3-22-2019

Machine Learning Models of C-17 Specific Range Using Flight Recorder Data

Marcus Catchpole

Machine Learning Models of C-17 Specific Range Using Flight Recorder Data

THESIS

Marcus Catchpole, Capt, USAF

AFIT-ENG-MS-19-M-016

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-MS-19-M-016

MACHINE LEARNING MODELS OF C-17 SPECIFIC RANGE USING FLIGHT RECORDER DATA

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Computer Engineering

Marcus Catchpole, B.S. Electrical Engineering

Capt, USAF

March 2019

AFIT-ENG-MS-19-M-016

MACHINE LEARNING MODELS OF C-17 SPECIFIC RANGE USING FLIGHT RECORDER DATA

THESIS

Marcus Catchpole, B.S. Electrical Engineering
Capt, USAF

Committee Membership:

Dr. Laurence D. Merkle
Chair

Dr. Brett Borghetti
Member

Col Adam Reiman, PhD
Member

AFIT-ENG-MS-19-M-016

Abstract

Fuel is a significant expense for the Air Force. The C-17 Globemaster fleet accounts for a significant portion. Estimating the range of an aircraft based on its fuel consumption is nearly as old as flight itself. Consideration of operational energy and the related consideration of fuel efficiency is increasing. Meanwhile machine learning and data-mining techniques are on the rise. The old question, "How far can my aircraft fly with a given load cargo and fuel?" has given way to "How little fuel can I load into an aircraft and safely arrive at the destination?" Specific range is a measure of efficiency that is fundamental in answering both questions, old and new. Predicting efficiency and consumption is key to decreasing unnecessary aircraft weight. Less weight means more efficient flight and less fuel consumption. Machine learning techniques were applied to flight recorder data to make fuel consumption predictions. Accurate predictions afford smaller fuel reserves, less weight, more efficient flight, and less fuel consumed overall. The accuracy of these techniques were compared and illustrated. A plan to incorporate these and other modeling techniques is proposed to realize immediate fuel cost savings and increase savings over time.

AFIT-ENG-MS-19-M-016

Many thanks to Dr. Merkle and the committee of advisors for both their teaching and guidance. Thank you to my father and mother who encouraged academics and engineering. A deep thanks to my wife and our children whose love and support made this work possible.

Table of Contents

Page

Page

List of Figures

List of Tables

MACHINE LEARNING MODELS OF C-17 SPECIFIC RANGE USING FLIGHT RECORDER DATA

I. Introduction and Background

This chapter gives a background of the research effort and details where this effort fits in the big picture. The contents and organization of each chapter are described. An overview of this data and and its origin is explained. To further understand the data source and relevant flight details, an example sortie is dissected from its flight recorder data with notes on key data details. Finally, the research questions and the conclusions of the research are summarized.

1.1 Introduction

Boeing C-17 Globemasters account for a significant portion of the United States Air Force's air mobility capacity. As such, they account for a significant amount of the Air Force's fuel usage. The cost of fuel is substantial for air freight in both civilian and military applications. Decreasing fuel consumption for the C-17 fleet would result in significant cost savings.

The amount of fuel an aircraft uses in a sortie increases with the aircraft weight. A significant part of the aircraft weight is that of the fuel itself. Fuel remaining at the end of a sortie has served no purpose. Instead, it has caused more fuel to be consumed. A way to minimize fuel consumption for a sortie is to put no more fuel in the aircraft than is necessary to reach its destination.

The current state of the Air Force's effort, as well as its motivation was summarized in a news release from the Air Force Office of Operational Energy. An effort to increase

the amount of data regarding aircraft fuel consumption began in 2018. The goal is to "enable data-driven decisions and better target opportunities to improve operations that deliver competitive advantages against adversaries." [1] Operational energy has been incorporated into war gaming. The C-17 fleet has been targeted for efficiency improvements because of the vast amount of fuel it consumes. The Air Force is giving attention to the great costs of fuel and gives special attention to the Globemaster fleet.

There are opportunities to use data to "target opportunities and ... deliver competitive advantages." One opportunity is to build a predictive model more accurate than what currently exists. Models to predict fuel usage already exist and are used in flight planning. Aircraft manufacturers provide aircraft owners with relatively simple fuel consumption models. There is much room for improvement in the models used.

A benefit of better models is best envisioned by considering the past. Railroads used to be the main source of transportation across the United States. The arrival of a train at its scheduled time was said to be so precise that one could accurately set their watch by its arrival. In a similar way, flight planners will be able to say "We will load the aircraft with x pounds of fuel. The aircraft will arrive at its final destination with between y and z pounds of fuel." Better models will have two results: the value of x for a given sortie decreases and the difference between y and z narrows.

Cruise is a key segment of flight when considering aircraft fuel use. The cruise segment occurs after the aircraft levels off at a target altitude. It ends before the aircraft descends to land at a target airport. For all but the shortest sorties, a majority of fuel is used in this segment. Flight data was gathered from recorders and used to make predictive models for this segment of flight.

1.2 Background

Fuel consumption is a major cost-driver for flying freight. Fuel consumption can be decreased by decreasing weight. Accurate predictions of fuel usage would allow mission planners to minimize fuel weight. This research applies machine learning techniques to make predictive fuel models in the cruise phase of flight. The cruise phase accounts for the majority of flight, both in terms of flight time and fuel used. The sources of variance that make one cruise segment different from another are relatively few in this phase. That makes this phase a particularly good candidate to which to apply machine learning techniques. Data from C-17 flight recorders is used to make and evaluate predictive fuel models. The scope of this research ties into larger organizational efforts. These goals have three levels of abstraction and are formatted differently for emphasis and clarity.

> This research is part of a vision to realize cost savings and enhance the use of operational energy. This increases the body of information needed to accurately predict fuel consumption for C-17s. Reliable, accurate predictions allow mission planners to optimize flight plans to meet mission requirements with the best fuel usage and most frugal expense of operational energy.

This effort can be translated to a technical goal. The technical goal is to predict fuel quantity remaining at the end of the sortie given its initial fuel load. This is summarized by an x-y-z numerical goal.

> For a given C-17 sortie with a value x, predict y and z. The value x is the takeoff fuel. The values of y and z are the minimum and maximum range of remaining fuel possible at the conclusion of the sortie. For a given x, the true fuel value must lie between y and z. Cost savings are realized by minimizing x. A minimum value of x can be achieved as prediction confidence, $z - y$, is minimized.

Finally, the goal of this paper is to apply machine learning techniques toward this effort. The goal can be articulated:

> Apply machine learning techniques on C-17 flight recorder data to make predictive models of specific range (fuel efficiency) during cruise flight. Models will make this prediction based on an aircraft's gross weight and altitude. Perform model selection of off-the-shelf machine learning techniques. Evaluate the best model(s).

1.3 Flight Recorder Data

Flight recorders store information from various sensors on the aircraft. The primary use of most of these sensors is to give information to aircraft operators. A secondary use is to create a record. Much of an aircraft's sensor data is kept in the data flight recorder, thus flight recorders capture a large volume of data.

The raw data from C-17 flight recorders is suitably presented in 250 fields. Each field represents a measured parameter. Some of the relevant parameters are listed.

1. Aircraft Time
2. Altitude
3. Angle of Attack
4. Autopilot Settings
5. Bank Angle
6. Fuel Flow to Each of Four Engines
7. Fuel Quantity
8. Indicated Airspeed

9. Weight on Wheels

The nature of the data is illustrated by a reproduction of an example sortie.

1.4 Anatomy of a Sortie

To understand both the nature of a sortie and the data recorded over the course of a sortie, a sortie's details may be reconstructed from its flight recorder data. A figure that summarizes key aspects of the sortie can be seen in Figure 1. In this sortie, the C-17 was loaded with 160,000 lbs of fuel. The aircraft used 121,000 lbs of fuel and landed with 39,000 lbs of fuel. The flight flew through 2,008 miles of air. The flight recorder collected 6 hours and 56 minutes of operation. The takeoff runway was at 784 feet in elevation and the destination runway was at 4,512 feet in elevation. Calculating the true air speed from the indicated airspeed it is determined that the aircraft kept an airspeed of approximately 500 knots for the majority of the sortie.

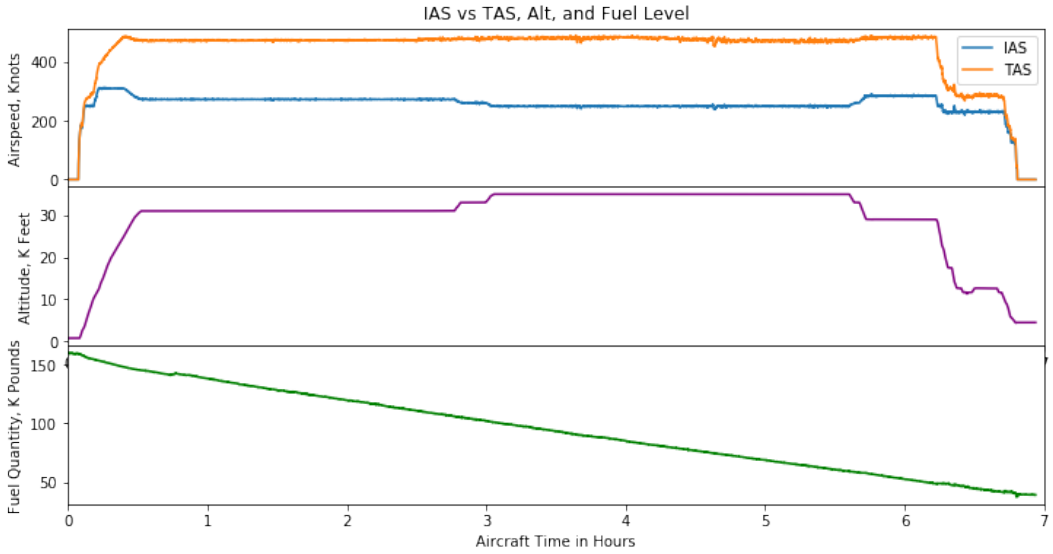

Figure 1. Various measurements recorded by the flight recorder over the course of an example sortie. IAS is indicated airspeed. TAS is true airspeed and is calculated from the altitude and indicated airspeed flight recorder values. Fuel quantity is shown in green.

The autopilot system was used for a majority of the sortie. There are two autopilot

settings that may be used independently or together. One is the altitude hold, which manipulates the control surfaces of the aircraft to maintain level flight. The other is auto thrust, which adjusts the throttle inputs. The throttle inputs affect the fuel flow to the engines which results in thrust. Figure 2 shows the portions of the sortie where parameters were controlled by the flight crew only. There was a very small amount of time where neither of the two settings was engaged.

