* increases, the amount of available demand points also increases, making the *SA* to consider more data.

b. Moving Average (MA)

As opposite to what happens in *SA*, this method averages a fixed amount of the most recent demand data. The mentioned fixed amount of observations is called as order of average. The *MA* equation follows:

$$f_{t+1} = \frac{1}{k}\sum_{i=t-k+1}^{t} a_i \tag{3.5}$$

where:

k = order of average

The smaller the order of average, the more responsive to peaks and shifts in demand the method turns. For this research, we used a *MA* of order 12, the exact size of the fit period, as a mean to keep the method smooth.

c. Single Exponential Smoothing (SES)

In this method, the forecast is a function of the immediate past forecast, adjusted by the last forecast error.

$$f_{t+1} = f_t + \alpha(e_t) \tag{3.6}$$

where:

α = smoothing factor. It is a chosen fixed value between zero and one;

e_t = forecast error, Equation (2.1)

The forecast error is used to correct the past forecast value to the opposite direction, when calculating the next forecast. Hence, α plays the important role of weighting the importance of the last forecast error. Higher values of α makes the impact of last forecast error, on the next forecast, to be higher. As α values increases, the method turns more responsive, or less smooth. The opposite condition also holds, as lower α values imply a more smooth method. Hence, an α value can be calculated to optimize the results in a specific accuracy metric. However, when the value is found, it is used as a constant throughout the time, thus disregarding any possible change in demand pattern.

Finally, this method implies the use of two parameters, before initiating the calculations. The first is α and the second is f_1 value, from which all the subsequent forecast values and forecast errors are generated and adjusted. Although we acknowledge the possibility of finding optimal values of the two parameters, our forecast model fixes $\alpha = 0.1$ and $f_1 = a_1$.

d. Adaptive-Response-Rate Single Exponential Smoothing (ARRSES)

This method aggregates the idea of a flexible α to the *SES* method. Therefore:

$$f_{t+1} = f_t + \alpha_t(e_t) \tag{3.7}$$

where:

$$\alpha_{t+1} = \left| \frac{A_t}{M_t} \right|$$

$$A_t = \beta e_t + (1-\beta) A_{t-1}$$

$$M_t = \beta |e_t| + (1-\beta) M_{t-1}$$

β is a constant value between zero and one and relates to the degree in which α values are allowed to vary, along the time. The initialization of *ARRSES* comprises a bigger set of fixed parameters, as opposed to the *SES* that needs only f_1 and α values. Our forecast model considers the same parameters used by (Makridakis et al., 1993):

$f_2 = a_1;$
$\alpha_2 = \alpha_3 = \alpha_4 = 0.2;$
$\beta = 0.12;$
$A_1 = M_1 = 0$

e. Combination

As mentioned in the literature review, there is an expected gain in applying a combination of forecast methods, when all of them individually generate poor results. Hence, this method is just a simple average of forecast values obtained by the other four methods exposed thus far. The corresponding equation is:

$$f_t = \frac{1}{m} \sum_{x=1}^{m} f_{t,x} \qquad (3.8)$$

where:

x = method index

$f_{t,x}$ = forecast generated by the corresponding method for the index x, at time t

m = amount of methods to be combined

Therefore, our forecast model applies indexes from one to four to the previous methods, resulting in the use of *m = 4.*

f. Exponential Smoothing with Backcasting

This method is a variation of *SES*, in which the initialization value of f_1 is obtained by applying the inverse process of forecasting. This particular way to initiate the *SES* was studied and recommended by (Ledolter and Abraham, 1984) and is currently used by the Navy's ERP. Hereafter, we will refer to this variation of *SES* as the *NAVY* method.

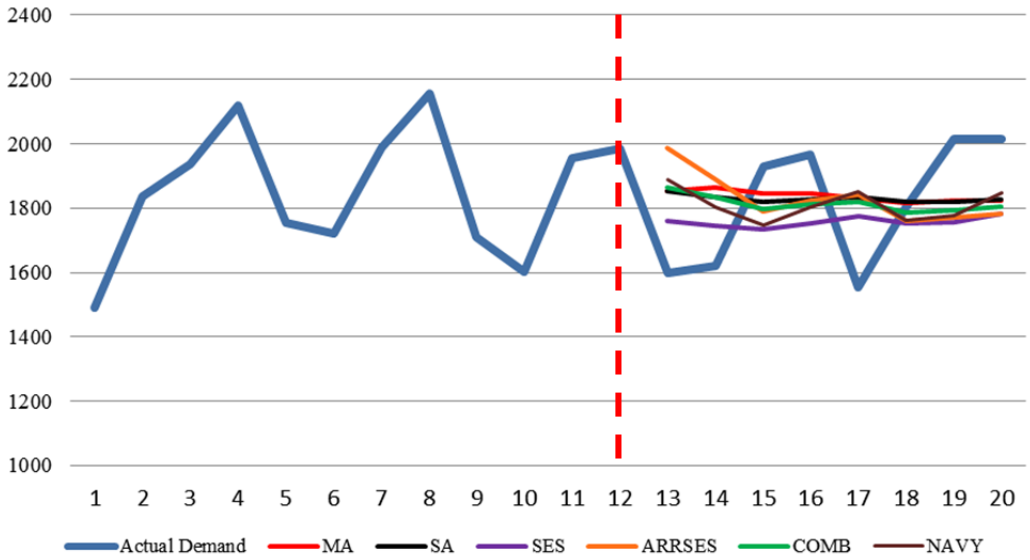
A short description of how the NAVY method follows: first, the condition $f_t = a_t$ is applied, meaning that the most recent forecast value equals to the most recent actual demand. Then, a fixed α value are applied to obtain backcast values for periods starting from $t - 1$ toward $t = 1$, as opposite to the generation of forecast value, which is calculated for the period $t + 1$. Our model applies the same smoothing factor as used by the Navy's ERP.

The process of generating backcasts is kept until the f_1 value is obtained. Thereafter, a regular *SES* forecast method can be initiated.

4. Measure Accuracy at the Item Level

After calculating all the different forecasts for the test period, some process has to take place to identify the most accurate method. The following chart aims to show the different forecasts generated and how difficult it can be to rank the methods by accuracy.

Figure 14. Sample of Forecast Generation



The vertical axel shows the demand sizes. The colored curves show the different forecasts generated for the same item exposed in the Figure 13. . It also shows that the differences in accuracy, among the methods, are not always visually identifiable.

In order to utilize a quantitative approach for the selection of the best forecast method for a specific item, we applied a pool of four accuracy metrics. All the accuracy metrics used in this analysis were discussed in detail in Chapters II and III.

First, we selected *MAE* and *MSE,* respectively Equations (2.4) and (2.2), as they are reported to be commonly used in real situations and can generate valid results when actual demands are zero. The fragility of generating numbers with units does not harm the result's quality at the item level. Additionally, we selected $CIMIP_{i*}$ *and MASE,* respectively Equations (3.3) and (2.23), because the first is currently used by DOD, to assess the component's performance, and the second is the alternative metric presented in Chapter III, while making the comparative analysis.

Table 24. summarizes the results of four forecast accuracy measurements for each one of the six forecast methods applied to a randomly selected sample item from the dataset.

Table 24. Summary of Accuracy Results

NIIN		Forecast Methods											
14646078		Simple average		Moving average		SES		ARRSES		Combination		NAVY	
Demand Description		MAE	174.59	MAE	171.72	MAE	182.20	MAE	218.27	MAE	185.01	MAE	195.23
Mean	1837.75	MSE	36796.09	MSE	36912.69	MSE	37129.76	MSE	57422.29	MSE	40122.25	MSE	43940.14
STD	196.8079146	MASE	0.79	MASE	0.78	MASE	0.83	MASE	0.99	MASE	0.84	MASE	0.89
CV	0.107091778	CIMIP	0.90	CIMIP	0.91	CIMIP	0.90	CIMIP	0.88	CIMIP	0.90	CIMIP	0.89

Highlighted in yellow are the accuracy metrics' choices of most accurate forecast methods.

5. Rank the Forecast Methods by Accuracy Metric

In order to identify the best and worst forecast method for any particular item, we generated rankings for each one of the accuracy metrics. *MAE, MSE and MASE* results are considered better when values are low. On the other hand, *CIMIPi** results are considered better as the values are high.

The following table considers the results exposed in Table 24. to form the rankings within each one of the accuracy metrics used.