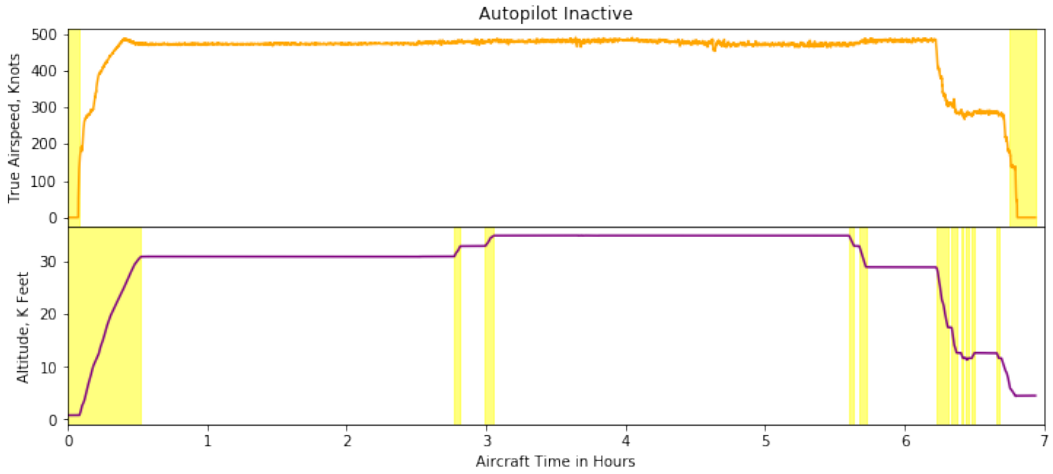

Figure 2. Sections of flight where the respective autopilot functions are NOT used. Highlighted areas indicate a particular autopilot function is not being used. Auto throttle directly impacts airspeed. Altitude hold directly impacts the altitude. The top graph shows the true airspeed and highlights where the auto throttle function is not engaged. The bottom shows the altitude and highlights where the altitude hold function is not engaged.

A sortie can be considered in terms of these phases.

1. Taxi and Takeoff
2. Climb
3. Cruise
4. Step Climb
5. Descent and Approach
6. Land and Taxi

Taxi, Takeoff, and Climb.

Taxi, takeoff and climb were completed 32 minutes after the flight recorder began. Figure 3 shows relevant recorded parameters in detail. The aircraft uses thrust from the engines to move to the beginning of the runway. Once in position, the aircraft increases thrust to takeoff. The timing of the takeoff for the sortie was estimated using the true airspeed and the weight on wheels signal. During taxi, the measured airspeed is near zero. The portion of the data that accounts for acceleration for takeoff can be estimated using two conditions. These conditions are when the airspeed is nonzero and the weight on wheels signal is positive. Weight on Wheels is the output of a sensor designed to indicate whether there is weight on the landing gear. Using these conditions, takeoff took approximately 30 seconds.

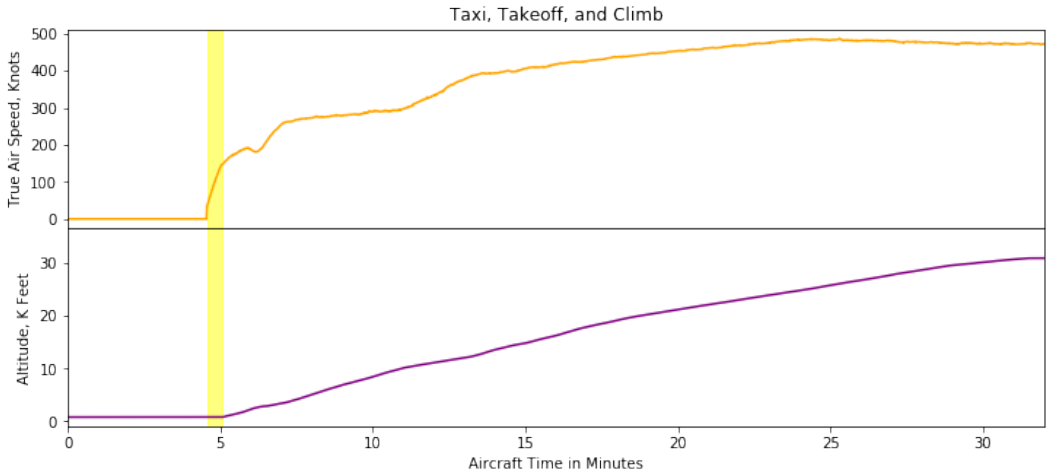

Figure 3. The first three phases of a sortie. The aircraft maneuvers on the ground to its takeoff position. The yellow highlights the time where the Weight on Wheels value was true and the true airspeed was nonzero. To the left of the highlighted segment is the taxi phase. To the right of the highlighted segment is the climb phase.

Sensor measurements are only estimates of the true variables they represent. An illustration of this important distinction was recorded in the first hour of flight. Two estimates of the fuel quantity can be seen in Figure 4. The approximations are similar, but they do not agree perfectly. The first and most direct measurement is from the

sensors in the fuel tanks. The second estimate of fuel quantity uses the recorded fuel flow values. To estimate the fuel quantity, the total amount of fuel that has flowed through the engines may be subtracted from the initial quantity. The fuel quantity Q for an initial quantity IQ at a given time T is calculated by the equation

$$Q = IQ - \sum_{t=1}^{T} FF(t) \tag{1}$$

Where

$$FF(t) = \text{the fuel flow of the aircraft at time } t$$

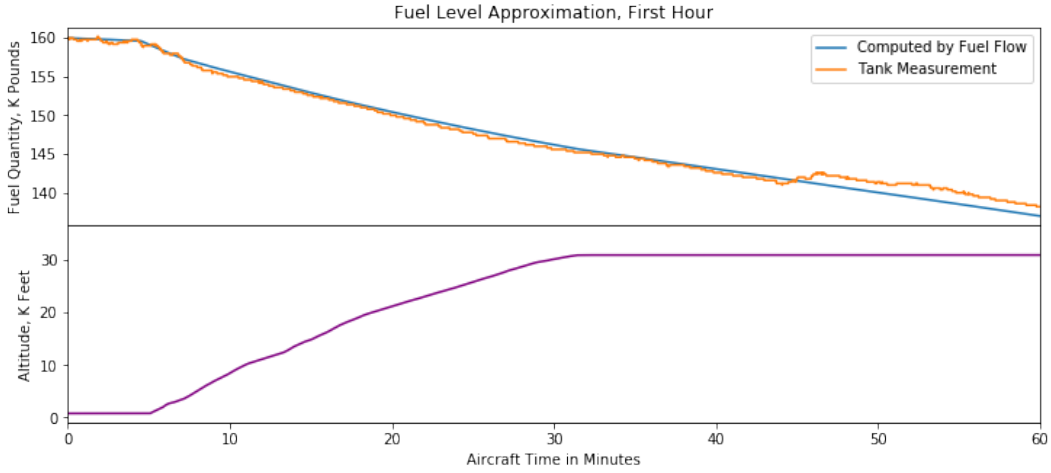

Figure 4. The top graph shows two approximations of the fuel quantity over time. Two sources of error are evident. First, the tank measurement (shown in orange) is discretized. Discretization is revealed by level portions of the graph interrupted by discontinuous jumps. The second source of error, also in the tank measurement, shows a fuel increase between 45 and 50 minutes. It may seem more fuel was added during the sortie, which is not the case. This is likely due to how fuel moved around the tanks with regard to the sensors. The blue estimate of fuel quantity is the initial fuel load minus the cumulative fuel flow to all four engines. These are two estimations. Which estimation is "better" is worth consideration.

The first increase in altitude is during the initial climb. At 32 minutes, the plane reached a cruising altitude of 31,000 feet.

Cruise and Step Climb.

Though the aircraft had begun the cruise section, it did not stay at the same altitude for the duration of the section. This aircraft began to climb again at 2 hours 46 minutes into the sortie. This is called a step climb. A step climb increases the aircraft's altitude. Air at higher altitudes has lower density and results in less drag on an aircraft for a given airspeed. Step climbs result in more efficient fuel use. The result of the step climb is that the aircraft flies at an altitude where fuel usage is more efficient. The aircraft's weight decreases over the course of the sortie as fuel is burned. The weight change causes the most fuel efficient flight altitude to change. Step climbs are accomplished mid-flight when the change in weight results in a better cruise altitude. These step climbs and descents can be seen in better detail in Figure 5. The cruise portion of this flight shows the difference between these two methods of estimation. Figure 5 shows both the sensor value of the indicated airspeed and the calculated value true airspeed.

Two pressure systems are used when estimating airspeed. These are the static pressure and the pitot pressure. The air in the static pressure system is designed to have a pressure equal to what would be measured in still air. The pitot pressure system is designed to have a pressure equal to the pressure created by the motion of the aircraft. The difference between these two systems allows a sensor to estimate the airspeed. Without compensating for altitude, the difference in these two pressure systems is used to estimate the *indicated* airspeed. The indicated airspeed sensor is calibrated to show the speed as sea-level air density. It is a value proportional to the difference between the static and pitot pressure systems. Conversely, the true airspeed is an estimate of aircraft's speed relative to the air through which it is flying. The indicated airspeed decreases compared to true airspeed at higher altitudes. True airspeed may be estimated by calculation using the altitude, temperature, and

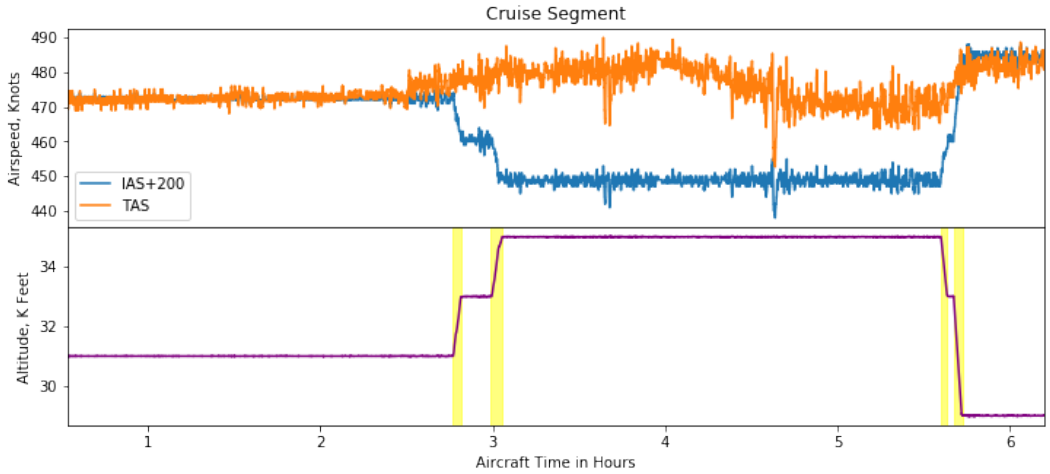

Figure 5. A part of the cruise section. The top graph shows both the sensor value of indicated airspeed (blue) and the calculated value of true airspeed (orange). Indicated airspeed is offset to appear in the same range as true airspeed. The bottom graph shows the altitude with the step climbs and descents highlighted in yellow. The altitude hold setting is off in the highlighted areas. The variance in true airspeed is larger in amplitude. True airspeed drifts while indicated airspeed is roughly constant. Also, true airspeed varies relatively little with altitude while the indicated airspeed decreases with increasing altitude. The low variation in true airspeed suggests the sortie's target was to maintain a true airspeed.

indicated airspeed sensors.

Descent, Approach, Landing, and Taxi.

In this sortie, the aircraft left the cruise segment and entered the descent phase when the aircraft descended from its maximum altitude. The aircraft crew no longer attempted to maintain an altitude to optimize fuel consumption. Instead, it prepared to enter the traffic pattern of the destination airport. Figure 6 shows the last phase of flight.

1.5 Selecting a Cruise Altitude

Specific range is a measure fuel efficiency. Specific range for aircraft has the dimensions of distance per unit weight of fuel. It is common to use units of nautical miles per thousands of pounds of fuel. A related and more familiar standard is

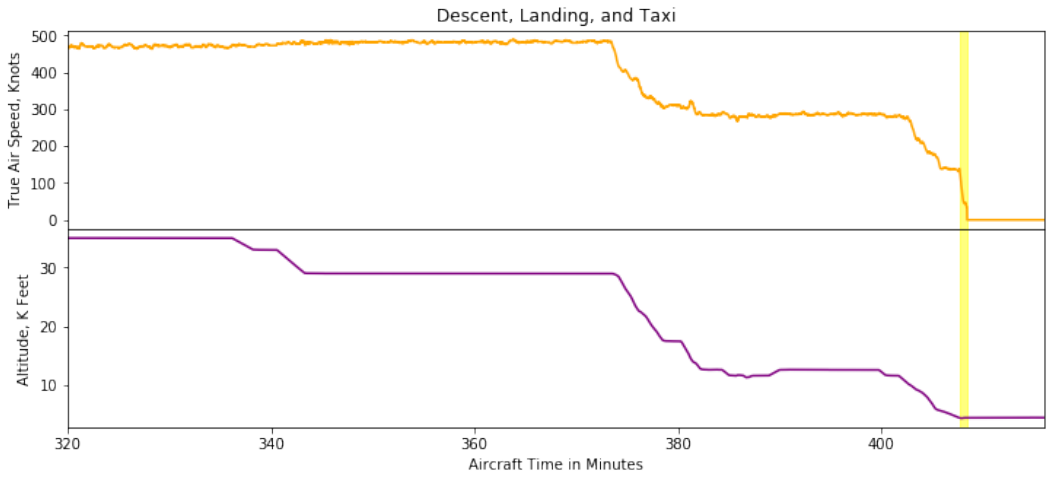

Figure 6. The final phases of the example sortie. Left of the yellow highlighted section is when the aircraft was in the air. The yellow highlighted section shows the conditions where the Weight on Wheels had a positive output and the airspeed sensor was nonzero. To the right of the highlighted area accounts for the taxiing to park.

that for automobile fuel efficiency. In the United States, automobile efficiency is typically measured in miles per gallon. Automobiles operating in town or on highway conditions result in different efficiencies. Likewise, different flying conditions affect specific range. Altitude and the gross weight of the aircraft are two of those factors. Gross weight is the sum of the empty aircraft weight, the payload weight, and the fuel weight. The empty aircraft weight does not change significantly from sortie to sortie. The payload weight changes for the mission. The fuel weight is often the estimated minimum required to confidently reach the destination plus a reserve. The weight is dependent on conditions over which mission planners have little control. Altitude may vary, thus an altitude is selected to optimize cruise for a given gross weight.

Another way to examine the example flight is in the domain of gross weight and altitude. This is done in Figure 7. This is a useful domain in which to examine data. Time increases from the higher weights near the bottom of the graph to lower weights near the top. The specific range varies over time and is generally higher at higher altitudes and lower gross weights.

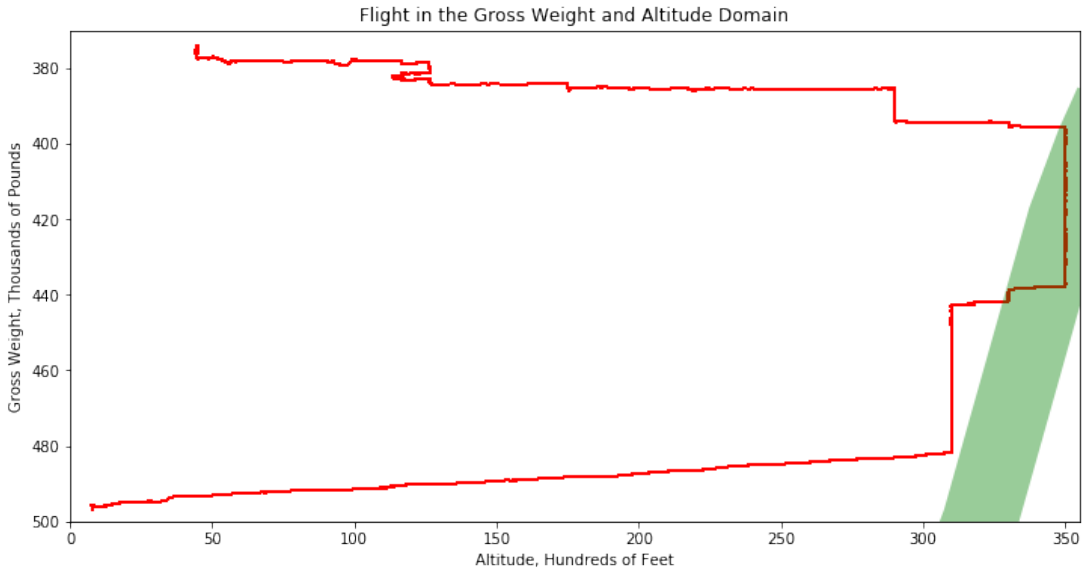

Figure 7. The sortie decreases in gross weight over time. Samples near the bottom of the figure represent those taken early in the sortie. Samples near the top are taken later in the sortie. The green highlighted area represents the C-17 technical manual's estimation of optimal fuel consumption under normal conditions. [2]

1.6 Research Questions

Flight recorder data was used to estimate specific range as a function of altitude and gross weight. This was done by following common predictive modeling processes. Several questions were answered as part of model selection and evaluation. Off-the-shelf modeling techniques have different strengths and weaknesses. The central question of this research is:

Can nonlinear models made from flight data make accurate predictions for fuel consumption in cruise segments of flight given initial gross weight and altitude?

To answer this requires two other questions to be considered.

How accurately can nonlinear models predict specific range?

Given a nonlinear model for specific range, how accurately can the distance traveled be predicted for the amount of fuel consumed?

The process of building these models requires several subsequent questions to be

answered. The central question evaluated is:

Given four common nonlinear modeling techniques, which makes the most accurate prediction of specific range?

Many off-the-shelf predictive modeling techniques require parameter tuning. Parameters were tuned for two techniques, k-nearest neighbors and forest regression. For the k-nearest neighbors technique two questions are evaluated:

Which weighting scheme is more accurate, Euclidian or uniform?

What is the value of k that results in the most accurate model?

For the forest regression technique one parameter was optimized. The following question was explored:

When building tree models for the forest, is the most accurate model made when features are selected at random or based on the best feature?

After the best parameters were found, the techniques were compared. Comparing the models answered these questions:

Which of the modeling techniques is most accurate in predicting specific range?

What are the relative strengths and weaknesses of the techniques?

1.7 Overview of Methodology and Results

The modeling techniques explored are listed.

1. Linear Regression
2. K-Nearest Neighbors (KNN)
3. Bagged Tree Regression (Random Forest)
4. Artificial Neural Network (ANN)

Specific range is calculated from several values including the aircraft's fuel flow. The aerospace theory describes specific range in terms of steady state conditions

where lift is equal to weight and thrust is equal to drag. This is never the case in real flight due to stochastic inputs to the system. Fuel flow is one parameter used to estimate specific range. A moving average was applied to fuel flow values before the specific range truth values were calculated. For each technique with multiple parameters, the best parameters were found. The modeling techniques tuned with their best parameter settings showed similar performance, especially in ranges of altitude and gross weight where there were many samples. The best overall modeling technique was linear regression. This is likely due to its relatively small capacity. The small capacity limits the effect of erroneous samples. KNN was found to be too computationally expensive with an accuracy no better than the other modeling techniques.

A prediction of range on the test data cruise flight segments showed consist results between the models. All of the techniques were shown to have similar errors in making this prediction. For estimating short segments of flight, the models tended to overestimate the range of the aircraft. For long segments, the models tended to underestimate the aircraft range. There errors in range prediction for each model are correlated. This indicates improvements in prediction accuracy are likely to come from data processing and selection techniques, than model refinement.

1.8 Organization of the Document

In many ways, this research is a data mining project. It is organized both as a research effort and as a project. The Cross-Industry Standard Process for Data Mining (CRISP-DM) [3] is a standard and guideline for organizing data-mining projects. CRISP-DM is formally covered in Chapter II. Each chapter can be described both in terms of a typical thesis structure and by the relevant elements of the CRISP-DM process.

Chapter I contains a background and statement of the problem. In project management terminology, it articulates the background, business objectives, and how those are related to the data mining goals. Much of the content that is considered Business Understanding fits into this chapter.

Chapter II contains a literature review. In project management terms, this is more background information and an assessment of tools and techniques.

Chapter III contains the methodology so the work accomplished can be reviewed and repeated. The procedure portion of the phases Data Understanding, Data Preparation, and Modeling are in this chapter. In these phases the bulk of a data-mining project is accomplished to prepare for the subsequent phase, Evaluation.

Chapter IV contains data descriptions and analysis. These are the numbers, visualizations, and technical details from the procedure outlined in Chapter III. The chapter describes the output of the three phases covered in Chapter III, Data Understanding, Data Preparation, and Modeling.

Chapter V contains findings and conclusions. Conclusions include recommendations on future work. The work is evaluated in its ability to meet the objectives and goals outlined in the Business Understanding phase. Possible actions and a review of the process was documented here.

II. Literature Review

2.1 Overview

This chapter provides information relevant to understanding the research efforts and results. The organizing standard for the research plan and the conceptual organization of the thesis is first. The "Universal Workflow of Machine Learning" enhances the understanding of the effort and is considered. Specific range, the value to be predicted as a result of the data mining process is explained in depth. The chapter continues with an overview of systems. Highlights from a research paper attempting to model specific range follow. A detailed review of models and the process is explained. The chapter concludes after considering other published works related to this effort.