Table 25. Ranking of Forecast Methods by Accuracy Metric

Simple average		Moving average		Simple Exponential Smoothing		ARRSES		Combination		NAVY	
MAE	2	MAE	1	MAE	3	MAE	6	MAE	4	MAE	5
MSE	1	MSE	2	MSE	3	MSE	6	MSE	4	MSE	5
MASE	2	MASE	1	MASE	3	MASE	6	MASE	4	MASE	5
CIMIP	2	CIMIP	1	CIMIP	3	CIMIP	6	CIMIP	4	CIMIP	5

For this particular item, using MAE as the selected accuracy metric, Moving Average is the forecast method that is expected to minimize the errors between forecast values and actual demand.

6. Count of Best Ranks

This analysis aims to investigate the skewness of best ranks distribution, considering the underlying methodologic differences of the four accuracy metrics mentioned. In other words, we test if a particular forecast method is considered the most accurate for the majority of items contained in the trimmed data.

7. Generate Overall Accuracy Ranking at the Item Level

We consider that the most accurate method to forecast demand, for the specific item considered, is the one that generates the lowest median of ranks, as shown in Table 26. and Table 27.

Table 26. Overall Ranks

	SA	MA	SES	ARRSES	COMB	NAVY
Overall rank	**2**	**1**	**3**	**6**	**4**	**5**

Table 27. Best and Worst Forecast Methods

Best Method	**MA**
Worst method	**ARRSES**

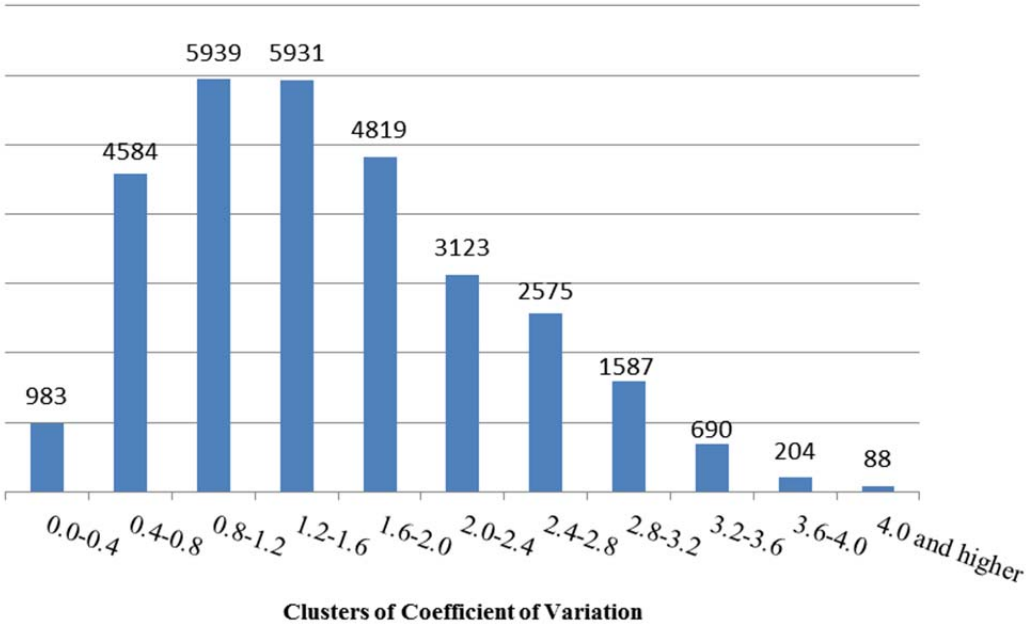
Considering all four accuracy metrics' results, Moving Average is considered the most accurate forecast method for this item, as it generates the lowest overall rank.

8. Build Clusters

In order to allow the investigation of the possibility of one forecast method to be capable of outperforming all the others for a specific group of items, we created 11 clusters of items, each one of those corresponding to a specific range of coefficients of variation *(CV)*. Hendricks and Robey (1936) explain the coefficient of variation as the ratio of the standard deviation of a number of measurements to their arithmetic mean. This ratio provides a standard for overall variability assessment since the number is scale free, and can be used to compare datasets.

The following histogram shows the *CV* clusters, along with the amount of items contained.

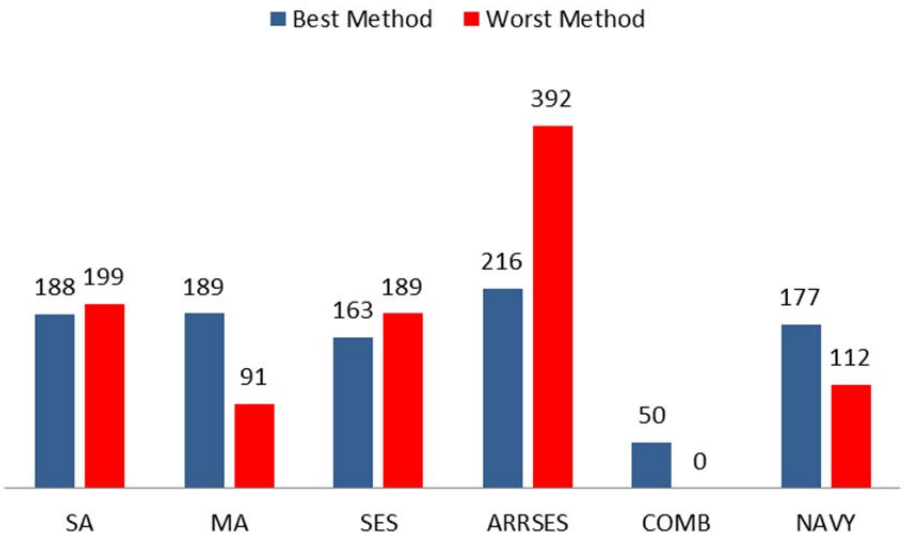
Figure 15. Histogram of Coefficient of Variation



9. Generate Rankings on Clusters of Coefficient of Variation

In order to elect the best forecast method for a specific cluster of coefficient of variation, we counted the number of items in which each of the forecast methods was considered the best and the worst option. The sample chart below shows how the rank results stored in a given cluster of *CV*.

Figure 16. The Best and Worst Methods Within a Cluster



Results collected in the *CV* cluster of 0.0-0.4. The vertical axel represents the amount of items, while the horizontal axel shows the forecast methods. In this case, *ARRSES* is the most frequently considered best and worst method. That information provides the idea of risk involved in the decision of selecting a specific forecast method.

10. Generate *MASE* Scores of Clusters

As a different approach to the use of ranks to track the performance of forecast methods, we calculated the average, minimum and maximum *MASE* values within each cluster of coefficient of variation. The intention is to identify a pattern of relative performance as the *CV* increases, compared to what naïve method produces. Moreover, those three values of *MASE*, measured along the time, provide the range of possible results to inform about the existing risk of choosing that specific method for the entire population.

11. Assess the Relative Performance of Navy's Forecast Method

We used the *MASE* accuracy metric in order to measure the potential gain of implementing different forecasting methods, instead of the Navy's status quo. First, we counted the percentage of items in which the *NAVY* method is not the best, meaning that there is opportunity to increase accuracy by using another forecast method.

Additionally, we counted the percentage of times that the Navy's forecast method performed worse than naïve; which means *MASE* values higher than one. Then, out of that, we counted how many times another method was capable of outperforming the naïve.

12. Measure the Level of Agreement between *MASE* and $CIMIP_{i*}$

In order to complement the comparative analysis conducted in the Chapter III, we measured the amount of times that rank results of *MASE* and $CIMIP_{i*}$ agree. The idea is to provide the magnitude of the existing theoretical difference among the metrics, using real data.

E. RESULTS

The model described is used to calculate forecast values, along with the respective accuracy scores as a means to identify the method that minimizes the expected error in each item. This section presents results grouped in to two categories: accuracy metrics and forecast methods. The first category utilizes real data to complement the theoretical comparative analysis among *CIMIP* and *MASE* accuracy metrics, conducted in Chapter III. The second utilizes accuracy measurements as a tool help forecasters in the task of optimizing the selection of a forecast method.

1. Accuracy Metrics

As mentioned, there are expected qualitative gains in choosing *MASE* as a substitute of *CIMIP* metric. As the comparative analysis used small sets of hypothetical items to demonstrate some characteristics of the metrics, a relevant question remained: do the results generated by the new metric represent a significant improvement?

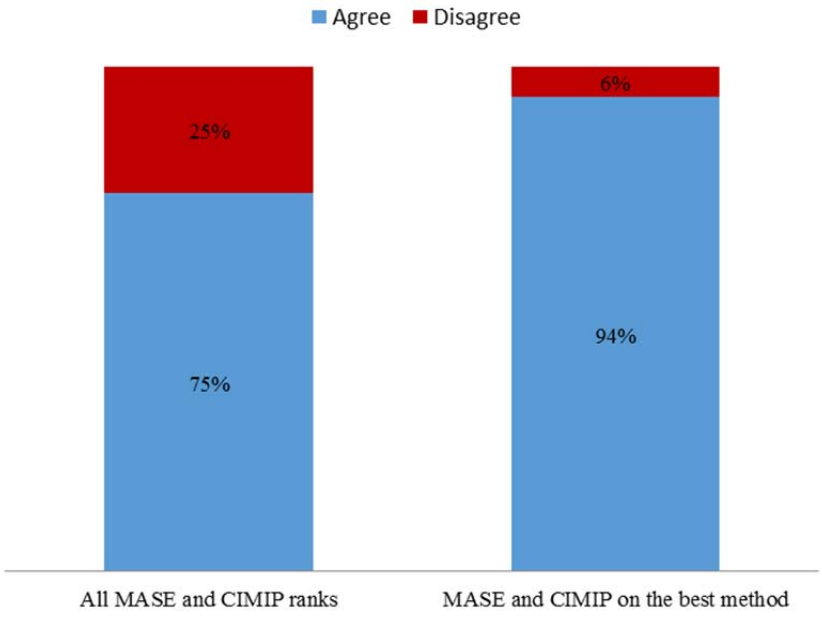
In order to answer that question, we have to consider that the current procedures do not formally involve the use of accuracy measures at the individual level. Components are just required to generate aggregated accuracy values to report to DOD as a representation of the overall forecast performance.

The Navy has tried to implement $CIMIP_i$ and *LASE*, respectively Equations (3.2) and (2.26), at the individual level as an internal effort to identify items that represent significant sources of inaccuracy within the big data. The vulnerabilities of those metrics were exposed in Chapter II, while Chapter III conclude that both *MASE* and $CIMIP_{i*}$ metrics, respectively Equations (2.23) and (3.3), can be used at the individual level.

However, the model presented in this chapter has a higher ambition on the use of accuracy metrics at the individual level. Assuming the generation of multiple forecasts per item in a given time, accuracy values can be used as inputs to support the decision of selecting the best forecast method.

Figure 17. shows the agreement level between *MASE* and $CIMIP_{i*}$ among themselves and with the overall rank generated. The agreement level can be explained by the percentage of times, considering all items, in which the results of two accuracy metrics lead to the same conclusion. This analysis uses ranks as the criteria to set a common ground for comparison among the accuracy metrics.

Figure 17. *MASE* and $CIMIP_{i*}$ Agreement



The first bar on the left represents the percentage of items that *MASE* and $CIMIP_{i*}$ results led to the exact same ranks for all six forecast methods used in the model. The second bar measures the agreement level on electing the most accurate forecast method.

When forming complete rankings of forecast methods, the methodologic difference between *MASE* and $CIMIP_{i*}$ led to the significant divergence of 25%. However, the main objective of the whole model is to provide useful information to optimize the selection of the most accurate forecast method for each item. For that matter, there is a high agreement level of 94% among the accuracy metrics.

2. Forecast Methods (Time Series)

We acknowledge the fact that parameters used to generate accurate forecasts in the past do not guarantee high performance in the future. However, based on the assumption of demand stationarity, we expect that the selection of the most accurate method in past data can result in improvements on future forecast performance.

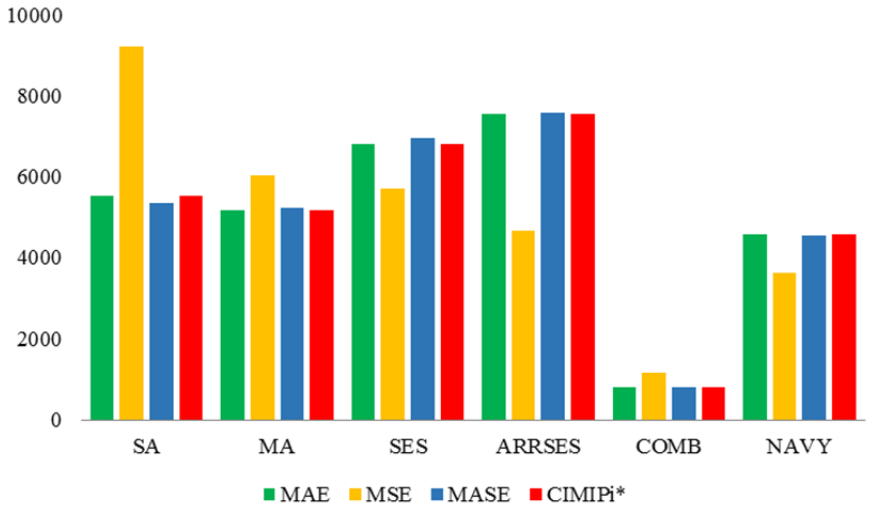
This section aims to uncover the existence of patterns that could be used to form a decision rule on the selection of the best forecast method. The tests were conducted under two main methods: analysis of ranks and *MASE* results analysis.

a. Analysis of Ranks

(1) Whole Population of Items

Considering the completely trimmed data, we first count the amount of items in which the forecast methods were considered the most accurate, by each accuracy metric. Results follow:

Figure 18. Count of Best Ranks by Accuracy Metric

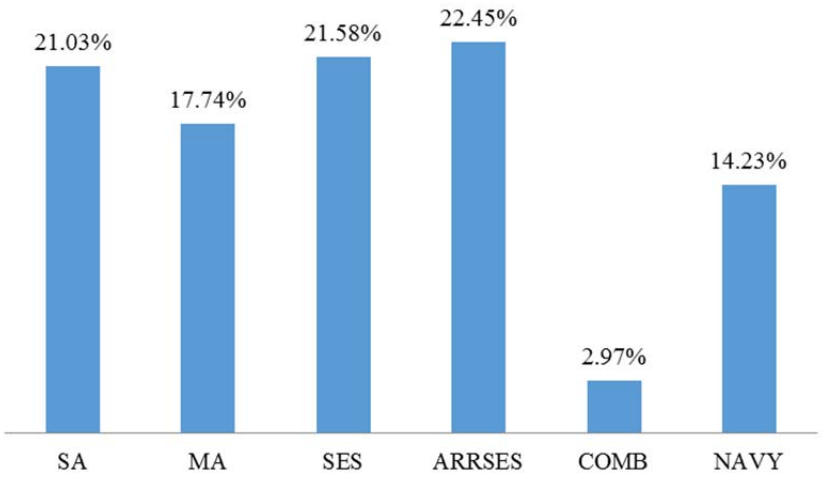


The vertical axel represents the amount of items.

There is no clear evidence, in the trimmed data, that one forecast method is mostly considered the best option. While *MSE* results are the most skewed toward *SA*, the other three accuracy metrics are slightly skewed toward *ARRSES*.

In order to enable a clear visualization of the overall skewness of ranks, among the forecast methods, we consolidated the counts of the four accuracy metrics. Results are shown in Figure 19.