2.2 Cross-Industry Process for Data Mining (CRISP-DM)

CRISP-DM [3] is a standard for organizing data mining projects. This research effort is a data-mining project. This standard was referenced when organizing this research effort and also serves as a baseline for logically organizing this paper. CRISP-DM organizes a project into phases. Each phase is divided into tasks. There are six phases in the data mining process listed here.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation

6. Deployment

The important aspects of each phase are explained in the remainder of this section.

Business Understanding.

In CRISP-DM, Chapman explains that the effort in the Business Understanding phase "focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives." [3, p. 10] Articulating the business objective is important. The business objectives give qualitative guidance throughout a project. Without a business goal in mind, it would be difficult to efficiently expend effort. Effort is efficiently expended when wise decisions can be made. Wise decisions are made when the expected trade-offs of decisions can be estimated and considered. The value of trade-offs is estimated based on the project's goals. With no goals, there is no guiding principle to evaluate the trade-offs different decisions will yield. Business objectives give a beginning point to establishing goals. The result of this phase is the ability to take advantage of opportunities and mitigate risks in efficient ways throughout a project. Business objectives, often qualitative, give a starting point for setting a quantitative goal.

Translating a project objective into something technical and quantitative is necessary. Translating a project objective this way has two benefits. First, a quantitative goal gives direction where qualitative objectives are imprecise. Second, quantitative goals give information regarding when an effort is complete. A goal is useless if there is insufficient criteria to determine whether or not the goal is met. Without a binary measure of success, projects could be improved or edited in perpetuity. Additionally, quantitative goals allow a better context to consider decision trade-offs.

Putting a qualitative goal in technical language allows effective decisions to be

made when outcomes may be estimated quantitatively. This guides the scientific effort of a project, especially where the business objective is insufficiently technical. An initial project plan may be drafted when the qualitative and quantitative goals are articulated. The tasks in this phase are enumerated here.

1. Determine Business Objectives
2. Assess Situation
3. Determine Data Mining Goals
4. Produce Project Plan

Determine business objective is the task where the project team considers big picture goals. *Assess situation* is the task where members account for the resources available to the team. *Determine data mining goals* is a task where members form quantitative goals from the qualitative big picture goals. Finally, the *produce project plan* is the task that, when complete, results in a draft plan that may be implemented to realize the business objectives.

Data Understanding.

In the data understanding phase, project team members should "become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information." [3, p. 10] Highlights from this phase were described in Chapter 1. CRISP-DM breaks this phase into four tasks.

1. Collect Initial Data
2. Describe Data

3. Explore Data

4. Verify Data Quality

The task *collect initial data* includes describing the method to attain data. Gathering this data gives insight to the cost of procuring additional data. *Describing data* is a task where the "gross" or "surface" properties of the data are articulated. *Exploring data* is a task where members focus on querying, visualization and reporting techniques. Finally, *data quality verification* is the task where the project team considers whether the data has errors, and how common errors are, if any are found.

Data Preparation.

Data Preparation is the phase where appropriate preprocessing techniques are applied in view of the desired project outcome. In this phase, "tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools." [3, p. 11] There are four tasks in the data preparation phase.

1. Select Data

2. Clean Data

3. Construct Data

4. Integrate Data

5. Format Data

Select data is a task that results in a decision on what data to use. The *clean data* task is where scientists raise the quality of data needed. *Construct data* is a task that includes producing derived attributes. *Integrate data* is the task that results in additional derivations of data attributes and combining data from multiple sources,

if necessary. Finally, *format data* is the step when syntactic changes to the data is accomplished.

Modeling.

In the Modeling phase, "various modeling techniques are selected and applied, and their parameters are calibrated to optimal values." [3, p. 11] A test to evaluate the quantitative goals is designed and then the model is built. This *build model* task is repeated for each modeling technique if more than one is tested. The technical goals are evaluated with regard to the model or models made. Technical measures are assessed in this phase. The four tasks in this phase are listed.

1. Select Modeling Technique
2. Generate Test Design
3. Build Model
4. Assess Model

Select modeling technique is a task where team members consider the technical specifics of the model and how it will be generated. Examples might be a forest regression or artificial neural network models. *Generate test design* is a task where the team considers the model or models' quality and validity. The plan to evaluate models is made during this task. *Build model* is the task where modeling techniques are applied to the data. Finally, *model assessment* is the task in which the team evaluates the target models in terms of accuracy and other relevant qualities. The qualities of each of the models are compared. Evaluation in this task means the technical evaluation, as separate from the phase with the same name. The evaluation phase considerations are more related to the business objectives rather than the technical goals.

Evaluation.

Unlike previous steps which consider the quantitative value of the models used, the Evaluation phase regards the business objectives. Qualitative considerations are articulated to identify reasons the selected model may not meet the business criteria. Tasks in this phase are enumerated.

1. Evaluate Results
2. Review Process
3. Determine Next Steps

These are tasks that consider a nearly finished product or service that may be implemented. *Evaluate results* is a task where the human factor is applied to determine whether the results of the process can achieve the qualitative objectives in the Business Understanding phase. The *review process* task concerns quality assurance. Considering and articulating lessons-learned is an important part of any business process and is emphasized in this phase. Decisions on whether to field the results of the project based on the business objectives are made during this phase. Observations made and information gained in the course of the project allows business objectives to be considered with the benefit of knowledge gained in the course of the project. In the task *determine next steps*, the decision on whether to proceed to deployment, or to begin more iterations of the process are made.

Deployment.

At the end of the data-mining process, fielding the object of the project must be completed to realize business success criteria. Much of the effort in this phase results in plans and reports. The tasks in this phase are listed.

1. Plan Deployment

2. Plan Monitoring and Maintenance

3. Produce Final Report

4. Review Project

The *plan deployment* task is when members consider applying the results of the project to achieve its business objectives. A plan to take the project results from evaluation to implementation is laid out in the deployment plan. The real world changes over time. A model that predicts well at release is subject to a decline in quality as the relevancy of the foundational data and assumptions change over time. *Planning monitoring and maintenance* is essential to handle this and ensure the long term utility of the project's output. The task *review project* is when the final report is issued. Lessons learned from the project perspective are considered.

2.3 The Universal Workflow of Machine Learning

Another perspective on the procedure for data-mining is from "Deep Learning with Python" [4, pp. 111-115] which describes a "Universal Workflow for Machine Learning." Machine learning and data-mining are often used interchangeably. The title implies the presented workflow is generalizable to any machine learning effort. The text considers artificial neural networks (ANNs), a subset of machine learning. The section in this text describes the data-mining process more briefly from a machine learning perspective. Chollet describes the workflow using these headings.

1. Defining the Problem and Assembling a Dataset

2. Choosing a Measure of Success

3. Deciding on an Evaluation Protocol

4. Preparing Your Data

5. Developing a Model that Does Better than a Baseline

6. Scaling Up: Developing a Model that Overfits

7. Regularizing Your Model and Tuning Your Hyperparameters

Many of these are sufficiently represented by a similar description in the CRISP-DM process, though there are nuances worth considering. The first nuance falls under the heading "Developing a Model that Does Better than a Baseline." The reason to include this is that there are two hypotheses for every attempt at data analysis and these must be tested. [4, pp. 111-115]

1. The outputs can be predicted given your inputs.

2. The available data is sufficiently informative to learn the relationships between the inputs and outputs.

This is the reason a baseline is created. If a naïve model or good guess can perform nearly as well as a machine learning technique, one or both of these hypotheses is false. Other nuances of this workflow apply specifically to ANNs. This is discussed in more detail in section 2.7.

2.4 Specific Range

A cornerstone of this research is Peckham's report "Range Performance in Cruising Flight." [5] It gives a framework for estimating the range of an aircraft. Calculating the range of an aircraft consists of three parts.

1. The aircraft's performance during climb, cruise and descent, for a range of conditions of weight, speed, and altitude.

2. Estimation of fuel available after taking into account payload and reserve fuel requirements.

3. Choice of flight trajectory such as cruise speed and height, climb and descent paths, distance for diversion, and time for holding. [5, p. 23]

Peckham's work focuses on performance, which is considered in terms of efficiency measured in distance traveled per amount of fuel burned. This measure of efficiency in the cruise portion of flight is defined as "specific range." It is "normally the subject of guarantees between the manufacturer and airlines, and checks on specific range performance at a number of speeds and altitudes form an important part of the flight-test program of a new aircraft." [5, p. 3]

Specific range is used to estimate aircraft range. "Integration of specific range over a given flight trajectory, for a change in aircraft weight equal to the fuel consumed gives the range." [5, p. 1] The way specific range, $\frac{dR}{dW}$, is used to calculate range, R is shown. [1]

$$R = -\int_{W_i}^{W_i - W_F} \frac{dR}{dW} dW \tag{2}$$

The value dR is the instantaneous distance traveled for the weight of fuel dW used in that instant. The integration is performed over dW. The variable W_i is the aircraft weight at the start of cruise and W_F is the weight of the fuel consumed.

Specific range in Peckham's report, and here, pertains only to cruise sections of flight. It is useful to consider these segments as, with the exception of short flights, cruise accounts for the most significant portion of a flight in both time and fuel

[1] The motivation for Peckham's inconsistent subscript capitalization is unclear.

consumption.

The report contains information about assumptions and restrictions to the calculation method. Significant considerations are quoted here.

> All theory in these sections is based on the assumption that the specific fuel consumption remains essentially constant along the cruise trajectory considered. [5, p. 4]
>
> In steady level cruising flight, because the incidence and altitude are small, it can be assumed that lift is equal to weight, and that thrust is equal to drag, so the expression for specific range becomes

$$-\frac{dR}{dW} = \frac{V}{cT} = \frac{1}{W}\frac{V}{c}\frac{L}{D} \tag{3}$$

> where R is range, W is weight, V is true air speed, c is specific fuel consumption, T is thrust, L is Lift, and D is drag. The resulting unit is distance per unit of fuel [5, p. 5].
>
> It is often sufficiently accurate to obtain cruise range by multiplying a mean specific range by the weight of the fuel consumed since the variation of specific range with weight is usually close to linear [5, p. 12].

It is reasonable to conclude that specific range calculation is appropriate under the three considerations:

1. $T = D$
2. $L = G$
3. Increase in specific range for a corresponding decrease in weight is approximately linear

2.5 Systems

"System" is a commonly used term that has a broad set of meanings. Kamen has a practical definition: "A system is a collection of one or more devices, processes, or

computer-implemented algorithms that operates on an input signal x to an output signal y." [6, p. 21] Kamen describes a signal as a function of the time variable t that has a real or scalar value. [6, p. 1]. Alternately, a system can be described as "an interconnection of elements and devices for a desired purpose." [7, p. 1081] A system consists of three things:

1. Inputs
2. Outputs
3. State

The state of a system is described by "[a] set of numbers such that the knowledge of these numbers and the input function will, with the equations describing the dynamics, provide the future state of the system." [7, p. 1080] The inputs and state determine the outputs of the system. Systems are often considered in theory where an assumption is applied to the theoretical systems. When a system is theoretical, *only* a system's state and inputs determine its outputs. Interference from the outside world is eliminated.