Figure 19. Consolidated Percentages of Best Ranks

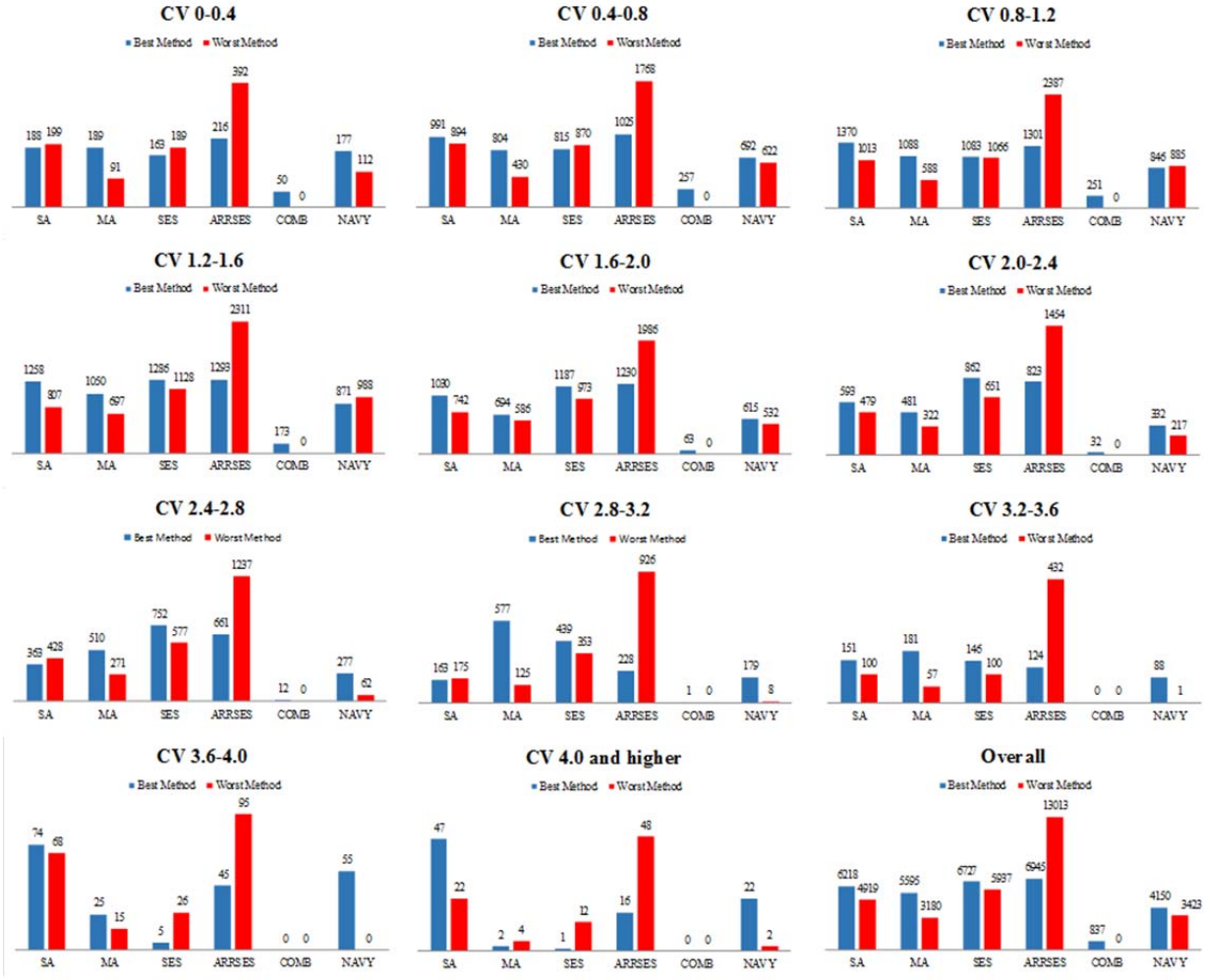


We found that there is no clear evidence, in the trimmed data used, to support that a particular forecast method is capable of outperform the others in a big majority of items. Hence, further analyses are needed to help in the decision of selecting the most accurate forecast method.

(2) Clusters of Coefficient of Variability

Rather than try to identify the most accurate forecast method for the whole population of items, the next analysis investigate the benefits of choosing a specific forecast method in groups of items that have similar demand behaviors, in terms of amount of variability. Hence, the following analysis applies a rank analysis, utilizing clusters of *CV* to group items and. Results follow.

Figure 20. Best and Worst Forecast Methods by Cluster



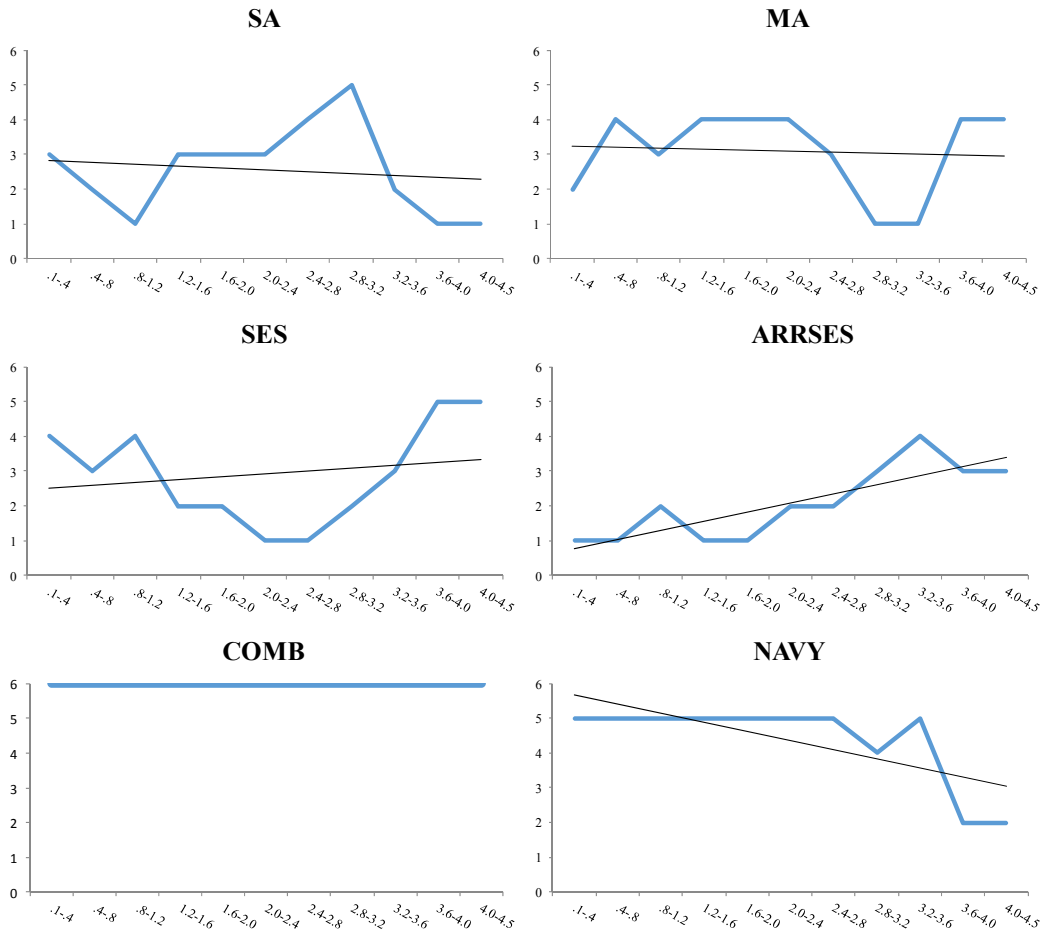
The relation between the blue and the red bars provides an idea of risk involved in the choice of one fixed method to be used in a whole cluster of items.

There is a pattern, along the clusters, of high risk in selecting one forecast method to be applied to the whole group of items. Just as an example, *ARRSES* was most elected best method, all clusters combined. At the same time, it was considered the worst option more times than all others. Hence, we can state that there is a significant risk of inaccuracy in choosing one method to be used in a cluster of *CV*.

Another relevant investigation is about the potential existence of upwards or downwards trends on forecast method ranks, as *CV* increases. Figure 21. shows how

each forecast method is ranked on the clusters of *CV*, based only on the amount of times it was considered the best option.

Figure 21. Average Rank Variation by Clusters



The vertical axes represent the aggregated rank, which is related to the number of times one method was considered the best option within each cluster. Trend lines are in black.

The *Combination* method shows a constant worst rank in all clusters of *CV*, that does not mean that it is the absolute worst method. What it does mean is that it is not often the best method, not considering the insignificant amount of items in which it was considered the worst method. Additionally, trend lines help to explain a significant amount of variance in results of two methods. *ARRSES* tends to lose rank as *CV*

increases, though not uniformly, while *NAVY* method, not uniformly, tends to gain ranks as variability increases.