Systems outputs are either steady-state or transient. If the state does not change for a sufficient amount of time, the output is called the steady state output for the given inputs and state. If the input is changed, either the output or the state may change, or both. For many systems, the output will not change instantaneously to a change in input, but will change over time. The output of a system in this transition phase is transient. A system that has left one steady state but has not arrived at the subsequent steady state is in a transient state.

A theoretical system is often a model of something in the real world. Establishing the theoretical properties to understand the system requires measurements. Often the steady state of a system accounts for a majority of the systems operation and

theoretical models are best approximated by steady-state measurements.

2.6 Military Flight Operations Quality Assurance Derived Fuel Modeling for the C-17

A linear regression model to predict specific range from C-17 flight recorder data was the object of Havko's research. [8]

Havko estimated specific range using the equation

$$\Theta = \frac{v_{adj}}{FF} \tag{4}$$

Where

$$\Theta = \text{Specific range in NMs per Klbs}$$

v_{adj} = Wind adjusted ground speed derived from GPS coordinates in NMs per hour

$$FF = \text{Total fuel flow in Klbs per hour}$$

The researcher fit a linear regression model to data derived from flight recorders:

$$\Theta = \beta_0 + \beta_1\alpha + \beta_2\alpha^2 + \beta_3\omega + \beta_4\omega^2 + \beta_5\alpha\omega \tag{5}$$

Where

$$\alpha = \text{Altitude in thousands of feet}$$

$$\omega = \text{Aircraft gross weight in Klbs}$$

Havko used 100 random samples of cruise sections of flights to fit the regression model and 100 samples from other flights to validate.

The author noted his model explained relatively little of the variance in terms of the coefficient of determination, or R^2. R^2 indicates the portion of the variance in the

dependent variable that was predicted by the independent variable. The R^2 of the prediction model with the validation set was 0.283. This indicates the portion of the variance in the dependant variable that was predicted by the independent variable. An apparent source of bias was also noted. The estimate for specific range during a segment spiked to unrealistically high values. This was due to a decrease in airspeed. The thrust was decreased by cutting fuel flow allowing drag to have a greater influence on the aircraft's motion. This observation highlighted a necessity to further process the data used in predictive models of specific range.

Havko used the model to estimate the specific range for the cruise sections of sorties. In one sortie, the estimated specific range had a significant positive offset. This offset can be seen in Figure 19 of the research paper, and Figure 8 below. The researcher speculated this was a section of deceleration required by Federal Aviation Administration regulation. This period of flight, therefore, is not consistent with Peckham's definition of cruise flight. [5, p. 5] This may be a problem with other flight segments as well. Whatever the cause of airspeed adjustments during cruise sections of flight, these errors are hazards when building accurate predictive models.

Many modeling techniques and machine learning algorithms can make good predictions when there is variation in the data set. This is not the case, however, when measurements are known not to fit theory. In Havko's research paper, an example of how the data that is incompatible with theory introduced bias in a predictive model is clearly seen.

Fuel flow that approaches zero in a phase of deceleration results in an instantaneous calculation of specific range that is unrealistically high. It is high because $T \neq D$. Thrust is much less than drag which results in deceleration. This is a transient response. The linear modeling technique used in Havko's research is sensitive to data with samples that greatly deviate from the assumption of steady flight in

smooth air.

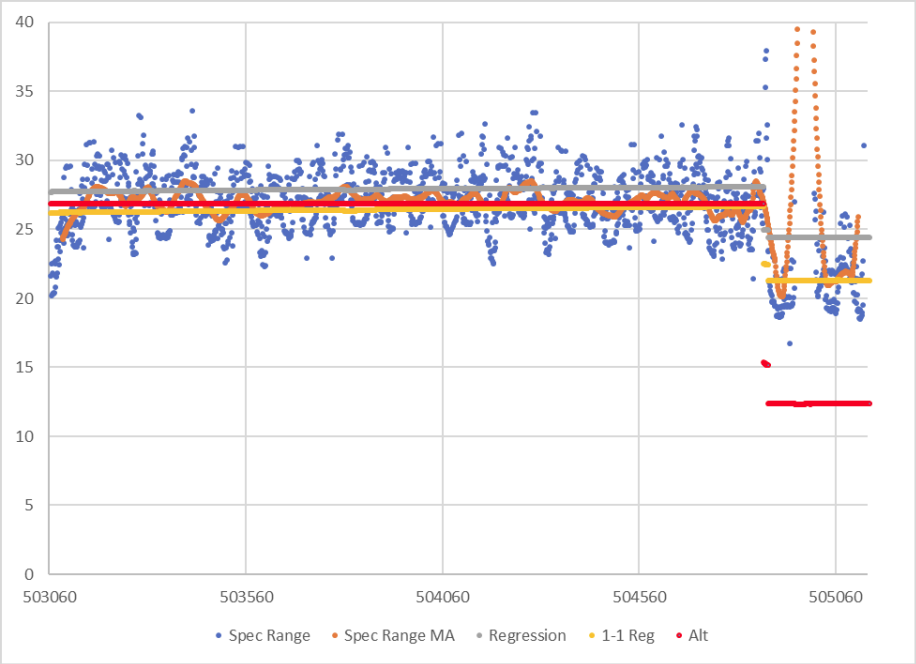

Figure 8. Calculated instantaneous specific range is in blue. Of note is the segment of flight where the specific range estimation is well above 40 nautical miles per 1000 pounds of fuel on the right side of the graph. [8, p. 37] The deviation can better be seen in orange which is a moving average of the blue points. The orange points deviate substantially from reasonable specific range values.

2.7 Models

For some models, the purpose is to be an accurate predictor. Models designed to make accurate predictions are called "predictive models." Predictive modeling applies tools that

> take our current information, sift through data looking for patterns that are relevant to our problem, and return answers. The process of developing these kinds of tools has evolved through a number of fields... and has been called 'machine learning,' 'artificial intelligence,' 'pattern recognition,' 'data mining,' 'predictive analytics,' and 'knowledge discovery.' [9, p. 6]

Vocabularies and word usage from different fields mix in this domain-agnostic field of predictive modeling. For clarification, this thesis will tend toward using terms typical for those who describe modeling in terms of machine learning. "Machine Learning is the science (and art) of programming computers so they can *learn from data*." [10, p. 4] This effort applies machine learning techniques to build predictive models.

"Predictive modeling is the process of developing a mathematical tool or model that generates an accurate prediction." [9, p. 2] Selecting an appropriate modeling technique is the goal of the modeling process. It is important to test multiple modeling techniques because, "in some sense, no [modeling] algorithm is universally any better than any other." [11, p. 116] Additionally, "it is seldom known in advance which procedure will perform best or even well for any given problem." [12, p. 350]

There are many modeling techniques. Without knowing the particulars of which technique may work best, a beneficial starting place is to take advantage of "off-the-shelf" modeling techniques. Each technique has some known strengths and weaknesses. [12, p. 350] Off-the-shelf techniques include neural nets, trees, and k-nearest neighbors. These off-the-shelf techniques as well as their strengths and weakness are considered in the relevant sections below.

Training is a term that describes the process by which a machine learning technique *learns* to make its prediction. Machines learn from data. There are two categories for data sets and, therefore, two categories of the machine learning algorithms that make use of them. The algorithms are either supervised or unsupervised.

> Unsupervised learning algorithms experience a dataset containing many features, then learn useful properties of the structure of the dataset ... Supervised learning algorithms experience a dataset containing features, but each example is also associated with label or target. [11, p. 102]

The algorithms used to make predictive models in this effort applied a supervised

machine learning approach. These *machine* learning approaches depend on the quality of the features and labels in the data. Machines in this context are synonymous with computers. Computers (machines) are programmed. What makes machine learning unique from other computer applications is summarized by Chollet. He describes machine learning as a new programming paradigm. Classical programming takes rules and data as inputs and outputs answers. Machine learning takes data and answers as inputs and outputs rules. [4, p. 5]

A doctrine held among programmers is that computers will only do exactly what you tell them. In machine learning we give an algorithm "truth" data and the machine makes a model. Machine learning techniques make models that are only as good as the data on which they are trained. Data is often made from measurements of values. A measured variable must be distinguished from the measurement to consider the quality of data. Measurements can be said to have a "degree of goodness." Measurements are "influenced by a number of elemental error sources." Some sources of error include: [13, p. 8]

1. errors in the standard for calibration
2. errors is the calibration process
3. variations in ambient temperature
4. variations in humidity
5. variations in pressure
6. variations in electromagnetic influences
7. unsteadiness in the "steady state" phenomenon being measured
8. imperfect installation of the transducer

An example in Coleman's text illustrates potential problems with measurements. In the example, students were asked to read the temperature from an analog thermometer immersed in a container of water to the nearest tenth of a degree. The thermometer was biased and "read high." The students were unaware of the bias. The true temperature was 96.0°F and the average value read by students was 97.2°F. To determine whether this error is problematic depends on how the measurement is used and for what. There is a condition where this error is harmless. Assume the thermometer always reads high by the same amount. If the students are trying to measure the difference between a before and after a physical change to the water takes place, the bias on the thermometer will cancel in the subtraction. Alternatively, if the air pressure is being estimated by measuring the temperature at which water boils, the actual temperature is needed and the bias will be a problem. More, the measurement will be applied to an equation that may exacerbate the error.

When variables are not measured directly but calculated from one measured value, the variable is said to be the result of a "data reduction equation." When these kinds of equations are used in experiments, Coleman explains, "we must consider how the systematic and random uncertainties in the measured variables propagate... to produce the systematic and random uncertainties associated with the result." [13, p. 21]

In this research the properties of a system are estimated by measurements and data reduction equations. An aircraft in flight is a system. The inputs of the system come from weather and other descriptions of the conditions through which the aircraft is flying. Other inputs come from controls like the throttle and yoke. The outputs are the airspeed, altitude, and other measurable parameters of flight. The inputs are always changing in unpredictable ways. A real-world aircraft system *never* reaches a true steady state. Data quality for models must be considered and improved where

reasonable. Once this is done, the data must be methodologically applied to create a predictive model.

A common hazard with data based modeling is that a model may be built that perfectly predicts the data set on which it is trained and make predictions with poor accuracy on new information. Models that do this are said to be overfit. An overfit model does not generalize well. An opposing hazard is a model which makes less accurate predictions than might be possible. Such a model is said to be underfit. Model tuning is the process by which Goldilocks settings are selected.

A Goldilocks model has hyperparameters that make the most accurate prediction possible given the data. Hyperparameters are "settings that we can use to control the [modeling] algorithm's behavior." [11, p. 117] Not only do some modeling techniques perform better than others, different variations of each techniques perform better. For this reason, the model selection process includes tuning hyperparameters for each technique. The Goldilocks hyperparameters must be estimated from the data.

A Goldilocks model's accuracy is limited by the technique used and "Bayes error." Bayes error is the "error incurred by an oracle making predictions from the true [probability] distribution..."[11, p. 113] The oracle is a predictor that knows the true probability distribution that generates the data. A model with perfect hyperparameters can do no better than an oracle.

The accuracy of a selected model is measured by a "test set" of data that has not been used in any way to create the model or select a modeling technique. The test set is unseen data. Model selection is done on the "training set." Misuse of the test set may result in false confidence in a model. The Goldilocks hyperparameters for each modeling technique must be done using only the limited amount of training data. To estimate modeling techniques' accuracy and ability to generalize before applying the test set, surrogate sections of the training set may be used as pseudo-test data.

This partition of the training set is called a validation set. A data-efficient validation approach is k-fold cross-validation.

K-fold cross-validation "involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set and the [model] is fit on the remaining $k-1$ folds." [14, p. 181] The "validation set" is the data taken from the training set to stand in as a surrogate for the test set. It is pseudo-unseen. An example of how this might work would be to split the training data into $k = 10$ parts. For each modeling technique, ten models would be built with $\frac{9}{10}$ of the data. For each of the ten models, the accuracy would be tested on the remaining $\frac{1}{10}$ of the data. Averaging the accuracy of the ten models gives a good estimate of how the modeling technique is likely to perform on the test data.

Linear Regression.

Linear regression is a family of modeling techniques. Training a linear regression model arranges the training data into an input matrix X with truth data in a vector y. A weight vector, w, is chosen to satisfy the following expression. [15]

$$\min_{w} ||Xw - y||_2{}^2 \tag{6}$$

This value can be found directly by applying the equation, assuming the inverse exists.

$$w = (X^T X)^{-1} X^T y \tag{7}$$

Where M^T is the transpose and M^{-1} is the inverse of a given matrix M.

To use w to predict a new output, $\hat{y}$ from a new vector input of independent variables, x_{new}, the vectors are simply multiplied. The result, $\hat{y}$, is a scalar.

$$\hat{y} = x_{new}^T w \tag{8}$$

The resulting model is always a linear combination of X. This makes the model effective for representing linear relationships between X and y and less accurate for relationships poorly expressed by matrix multiplication. There are techniques that may be used to enable linear regression to represent more complex relationships. One is to add squared and interaction terms of the original inputs to the vector X. Havko used this technique when he made a prediction model based not only on the altitude and gross weight, but also on altitude squared, gross weight squared, and gross weight times altitude. [8] Havko's β values were the elements in the weight vector w.

K-Nearest Neighbors (KNN) Regression.

K-nearest neighbors is a modeling technique that references data points. The constant k is the number of neighbors used to estimate the output value for an input. An algorithm finds the closest k data points for an input x. The average of the k neighbors is returned as the output value. Euclidian distance is a typical method to calculate the distance to a neighbor. It may be useful to scale the data when the units for different variables in X are not evenly spaced or are dissimilar units. A variation on k-nearest neighbors is to weight the average of the neighbors based on the distance to that neighbor.

This modeling technique may be best to model some phenomena as it directly applies the training data to estimate the prediction value. There are some potential drawbacks, however.

A potential drawback of a nearest neighbor model is that it is discontinuous. As X is changed, one or more neighbors will be dropped for the point estimate and as many neighbors added for the replacement set. This causes a discrete jump in the

prediction value for a change in X proportional to the difference in lost and gained neighbors' values.

Another potential shortcoming of this modeling technique is the "curse of dimensionality" has a greater impact on KNN models than other techniques. The "decrease in performance as the dimension increases is a common problem for KNN, and results from the fact that in higher dimensions there is effectively a reduction in sample size." [14, p. 108]

A final concern for this modeling technique is its computational complexity on large data sets. Each estimate can require as many calculations of distance as there are data points. There are ways to decrease the computational cost, but the cost is significant for large numbers of data points. Unlike many models which take more computational time to build and less time to produce an estimation for a given x, predictions take significant computation time.

Tree Methods.

Forest regression comes from a family of tree models. It takes advantage of several ideas. A tree model is a decision based approach which attempts to split data into homogeneous groups. [9, p. 370] A decision tree is a series of binary splits based on features-value pairs. For each split, the best feature and decision boundary for the feature is *typically* selected to segment the region of data. This process continues until a stopping condition is met. Two such stopping conditions are when the algorithm has met a maximum number of splits or a minimum number of data points remain in each leaf. Trees perform poorly compared to other supervised techniques. Combinations, or ensembles of trees compensate for shortcomings of single-tree models. Bootstrapping and bagging are methods used to combine trees to make improved predictions.

Bootstrapping is a process of creating multiple data sets by randomly selecting samples from the original data set *with replacement*. These are often made to be the same size as the original data set. This decreases the variance in the bootstrapped data sets. Bootstrapping can be used to make better estimations of statistical parameters of a data set. Models can also be created using bootstrapped data. Bootstrap aggregation or "*bagging* averages" is a prediction over a collection of bootstrap samples that reduces variance. [12]

In bagged multi-tree regression, an ensemble, or collection, of trees are made from separate sets of bootstrapped data. The estimation for the ensemble of bagged models comes from the average of the trees' predictions. For each split, each tree selects the best feature-value pair.

A random forest is an additional improvement to ensembles of bagged trees. In a random forest an adjustment to the tree building algorithm is made to decorrelate the trees. Decorrelated trees output predictions that vary more widely because the splits are based on different data. Each split in a given tree, as usual, consists of a feature and a value. In *typical* trees the feature is selected based on the best expected split. The split feature for trees in a random forest is the best of a random subset of the features. The best feature for a given split in a given tree is often not used. The result is often a better ensemble from the sets of bagged data.

Artificial Neural Networks (ANN).

Perceptrons are mathematical functions that are inspired by the neurons found in biological brains and make up the "nodes" of a neural network.

> The term *neural network* has evolved to encompass a large class of models and learning methods... There is a great deal of *hype* surrounding neural networks, making them seem magical and mysterious... they are just nonlinear statistical models. A neural network is a two-stage regression

> or classification model typically represented by a network diagram. [12, p. 292]

Forward propagation is the use of an ANN to come up with a prediction. If a true value can be compared to the prediction value, "back propagation" may be used to find an error gradient. An algorithm like "Adam" makes use of the gradient to update the network's parameter values. [16] Adam has been shown to be efficient in terms of computation time and the number of forward and back propagation iterations to be performed before a network is "fully trained."

One subset of neural networks is that of dense neural networks. Dense networks can be described in terms of "layers" of perceptrons. Each layer has a width value representing the number of perceptrons in the layer. Dense layers are layers in which every input to the layer is connected to every output from the previous layer. A dense neural network is made from dense layers.

As with other modeling techniques, ANNs can overfit the training data. "Machine learning algorithms will generally perform best when their capacity is appropriate for the true complexity of the task." [11, p. 109] One way to prevent a model from overfitting the data is by limiting its capacity. Though capacity is a qualitative concept, the capacity generally scales with the number of perceptrons and parameters in a network.

There are many hyperparameters that can be adjusted for building neural networks. In a simple multi-layer perceptron model one can "change the number of layers, the number of neurons per layer, the type of activation function to use in each layer, the weight initialization logic, and much more." [10, p. 272]

Géron concludes, however that "for many problems, you can just begin with a single hidden layer and you will get reasonable results. It has actually been shown that a multi-layer perceptron with just one hidden layer can model even the most complex functions provided it has enough [perceptrons]." [10, p. 273]

Another useful approach is to apply the "Universal Workflow of Machine Learning" outlined by Chollet. [4] This workflow includes developing a model that overfits, ensuring it has the capacity to fully represent the input-output relationships, followed by using regularization techniques and tuning hyperparameters to Goldilocks settings. Approaches to accomplish each are listed. [4, pp. 111-115]

To build a model that overfits:

1. Add Layers.
2. Make the layers wider.
3. Train for more epochs.

Techniques that may be tried to regularize the model and tune hyperparameters include:

1. Add dropout.
2. Try different architectures: add or remove layers.
3. Add L1 and/or L2 regularization.
4. Try different hyperparameters.
5. Optimally, iterate on feature engineering by adding new features or removing apparently uninformative features.

Géron notes that selecting a model with a known high capacity and applying dropout is a common "stretch pants approach" where, "instead of wasting time looking for pants that perfectly match your size, just use large stretch pants that will shrink down to the right size." [10]

Models are trained over epochs. An epoch is "An arbitrary cutoff, generally defined as 'one pass over the entire dataset,' used to separate training into distinct

phases, which is useful for logging and periodic evaluation." [17] Within each epoch are several batches. A batch is a set of samples that is computed independently from other batches. The gradient for the batch is the sum of the individual prediction gradients. A batch will result in one calculated gradient and one update to the network.

Dropout is a technique to prevent overfitting by randomly removing perceptrons during training. Which perceptrons are removed changes from batch to batch. Dropout effectively trains the network as several smaller networks with different architectures. After training, the final network makes predictions without dropping perceptrons. Weights inside the network are adjusted so that predictions from smaller dropout networks scale to be used in the full network. Authors of this technique suggest that a 50% perceptron dropout rate on hidden layers was optimal for mitigating noise in autoencoders. [18]

Modeling Tools.

Various software tools are available to build machine learning models. In this research, models were created using the Python programming language. The libraries most heavily used in the modeling process were the software libraries Scikit-Learn [15], Keras [17], and SciPy. [19] Visualizations used the Python libraries Matplotlib [20] and Seaborn. [21] Python tools were selected because they met the needs to build machine learning models and the researcher was familiar with the language and tools.

Scikit-Learn.

Scikit-Learn is a free software machine learning library. It was started in 2007 as a Google Summer of Code Project. [15] The goal of Scikit-Learn was to bring "machine learning to non-specialists using a general-purpose high-level language." [15] A notable

trait of this library is its consistent application of the "fit" and "predict" functions for its variety of machine learning estimators. This makes implementing different models relatively easy. This fit and predict scheme is also common in the Keras library.

Keras.

Keras is a deep-learning application program interface that has user friendliness, modularity, easy extensibility, and the ability to work with Python. [17] This library allows for a relatively easy interface to build and train ANNs.

SciPy.

Scipy is a library that interfaces with many other Python libraries. It is a "collection of algorithms and domain-specific toolboxes including signal processing, optimization, statistics and much more." [19] A key utility for this research was the signal processing sub-library which was used for data processing.

Matplotlib.

Matplotlib is a library for plotting 2D arrays in Python. It is open source and used extensively. It was made to emulate MATLAB plotting capabilities. [20] Many of the plots in this work used this library.

Seaborn.

Seaborn is a Python software library based on Matplotlib. It serves to provide a high-level interface for making attractive and informative statistical graphics. [21] Many of the figures in this work were created using Seaborn. Seaborn has refined many common data visualizations to be coded quickly while looking professional.