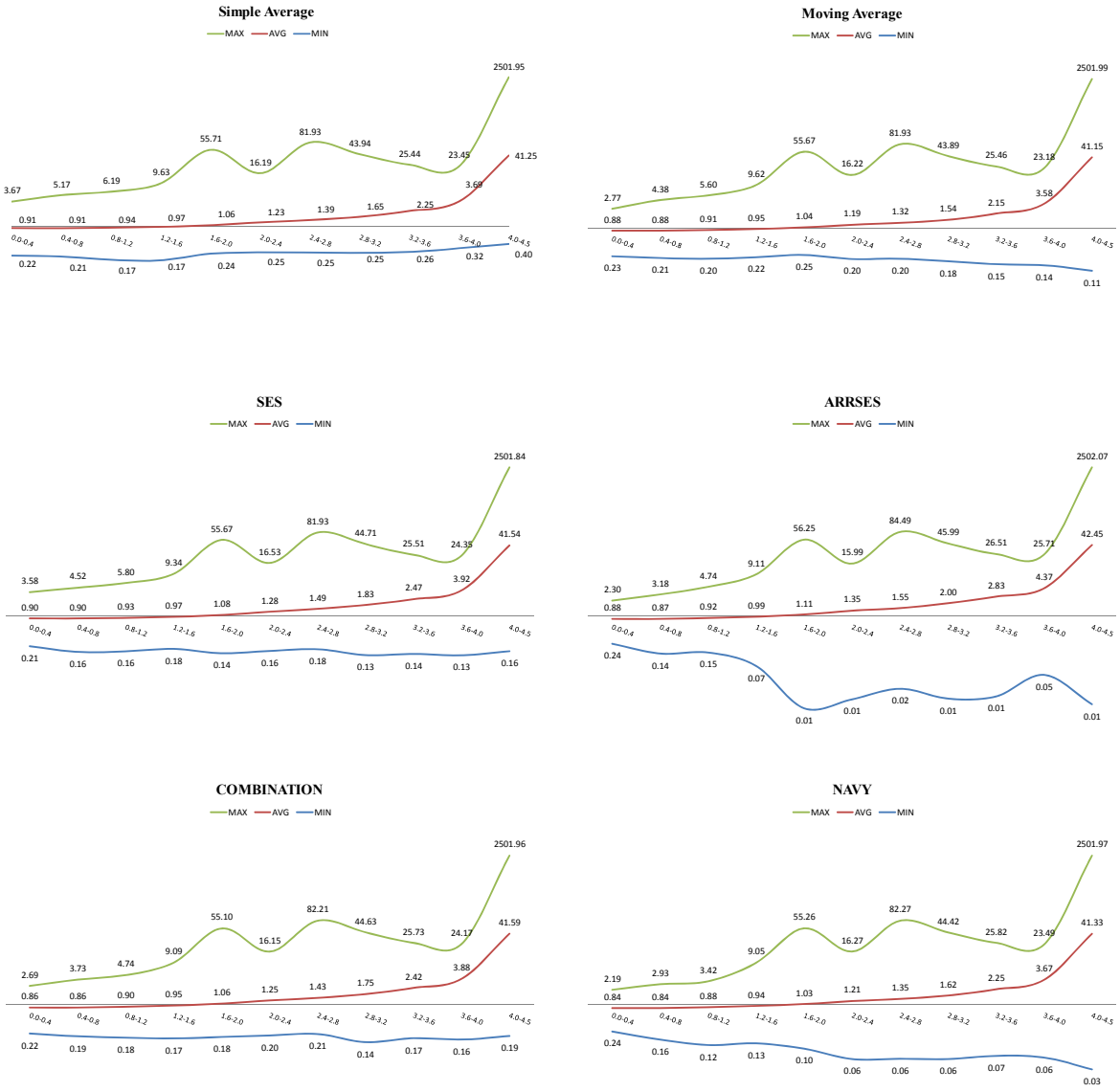
The drawback of the analysis of ranks is that it does not provide an accurate sense of differentiation between methods. As shown in Figure 20. , differences in counts of best rank, among the methods, sometimes are significant or clearly irrelevant. Therefore, analysis of ranks may distort the existing accuracy difference between the methods.

b. Analysis of MASE Results

In this section we analyze *MASE* results collected in the test period to select the forecast method to be used thereafter. The first analysis is set to investigate whether forecast methods behave differently as the coefficient of variation increases, in order to indicate the use of one for items with less variable demands and another for items with more variable demands.

Figure 22. shows how *MASE minimum, maximum and average* values of each forecast method change as *CV* increases.

Figure 22. *MASE* Values per Forecast Method



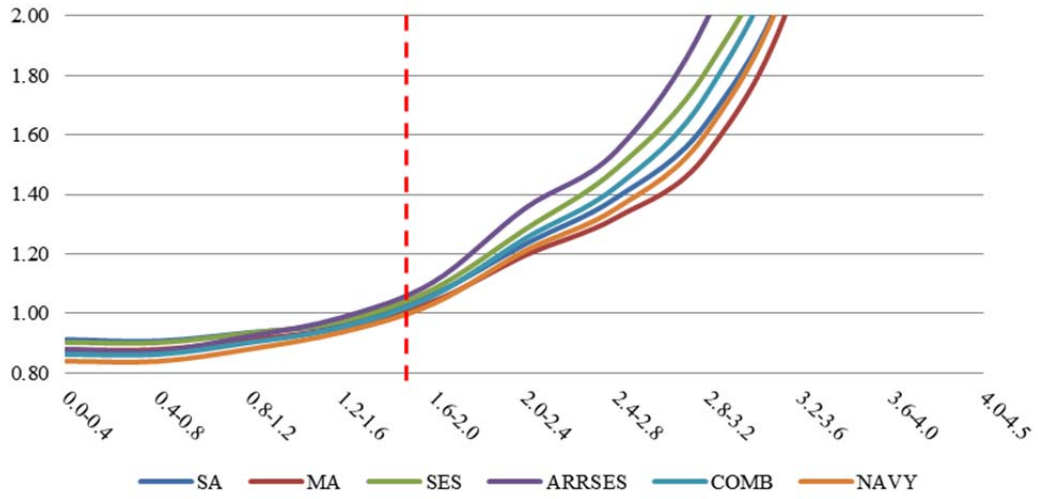
All the charts utilize exponential vertical axes to capture the entire spectrum of possible results. The horizontal axes correspond to the clusters of coefficient of variation. Maximum and minimum values provide the idea of the risk involved in the selection of the method as a fixed solution.

All forecast methods considered in the model generate similar shapes of maximum and average curves. However, *ARRSES* and *NAVY* methods are capable of generate the lowest minimum values, thus spreading the range of possible values by allowing significantly accurate forecasts at high values of *CV*.

The similarity of accuracy curves' shapes shows that there is no evidence that the selection of forecast method according to low or high variability will represent in any accuracy improvement. That similarity is partially explained by the fact that the forecast methods used in the model are classified as quantitative and time series. Hence, they are all based on the same assumption of demand stationarity, as they use historical data to predict future values. Furthermore, time series forecast methods can be considered responsive or smooth, depending on the parameters used. *SA* is a smooth method by nature, while the k, α and β values used respectively in *MA, SES* and *ARRSES*, made them behave as smooth methods as well. *Combination* method can also be considered smooth as it averages the forecasts of previous four methods. *NAVY* method is the most responsive in the model, as it uses $\alpha = 0.3$.

Figure 23. shows the six *MASE* average curves together, corresponding to the forecast methods applied in the model, to evidence the similarity in terms of forecast accuracy values.

Figure 23. Average *MASE* Results by Forecast Method



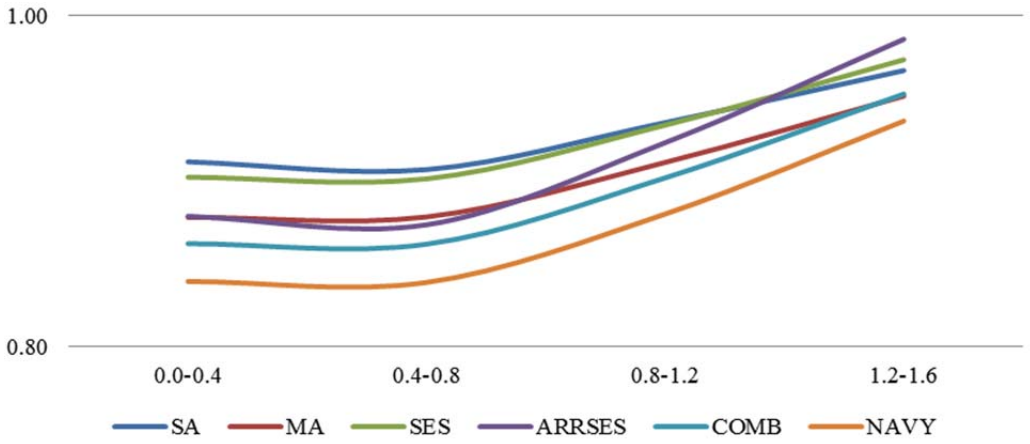
The vertical axel is comprised by *MASE* values and was intentionally cut at 2.0, as the values continue to increase and values higher than 1.0 are considered worse than naïve method. For low *CV* values, accuracy results are similar, but they tend to diverge as *CV* increases.

In order to optimize the quality of forecast results, we can apply the average *MASE* value of 1.0 as a threshold to consider that the use of one specific forecast method is recommendable, because it is capable of outperforming the naïve method systematically. For items with higher values of *CV*, deeper attention is needed to support the forecasting process.

Applying that threshold, we found that none of the forecast methods used in the model has systematic superior performance than naïve method for *CV* values higher than 1.6, while all of them can outperform, on average, the naïve method for *CV* values lower than 1.6. Hereafter, we will refer to the range of $0 < CV < 1.6$ as the "selected data".