2.8 Related Work

There are several avenues of research related to this effort. Many are attempts to model fuel consumption using flight recorder data.

Modeling Fuel Fllow-Rate.

Baklaciouglu used genetic algorithms to select artificial neural network architectures to predict fuel flow given other parameters from flight data. [22] There is cause for concern with this paper. First, this is a very computationally expensive approach. It seems like there is little justification for applying an advanced and complex technique when it is possible a simpler approach would have yielded as good or better results. The author used two parameters, flight altitude and true air speed as inputs. Fuel flow was the output. A second concern is with the data used. Backlaciouglu used 1,234 data points. How many separate flights this is from is not stated. Sampling frequency of the flight recorder is also not included. If the sampling rate was one second per sample, this data may be from a single 20-minute flight. Splitting this data into training and validation sets is problematic because it is unlikely the model will have any ability to predict fuel flow for other flights.

Analysis of Flight Fuel Consumption.

The authors used data from Airbus A-330 flight recorders to estimate fuel consumption using linear regression models. [23] Their goal was to identify the main factors that contribute to fuel consumption so that flight plans may be optimized for fuel efficiency. The data was gathered from a company and contained one year of operation. Flight phases were considered separately. The researchers found a high correlation with takeoff weight for fuel consumption during takeoff. The next phase they evaluated was climb. They used linear regression on the values of takeoff weight,

atmospheric temperature, climb rate, and climb distance. They found the highest correlation coefficient to be in climb rate and fuel consumption with a correlation coefficient of .83. Of special interest is how the authors modeled fuel consumption performance for the cruise section. Though specific range is not mentioned, they measure performance in terms of kilograms of fuel per kilometer, an appropriate unit for specific range. They noted that performance tended to increase for the cruise sections at higher altitude. They also separated their cruise data analysis into two categories: one with eastbound flights and the other westbound. The reason for this separation was to get more consistent data for each due to air traffic regulations in the country in which the flights occurred as well as persistent wind effects. The consideration of wind indicates their range was not air miles, but ground miles and was subject to variation from wind. The final analysis considered fuel consumption by descent rate, finding a fuel optimal descent rate for the A-330 to be about 2,000 feet per minute.

Fuel Consumption Estimation of Cruise.

This article had the same group of authors as "Analysis of Flight Fuel Consumption Based on Nonlinear Regression." The authors attempted to build a fuzzy neural network model to estimate fuel consumption. [24] The fuzzy network may combine the advantages of a neural network to conform to non-linear relationships with the benefit of dealing with low precision data. Their data came from Chinese usage of the Airbus A-330. The input to their model is a flight's cruise distance, Mach number, cruise time, and starting weight. The output of their model is the total fuel consumption for the cruise segment. Their cruise segment seems to be defined as the portion of flight between when the aircraft stops climbing to descent. The root mean squared error of their selected model was 6.29 kilograms (13 pounds). Between the translation from Chinese and the brevity of their modeling processes description, it is

unclear whether the training and test sets were split in a way that made them more correlated than a prediction of a new flight may be. It is also unclear where the data comes from as it is from 'PEK-SHA route 435 times flight." There were 355 sets of data used for training. This may mean the data was taken from 435 flights on the same route or 435 is a route identifier, but which is the case is unclear.

C-5 Fuel Efficiency Through MFOQA Data Analysis.

The study in this thesis used flight recorder data on C-5M military cargo aircraft to obtain a way to estimate the appropriate fuel load for individual aircraft more precisely. [25] For the C-5M aircraft, the total fuel is calculated by computer and multiplied by 1.04 to create a 4% safety margin in addition to the estimated fuel load. This, especially for long flights, may be costly. One source of uncertainty for which the safety margin compensates is engine degradation. Knowing, for a particular aircraft, the combined efficiency of its four engines may account for some of deviation from the expected fuel usage. Flight recorder data may be used to factor in the aircraft's engine degradation and decrease the safety margin multiplier.

Specific range was estimated with flight recorder data. These values for particular aircraft were compared to the technical instructions that illustrated specific range. Preprocessing was necessary to approximate specific range information from fuel flow data recorded in flight recorders. Local regression on time series information was applied. This was applied to the fuel flow values. The criteria for inclusion of a flight segment was one hour of cruise flight that deviated in altitude less than 25 feet. The first and final five minutes were removed from each of these 60 minute segments. The selection resulted in 40 cruise segments from one aircraft.

The research found that, for the given aircraft, the fuel flow through one of the engines is higher than the 4% safety margin. The other three were within the 4%

safety margin. The state of a given engine accounts for a significant portion of fuel required to ensure safe arrival at the destination.

III. Methodology

The chapter is generally organized into CRISP-DM [3] phases. Chapter II covered the Business Understanding phase and part of the Data Understanding phase. The more technical aspects of data understanding is where this chapter begins. Calculating specific range, though part of the Data Preparation phase, is significant enough to merit its own section. The procedure is subsequently explained. It is separated by section into the Data Preparation phase and the Modeling phase. The final considerations of the Modeling phase contain how the models are evaluated. Chapter IV contains results from the procedure and Chapter V contains information best suited in the Evaluation phase.

3.1 Understanding Flight Recorder Data (Data Understanding)

An introduction to the data is found in Chapter I. Important details of the data are discussed in that example. Possible segmentation of the data and the nature of the distribution are demonstrated in the chapter's figures. Methods to segment the data, the result of which is the many highlights in the figure, are shown by the example sortie. As in the example sortie, cruise segments were selected on binary values contained in the data. Some of the details are stated in the research paper by Havko. [8]

A modeling method similar to that used by Havko was applied. "Military Flight Operations Quality Assurance" (MFOQA) is the reason this data was gathered and stored. Flight data recorders captured this data in the first half of 2016. Raw data was then uploaded to a database. Once downloaded by a user from the secure online database, processing expanded the data from its flight recorder format to a time-labeled format. Havko further applied the criteria: "fuel flow is not blank, ground

speed above zero, pressure altitude above 10,000 feet, landing gear up, flaps and slats position less than 14 degrees, and pitch engage setting of 'Altitude Hold."' [8] A data reduction equation yielded true airspeed from indicated airspeed, altitude, and air temperature. The other variables used were sensor values. Each flight, or sortie, was broken into flight segments. Each segment contains continuous data where the processing did not result in internal segmentation. That is, each segment represents contiguous samples of flight data. One sample per second was the sampling rate. Havko used 200 segments in total.

Details of how the data was processed in previous research and the current process is illustrated in Table 1. Havko used a file format compatible with Microsoft Excel (Form: xlsx). In this format, different sorties were stored in separate files. Each file represented a unique sortie, though not necessarily a unique aircraft. Havko processed the data and accomplished his research in Step 5. This is the initial format of the data used in this research. The example sortie in Chapter I used data from Step 4, before non-cruise portions of flight were omitted from the data.

In Step 6, the test set was sequestered. Approximately every third sortie file was withheld until Step 11.

In Step 7, the remaining Excel files (training set) were converted to a single comma separated value (csv) file. Segments with fewer than 31 samples were omitted. All values computed by Havko were dropped, with the exception of true airspeed. The values retained for use in this research are listed. The items listed in 1-5 are parameters, calculated or captured by sensors. Items 6-8 were used for indexing.

1. Altitude (Thousand Feet)

2. Fuel Flow (Thousand Pounds per Hour)

3. Gross Weight (Thousands of Pounds)

4. True Airspeed (Knots)

5. Temperature (Degrees Celsius)

6. Sortie

7. Sortie Segment Number

8. Segment Index

Table 1. Steps to process and apply the data. The first column indicates the sequence in which data was separated, dropped, or altered. Location describes who had control of the data or in which media the data existed. Form indicates the data format or file extension. The flt label indicates the flight recorded format, xlsx indicates the file extension compatible with Microsoft products, and csv indicates comma separated value. The 'Data is' column indicates whether the data is processed (P) or unprocessed (UP). The column "Data" indicates what portions of the data exist in each step. "Actions taken" is a short description of what was done to the data to put it in the state indicated by its row.

Step	Location	Form	Data is	Data	Actions Taken
1	Flight Recorder	flt	UP	All	
2	Web Database	flt	UP	All	
3	Havko	flt	UP	All	
4	Havko	xlsx	UP	All	
5	Havko	xlsx	UP	Cruise (C)	Seg, C Only
6	Author	xlsx	UP	C and Train (TR)	Train Only
7	Author	csv	UP	C and TR $>$ 30 S	Drop Seg $\leq$ 30 S
8	Author	csv	P	C and TR $>$ 30 S	Processing
9	Author	csv	P	C and TR $>$ 30 S	Val (no change)
10	Author	csv	P	C and TR $>$ 30 S	Models Trained
-	-	-	-	-	-
11	Author	xlsx	UP	C and Test (TE)	
12	Author	csv	UP	C and TE $>$ 30 S	Drop Seg $\leq$ 30 S
13	Author	csv	P	C and TE $>$ 30 S	Processing
14	Author	csv	P	C and TE $>$ 30 S	Add Pred

When the files were combined into one, the Sortie index was assigned so data points from a unique flight could be identified. Each segment in a given sortie had a unique value in that sortie. Sortie Segment Number is that index. Each combination of Sortie and Sortie Segment Number is unique for each segment. The Segment

Index was added so that each segment had a unique index value. Data in this step has not been altered from either Havko's research or the format that was the result of unpacking from the flt format. With the exception of keeping the True Airspeed values, the only alterations to the data from its flight recorder format was by omission of samples. Because the values are unaltered, and only one computed, the data in this step may be referred to as "unprocessed."

In Step 8, the filtering was accomplished, specific range was calculated for each sample, and an index indicating which of the k-folds into which each sample belonged was added. This is the format of the data used for validation in Step 9 and for training in Step 10. This accounted for a majority of the the work done to prepare data for research. Training data that is said to be "processed" is that which results from Step 8.

The test data in Step 11 is the same as it was before Step 6. Formatting accomplished on the test data in Step 12 is identical to the formatting done on the training data in Step 7. Test data that is said to be "unprocessed" is the result of Step 12.

Alterations to data applied to the training data in Step 8 were also applied to the test data in Step 13 with one exception. No cross validation indices were needed. Test data from this step is "processed."

For evaluation on the test set, each sample had a prediction from each model. These predictions were simply added to the data in Step 14. The values that were present in Step 13 were unaltered.

A number of details were captured on the unprocessed data. This includes histogram visualizations of the number of samples in each segment. Additional details collected for each the training and test sets are listed. The excluded values are those segments that are omitted due to length in Table 1 Steps 7 and 12.

1. Total Sorties

2. Total Segments Included for Modeling

3. Total Samples Included for Modeling

4. Total Segments Excluded

5. Total Samples Excluded

3.2 Calculating Specific Range

It is useful to compare theoretical cruise consistent with Peckham's observations and assumptions with unprocessed flight segment data. In the unprocessed data, altitude is sufficiently constant due to Havko's requirement that the altitude hold be engaged. [8, p. 20]. Weight is approximately equal to lift. There are concerns as to whether the unprocessed data sufficiently meets the equality requirement of thrust and drag.