Figure 24. shows the same results as in Figure 23. , but in a different scale, as its *MASE* values are limited to 1.0.

Figure 24. Average *MASE* Results in the Selected Data



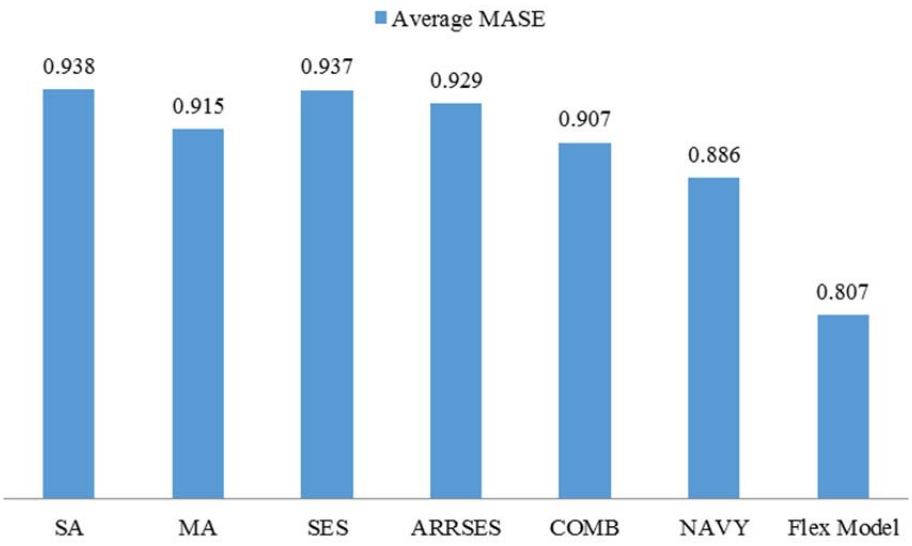
Within the range of *CV* that the forecast methods can be used to outperform the naïve benchmark, *NAVY* method is systematically considered the best option.

Although the *NAVY* method had better performance in all clusters of *CV* in the selected data, we identified a risk in using a fixed forecast method for a group of items. Hence, we investigated the potential benefit on accuracy when the most accurate forecast

method is selected for each item, what we call as *Flexible Method*, instead of working with a fixed method.

Figure 25. shows that the adoption of *Flexible Method* in the selected data resulted in a significant gain of accuracy, when compared with each one of the forecast methods applied individually.

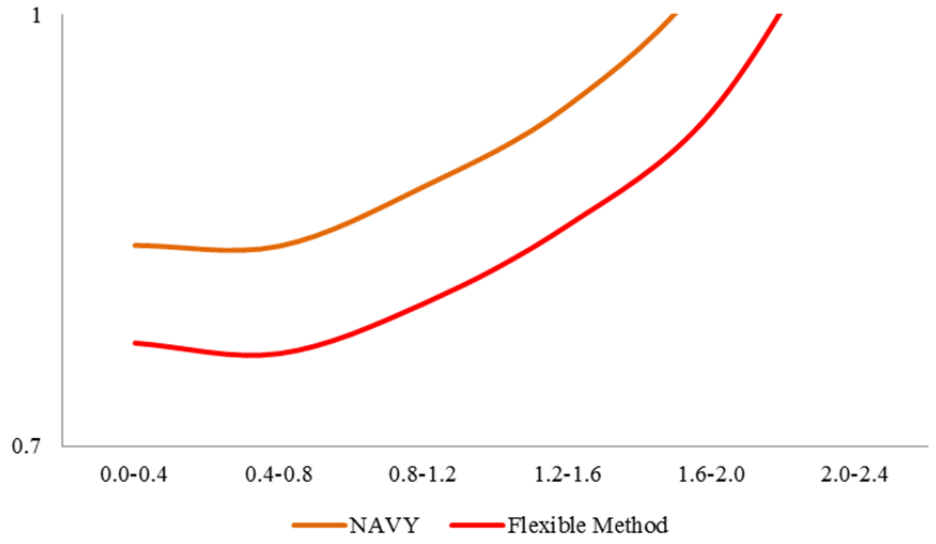
Figure 25. Accuracy Gain of Flexible Method



The bars represent the average of *MASE* results for items with *CV* < 1.6.

Additionally, Figure 26. shows that the *Flexible Method* not only has superior accuracy than the *NAVY* method, which was considered the most accurate among the six methods applied in the selected data, but it is capable of extending the range of *CV* in which it can be used to systematically outperform naïve benchmark.

Figure 26. *MASE* Values of *NAVY* and *Flexible Method*



Flexible Method resulted in a significantly superior accuracy in all clusters of *CV* in the selected data. Additionally, it generates average *MASE*<1 for the *CV* cluster (1.6-2.0), what extend the overall range of *CV* values in which the use of time series forecast methods is expected to outperform naïve the method.

Therefore, considering the data used, the adoption of *Flexible Method* represented a significant gain in forecast accuracy as well as an extension in the number of items that time series forecasts were considered recommendable. Implementing our findings, all six time series forecast methods, if applied as a fixed solution, were recommendable for 17,437 items, what represents 57.22% of the trimmed data. Meanwhile, utilizing the same criteria, the *Flexible Method* is considered recommendable in 22,256 items, thus representing 73.04% of the trimmed data.

F. CHAPTER SUMMARY

After applying a model that calculates demand forecasts and accuracy values in all items' data, the most significant findings were:

- Despite the methodologic differences and theoretical superiority of *MASE* over $CIMIP_{i*}$, both generated a very high level of agreement, while selecting the most accurate forecast method;
- The calculation of forecast accuracy can be used by the forecasters as a managerial tool, instead of just fulfilling the need of reporting;

- In order to provide information that helps to improve the forecasting processes, accuracy has to be calculated at the item level;
- All forecast methods applied in the model tend to be less accurate than the naïve method, as *CV* increases;
- Using averages of *MASE* values, the *NAVY* was considered the most accurate of all six forecast methods used in the model for all clusters of $CV < 1.6$.
- The use of *Flexible Method* resulted in a significant gain of accuracy, when compared to any of the other forecasting methods applied individually.

V. FINDINGS, RECOMMENDATIONS AND FUTURE RESEARCH

A. FINDINGS

While many of our findings throughout this research are detailed at the point of discussion, the major findings of our research in regard to $CIMIP_f$ and forecast accuracy measurement are summarized below for ease of access.

1. $CIMIP_f$ Weaknesses

$CIMIP_f$ is not able to produce accuracy results for individual line items when the actual demand for that item during the period, usually one year, is zero. This complicates the individual line item assessment of forecast accuracy, since $CIMIP_f$ returns an invalid division-by-zero result. This weakness does not prevent the aggregation of results for multiple line items because of the summation that occurs in the denominator prior to the final calculation.

$CIMIP_f$ results are significantly affected by the unit costs that are included in both the numerator and denominator of the equation. The inclusion of unit cost as an independent variable in $CIMIP_f$ detracts from the primary purpose of measuring forecast accuracy performance.

$CIMIP_f$ produces aggregated results that are not inherently intuitive and are disproportionately affected by over-estimations. This is especially evident with low demand items where the possibility of the size of the error exceeding actual demand is greater. We found that the aggregate $CIMIP_f$ for 28,235 low demand items produced a large negative result (-314%), while the aggregate $CIMIP_f$ for 15,690 high demand items produced a modest positive result (58%). As another example of the effect of unit cost, due to the high dollar weighting for the high demand group the total $CIMIP_f$ result was 48%.

$CIMIP_f$ does not consider the difficulty of accurately forecasting the entirety of material that the services and DLA are charged with managing. Its lack of a benchmarking function, similar to the one found in *MASE*, results in $CIMIP_f$ directly

comparing the forecasting performance of the services and DLA against each other. Although we did not compare the performance of the Navy versus DLA, without consideration of performance benchmark, the services could be penalized for what is considered to be poor performance or incentivized to make risky decisions in an effort to improve forecasting performance.

2. Forecast Accuracy

There has been significant study on the topic of forecast accuracy within the academic world. Among a large amount of forecast accuracy metrics currently available in literature, *MASE* was considered useful and theoretically superior than all variants of *CIMIP*.

From the perspective of IM's at WSS, the measurement of accuracy at the item level generates more value than one aggregated accuracy number, as currently required by DOD.

Item accuracy measurements enable a better identification of poorly forecasted items and can also be applied as a managerial tool for determining which forecast method to utilize.

3. Demand Forecasting

The task of demand forecasting within the DOD is very complex because demand patterns are significantly heterogeneous. Using *MASE* as the forecast accuracy measurement, we found that the Navy's preferred forecasting method, on average, out-performed the other five methods when compared to the naïve method and when CV was less than 1.6. Additionally, flexibility in the choice of forecasting method at the individual item level, enabled our test data to outperform the naïve method when CV was less than 2.0.

B. RECOMMENDATIONS

1. DOD

The following are recommendations for the DOD to improve demand forecasting:

a. Replace $CIMIP_f$ with MASE as the Aggregate Forecast Accuracy Measurement of Record

As we have shown, *MASE* is superior to $CIMIP_f$ in its ability to provide intuitive results across more demand patterns, while also avoiding distortions from unit cost and demand volume. The built in benchmarking of the *MASE* equation will also enable the DOD to more accurately assess the forecasting performance of the services and DLA.

b. Consider the Naïve Method as a Basis for Department Benchmarks

Direct comparison of demand forecasting performance between the services and DLA using an absolute error metric, such as $CIMIP_f$, does not consider the difficulty of forecasting for the unique materiel populations. A department-wide goal that arbitrarily declares a certain accuracy percentage as acceptable does not accurately reflect the complexity of the task and has the potential to drive counter-productive behavior in an effort to reach the goal. A better measure of demand forecasting performance would utilize a benchmarked metric, such as *MASE*, and then set the standard as outperforming the benchmark. In the case of *MASE*, which uses the naïve method as a benchmark, this would encourage the services and DLA to attain an aggregate forecast accuracy score equal to or less than some number less than one.

2. Navy

The following are recommendations for the Navy to improve demand forecasting:

a. Transition to Flexible Forecasting Methods at the Item Level

As we have shown, the Navy's current forecasting method of exponential smoothing with backcasting outperforms the naïve method on average when the CV is less than 1.6. If NAVSUP's forecasters had flexibility in their choice of forecasting method, then on average, they would be able to select an analytical forecasting method that outperformed the naïve method when the CV of an item is less than 2.0. The complexity of generating accurate demand forecasts for such a diverse set of items does not lend itself to using only one analytical forecasting method. As the ERP program improves its capabilities, the Navy would benefit from more flexibility in its forecasting

methods. The ideal approach would be to apply multiple forecast methods to the historical data of each line item and then choose the forecast method that optimizes the *MASE* result, or whichever accuracy metric the Navy utilizes.

b. Utilize MASE to Analyze Forecast Accuracy at the Item Level

MASE has advantages over both the $CIMIP_f$ and *LASE* equations and utilizing it as a forecast accuracy measurement will enable WSS to better identify specific line items that have not been well forecasted over time even when actual demand is zero.

c. Publish a NAVSUP Demand Forecasting Procedures Instruction

During the course of our research we could not locate a NAVSUP instruction that detailed the procedures that WSS shall use to generate demand forecasts for all of the various situations and how to measure those results. While there are internal business rules and other technical ERP documents, an instruction of this type would ensure a broader understanding of demand forecasting across the Navy and open up the process for constructive criticism that could lead to improved results.

C. AREAS FOR FUTURE RESEARCH

The challenge of accurately forecasting demand across the DOD is not a simple matter and the recommendations we have offered here are not likely to solve all of the issues that prevent the DOD from improving forecast performance. During the course of our research we looked at many segments of this issue that we did not have the opportunity to explore further. Some of these ideas may generate constructive improvements while others may not. The following are non-mutually exclusive ideas that we feel deserve further study in order to improve demand forecasting within the Navy and DOD.

1. Item Manager Discretion to Adjust ERP Derived Forecast

In our discussions with NAVSUP we learned that after ERP develops demand forecasts using the exponential smoothing with backcasting method these forecasts are subject to IM review and possible adjustment. We feel that it would be worthwhile to

compare the effectiveness of the IM adjusted forecasts to the original ERP developed forecast. A comparison of the actual demand data to the original and adjusted forecasts should reveal if the IM adjusted forecasts result in more or less accurate forecasts than the original ERP derived forecast. The scope of this research could examine all LCI's or just a specific LCI-subset, since NAVSUP uses different forecasting methods to generate forecasts for each LCI group

Additionally, surveys of the IM's could determine the leading reasons for adjusting an ERP-derived forecast. A comparison of these IM provided reasons with the actual forecast performance could help determine which reasons generally result in more accurate forecasts and which generally result in less accurate forecasts. If the human survey portion is included, the NPS researcher would need to attain permission from the human research protection program office and the institutional review board. A study of this kind would also require the full support of NAVSUP and access to the IM's.

2. Explore the Use of Retail Level Demand in Forecast Development

To develop demand forecasts, NAVSUP uses quarterly wholesale level demand over a five-year period. While this data provides a good proxy for aggregated retail demand and is easier to obtain, it also results in less frequent demand occurrences and could hide demand patterns. Although retail level demand can be challenging to organize and interpret, it may provide a better data set to generate demand forecasts. In multi-echelon supply chains, demand information from the end user level must be tracked in order to mitigate the negative impacts of the bullwhip effect. When demand variability at the retail level is combined with a lack of communication up the supply chain, excess inventory is likely to form at all levels. CIMIP has addressed inventory visibility challenges, but sharing of end-customer demand information can also help to reduce unnecessary inventory. We propose an analysis of whether properly trimmed retail level demand can provide a better demand forecast for items that have traditionally been difficult to forecast with only wholesale level demand.

3. Explore Alternatives to Managing Material by Life Cycle Indicator

The Navy currently uses LCI's from one to six to segregate material based on the maturity of the parent program that it supports. Initially for LCI-1, when demand is non-existent, engineering estimates are used to develop forecasts. As the item progresses to the next LCI categories these engineering estimates begin to factor in observed demand in order to develop forecasts. By the time an item is classified as an LCI-4 or -5 the analytical forecast is based solely on observed demand. While in general this makes sense, it may be possible that items could be more effectively managed and forecasted if they were placed into groups based on other criteria, instead of their parent programs' life cycle. We propose a study to determine what these more effective sorting criteria are and how best to employ them.

4. Time Periods and Fractions

The Navy currently uses five years of wholesale level demand, sorted into 20 quarterly buckets, to generate a single number demand forecast for the 21st quarter. To obtain a 12-month forecast the quarterly forecast number is multiplied by four. This single number is not always a whole integer. We propose a study of the effect of using different time buckets (days, weeks, months, etc.), different historical time periods (1, 3, 7, etc. years) and the treatment of fractional demand forecasts (round up, round down, no rounding, etc.) to potentially generate more accurate forecasts.

5. Investigate the Use of Alternative Forecasting Methods

The mathematical model presented in Chapter IV aims to generate improvement in forecast accuracy. However, it is not sufficient to select methods with the best *MASE* values throughout the entire curve, disregarding the fact that they can be worse than the naïve method. That method is considered to be a rudimentary prediction tool and still systematically outperforms the simple forecast methods used in this research for items with $CV > 2.0$. While we cannot recommend its blanket utilization for those items, we propose an investigation of the potential benefits of using either more complex time-series forecasting methods or alternative forecasting methods such as causal, qualitative, and expert estimates.

6. Analyze the DOD Bias Metric

The initial concerns of Congress and GAO, in dealing with the issue of excessive secondary inventory, seemed to be more focused on reducing the bias to over-forecast instead of improving forecast accuracy. While the focus today seems to have shifted away from bias toward accuracy, there is still a requirement to measure bias in forecasting. The DOD business rules that defined the accuracy metric also laid out the procedures for utilizing the bias metric. As we have discussed, our research centered on the accuracy metric, but the bias metric, as defined in Equation (2.25), could also benefit from a further analysis of its strengths and weaknesses.

7. Portfolio Theory Approach

Portfolio theory indicates that an investor can optimize the trade-off between risk and reward through diversification. If we apply that rationale to the flexible forecasting model, better results are possible when the pool of forecasting methods reflects a large spectrum of responsiveness, and is comprised of specific methods to deal with trends, seasonality and intermittent demand. We propose an investigation of the benefits of applying a portfolio theory rationale to the flexible forecasting model.

8. Grouping Method

In our research, we grouped items into CV clusters as an attempt to identify methods that are expected to outperform others for a particular range of variability. However, that grouping method was not able to segregate items in a way that one specific forecasting method outperformed the others. We acknowledge the possibility of grouping items in different ways, like demand patterns, clusters of unit costs, clusters of dollar demand, etc. However, forecasting method selection at the item level is more likely to produce more accurate forecasts than any other kind of grouping. Individualized forecasts are likely to require significantly more effort, so we propose an analysis to determine if this additional effort at the item level pays-off, in terms of marginal gains in accuracy.

9. Optimization of Parameters

Parameters used to initiate the calculations of forecast values in each of the methods that we tested were arbitrarily chosen. The intent of our research was to uncover potential opportunities of improvement by applying a flexible forecasting model. We propose further investigation of the results generated if the parameters were optimized for each item.

10. Apply Statistical Tools to Generalize Results

During our analysis of the DOD's accuracy metric, we utilized quick, hypothetical tests to uncover evidence of inherent flaws within $CIMIP_f$. The simplicity of these tests unfortunately means that the findings are not supported by any statistical analysis and cannot be generalized to larger datasets. Therefore, we propose statistical analyses on the impacts of the $CIMIP_f$ flaws that we identified.

LIST OF REFERENCES

Armstrong, J. S. (Ed.). (2001). *Principles of forecasting: A handbook for researchers and practitioners* (Vol. 30). New York: Springer Science & Business Media.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

Armstrong, J. S., & Lusk, E. J. (1983). Commentary on the Makridakis Time Series Competition (M-Competition) introduction to the commentary the accuracy of alternative extrapolation models: Analysis of a forecasting competition through open peer review. *Journal of Forecasting*, *2*, 259–311.

Assistant Secretary of Defense for Logistics and Material Readiness (ASD[L&MR]). (2010). *Comprehensive Inventory Management Improvement Plan*. Washington, DC: Department of Defense.

Baker, J. W., Schubert, M., & Faber, M. H. (2008). On the assessment of robustness. *Structural Safety*, *30*(3), 253–267.

Bencomo, L.A. (2010). *Demand forecast accuracy metric: LASE (Lead-time Adjusted Symmetric Error)*. Mechanicsburg, PA: NAVSUP Naval Inventory Control Point.

Cooper, J. P., & Nelson, C. R. (1975). The ex-ante prediction performance of the St. Louis and FRB-MIT-PENN econometric models and some results on composite predictors. *Journal of Money, Credit and Banking*, *7*(1), 1–32.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Department of Defense. (2013, August 13). *DOD Demand forecasting accuracy and bias metrics: Business rules*. Washington, DC: CIMIP Forecasting/TAV/ME Modeling Working Group.

Deputy Assistant Secretary of Defense for Supply Chain Integration (DASD(SCI)). (2010). *Logistics strategic plan*. Washington, DC: Department of Defense. Retrieved from http://www.acq.osd.mil/log/images/DOD%20Logistics%20Strategic%20Plan/DoDLogStratPlanFinalSigned-100707.pdf

Federal Managers Financial Integrity Act of 1982, Pub. L. No. 97–255, H.R. 1526 (1982). Retrieved from https://www.whitehouse.gov/omb/financial_fmfia1982

Ferber, R. (1956). Are correlations any guide to predictive value?. *Applied Statistics*, 113–121.

Fildes, R. (1989). Evaluation of aggregate and individual forecast method selection rules. *Management Science*, *35*(9), 1056–1065.

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review/Revue Internationale de Statistique*, 289–308.

Fildes, R., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and operational research: A review. *Journal of the Operational Research Society*, *59*(9), 1150–1172.

General Accounting Office. (1988). *Defense inventory: Growth in secondary items*. (GAO/NSIAD-88-189BR). Washington, DC: Author. Retrieved from http://www.gao.gov/products/NSIAD-88-189BR

General Accounting Office. (1989). *Financial Integrity Act: Inadequate controls result in ineffective federal programs and billions in losses*. (GAO/AFMD-90-10). Washington, DC: Author. Retrieved from http://www.gao.gov/products/AFMD-90-10

General Accounting Office. (1990). *Government financial vulnerability: 14 Areas needing special review*. (GAO/OCG-90-1). Washington, DC: Author. Retrieved from http://www.gao.gov/products/OCG-90-1

Gilchrist, W. G. (1979). Discussion of the paper by Professor Makridakis and Dr. Hibon. *Journal of the Royal Statistical Society, A*, *142*, 146–147.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*(4), 405–408.

Government Accountability Office. (2007). *Defense Inventory: Opportunities exist to save billions by reducing Air Force's unneeded spare parts inventory*. (GAO-07-232). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-07-232

Government Accountability Office. (2008). *Defense inventory: Management actions needed to improve the cost efficiency of the Navy's spare parts inventory*. (GAO-09-103). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-09-103

Government Accountability Office. (2009). *Defense inventory: Army needs to evaluate impact of recent actions to improve demand forecasts for spare parts*. (GAO-09-199). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-09-199

Government Accountability Office. (2010). *Defense inventory: Defense Logistics Agency needs to expand on efforts to more effectively manage spare parts*. (GAO-10-469). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-10-469

Government Accountability Office. (2011). *DOD's 2010 Comprehensive Inventory Management Improvement Plan addressed statutory requirements, but faces implementation challenges*. (GAO-11-240R). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-11-240R

Government Accountability Office. (2012). *Defense inventory: Actions underway to implement improvement plan, but steps needed to enhance efforts*. (GAO-12-493). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-12-493

Government Accountability Office. (2015a). *High-Risk series: An update*. (GAO-15-290). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-15-290

Government Accountability Office. (2015b). *Defense inventory: Services generally have reduced excess inventory, but additional actions are needed*. (GAO-15-350). Washington, DC: Author. Retrieved from http://www.gao.gov/products/GAO-15-350

Hendricks, W. A., & Robey, K. W. (1936). The sampling distribution of the coefficient of variation. *The Annals of Mathematical Statistics, 7*(3), 129–132.

Hoover, J. (2006). Measuring forecast accuracy: Omissions in today's forecasting engines and demand-planning software. *Foresight, 4,* 31–35.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*. [E-book version]. Retrieved from https://www.otexts.org/fpp

Jackson, Keneth J. (2011). Forecast Error Metrics For Navy Inventory Management Performance (Master's thesis). Retrieved from http://hdl.handle.net/10945/5756

Jenkins, G. M. (1982). Some practical aspects of forecasting in organizations. *Journal of Forecasting*, *1*(1), 3–21.

Kendall, F. (2011, December 14). DOD supply chain materiel management policy (DOD Instruction 4140.01). Washington, DC: Department of Defense.

Koehler, A. B. (2001). The asymmetry of the sAPE measure and other comments on the M3-competition. *International Journal of Forecasting*, *17*(4), 570–574.

Ledolter, J., & Abraham, B. (1984). Some comments on the initialization of exponential smoothing. *Journal of Forecasting*, *3*(1), 79–84.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications*. New York, NY: John Wiley and Sons.

Makridakis, S., Hibon, M., & Moser, C. (1979). Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society. Series A (General)*, 97–145.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*(1), 5–22.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(9), 987–996.

National Defense Authorization Act for Fiscal Year 2010, Pub. L. No. 111–84 § 328, 123 Stat. 2255 (2009). Retrieved from https://www.gpo.gov/fdsys/pkg/PLAW-111publ84/pdf/PLAW-111publ84.pdf

Nelson, H. L., & Granger, C. W. J. (1979). Experience with using the Box-Cox transformation when forecasting economic time series. *Journal of Econometrics*, *10*(1), 57–69.

Nelson, C. R. (1972). The prediction performance of the FRB-MIT-PENN model of the U.S. economy. *The American Economic Review*, *62*(5), 902–917.

Nelson, C. R. (1984). A benchmark for the accuracy of econometric forecasts of GNP. *Business Economics*, 52–58.

Newbold, P., & Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, 131–165.

Reid, D. J. (1968). Combining three estimates of gross domestic product. *Economica*, *35*(140), 431–444.

Schupack, M. B. (1962). The predictive accuracy of empirical demand analyses. *The Economic Journal*, *72*(287), 550–575.

Tayman, J., & Swanson, D. A. (1999). On the validity of MAPE as a measure of population forecast accuracy. *Population Research and Policy Review*, *18*(4), 299–322.

Thrift Savings Plan. (n.d.a). TSP and index annual returns 2011–2015. Retrieved from https://www.tsp.gov/InvestmentFunds/FundPerformance/annualReturns.html

Thrift Savings Plan. (n.d.b). Fund comparison matrix. Retrieved from https://www.tsp.gov/InvestmentFunds/FundsOverview/comparisonMatrix.html

Wheelwright, S., Makridakis, S., & Hyndman, R. J. (1998). *Forecasting: Methods and applications*. New York, NY: John Wiley & Sons.

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