Peckham's $L = W$ requirement is sufficiently met by use of the altitude hold function of the autopilot. Peckham's $T = D$ requirement is violated by the auto thrust function of the autopilot. [5]

Thrust is controlled by fuel flow. In the unprocessed data, fuel flow oscillates during the cruise segments. This is not behavior that would be seen in theoretical steady state flight through smooth air. Some of the oscillation is due to the auto thrust control system. When auto thrust is used to maintain airspeed, the control system varies the fuel flow to maintain the desired airspeed. Without the auto thrust engaged, any changes in thrust are from the crew's throttle adjustments. The human adjustments may oscillate, but at a lower frequency than that of the auto thrust system. Whether these oscillations are from the autopilot or human pilot, these changes violate assumptions about cruise flight. Specifically, these changes violate the assumption about estimating specific range that "thrust is equal to drag." [5,

p.4] The reason $L = W$ is sufficiently met and $T = D$ is violated has to do with the method of estimating specific range.

This research uses training values for temperature adjusted specific range (TASR). TASR is the calculation for specific range as used by Havko based on direction from the C-17 general technical order. [8] The first step in calculating TASR is to calculate specific range, SR, where

$$SR = \frac{TAS}{FF} \tag{9}$$

In the equation, TAS is true airspeed and FF is fuel flow. This value is adjusted slightly based on temperature to get the $TASR$. The adjustment is detailed in Section 3.3. The parameters in the reduction equation to estimate SR must be considered.

The fuel flow variable that is described in theory would be constant and result in a constant airspeed. To find how fuel flow that suits theory may relate to the measured fuel flow, it is useful to consider the control system that sets the fuel flow. A naïve but useful understanding of this control system will be explained. The system takes a reading of the airspeed, compares it to the desired airspeed, and adjusts the thrust to decrease the difference between the desired airspeed and measured airspeed. The naïve control system has two states: too slow and too fast. In the too fast state, the control system decreases fuel flow to cause the measured airspeed to decrease towards the desired airspeed. Once the aircraft slows sufficiently, the airspeed and desired airspeed will be momentarily equal. However, this simple system overshoots the desired airspeed and enters the too slow state. The control system now increases fuel flow to increase airspeed.

The naïve control system does not have drag as an input, and therefore at no time explicitly attempts to maintain an equivalence between thrust and drag. However, the thrust that would maintain equivalence between thrust and drag for the desired

airspeed must lie between the control system's resulting maximum and minimum fuel flow setting, presumably in the vicinity of the mean. The real control system on the C-17 is more complex than the one described here, but the effect on the fuel flow variable is comparable to that of the naïve system. A visualization of fuel flow for the duration of a flight segment can be seen in Figure 9. It is not unlike what the naïve control system would produce.

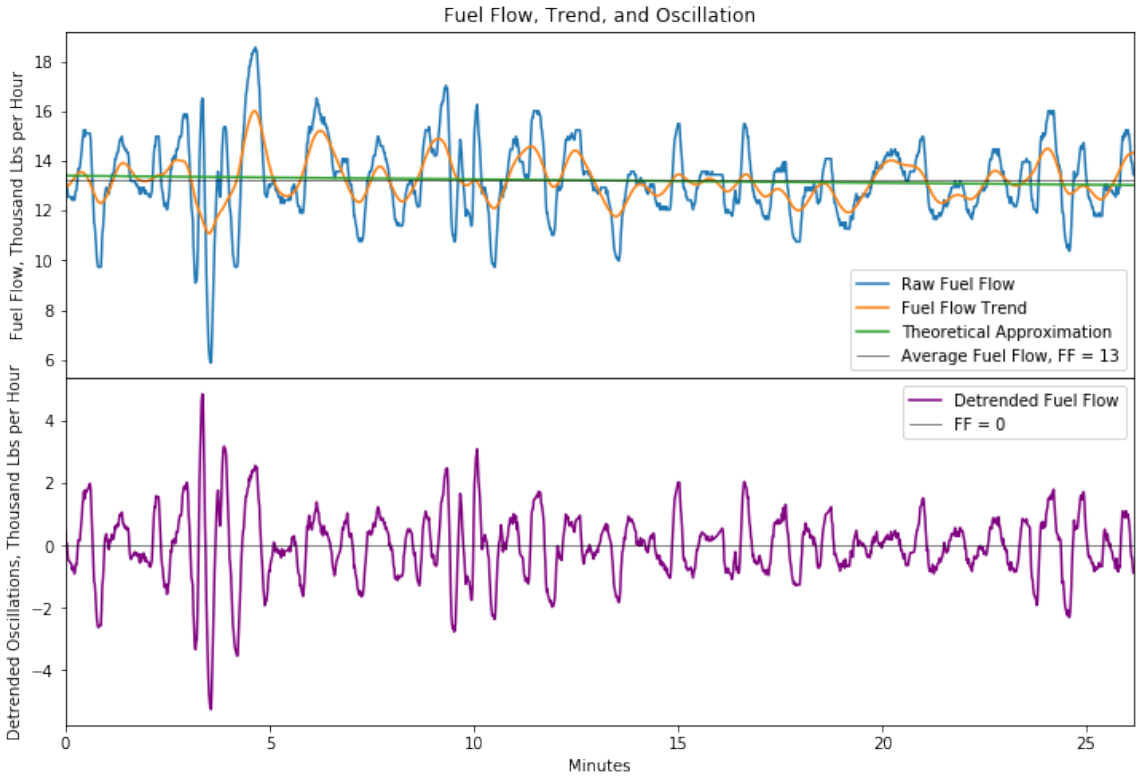

Figure 9. The top graph shows the raw total fuel flow data. The "Fuel Flow Trend" (orange) is the application of two moving average filters to the "Raw Fuel Flow" (blue). This filter is applied once 30 samples forward in time and once the same number of samples backwards in time. The green line is an approximation of the fuel flow values that may be expected if the aircraft were flying in theoretical air with a constant fuel flow. The average value for the fuel flow over the duration of the cruise segment (black) is included so the slight slope in the conjectured fuel flow line may be more easily identified. The bottom graph shows the difference of the raw signal and the filtered signal. The filtered data can be said to show a trend. Subtracting the trend results in isolating the non-trend data (purple). If the non-trend data is undesirable, it is often referred to as noise.

Oscillation, or even stochastic behavior, is not necessarily problematic for some

modeling techniques. Most such techniques used in practice result in predictions that tend toward the mean of the distribution. These models are reasonably accurate as long as the noise is symmetrically distributed about the mean. However, the error shown in Figure 9 is not evenly distributed, i.e. the error contains bias. The noise exists *not* in the prediction value, but in a value used in a data reduction equation. Consider error to be the difference between a transient state fuel flow (where $T \neq D$ or $W \neq L$) and that which would result in equity of the fundamental forces. Assume the true value of specific range is that which is calculated from the latter. Because of the bias, the sign of the error in fuel flow affects the magnitude of error. An incorrect approximation of the variable results in an error that is asymmetric and significant. The resulting bias in specific range would result in an inaccurate model. This is because the calculated training values for the model itself would be incorrect. Put another way, the truth data on which the model is trained would be invalid. The consequence of using a value of fuel flow that deviates from a theoretically appropriate value to calculate specific range is shown below.

In the case where Specific Range (SR) is estimated as in Equation 9, the effect of erroneous measurements can be demonstrated by replacing the fuel flow value with a measurement. That measurement can be taken as the fuel flow sample, FF, plus an error term. The new term is $FF + e_1$ where e_1 is the error term. Of interest is the effect of this error on specific range. We can represent this error by replacing SR with its true value and error term, e_2. For demonstration, assume no error in the measurement of TAS. The new equation is

$$SR + e_2 = \frac{TAS}{FF + e_1} \tag{10}$$

The real values of TAS and FF can be used in place of SR and the equation solved for e_2 leaving

$$e_2 = \frac{TAS}{FF + e_1} - \frac{TAS}{FF} \tag{11}$$

For large positive e_1 values, e_2 approaches $-\frac{TAS}{FF}$, while for large magnitude negative e_1 values, e_2 approaches infinity. The range of the error is approximately $\frac{1}{2}$ the amplitude of oscillations in the fuel flow value if the steady state value is in the vicinity of the mean. Oscillations, if they do not represent the thrust-drag equilibrium value of fuel flow, cause bias. This can be seen in Figure 10. If the green line in Figure 9 is the best fuel flow value, the range of e_1 is at least from -7 to 4 thousand pounds per hour. For these values, the resulting e_2 values would be approximately 8.5 and -3 nautical miles per 1000 pounds of fuel. At the extremes of the range, the resulting absolute error, $\mid e_2 \mid$, for the same magnitude error in FF, $\mid e_1 \mid$ is more than double.

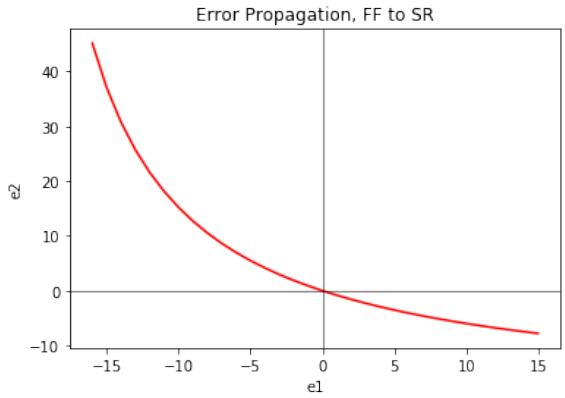

Figure 10. Error in specific range, e_2 for a given error in fuel flow e_1. The fuel flow values and true airspeed are typical of recorded values. The error when e_1 is bounded when positive and unbounded when negative.

Three solutions to this problem were considered and summarized here.

1. Omit data that deviates from steady state flight.

2. Eliminate data where auto thrust is not engaged.

3. Apply a moving average filter on FF, which was selected.

Omit Data that Deviates from Steady State.

Omitting data that violates steady state flight described by Peckham [5] was considered. The solution follows logically from previous processing applied to the data. Previous processing of the flight data omitted sections where the altitude hold was not engaged. The remaining data contains a flight profile where a control system was articulating control surfaces to maintain a state of equilibrium between lift and gravity, violating assumptions. It is reasonable to consider eliminating subsequent data that also violates cruise definitions. The downside of this approach is the difficulty in establishing a rule for which data to include. A case can be made that all of the real flight data violates one more definitions of steady state to some degree. Attempts to create an algorithmic threshold for omitting data were explored.

An algorithm that identified data for omission was created. For a given parameter in the flight segment, it calculates the standard deviation for a given "window," or contiguous set of data points. The window moves over each set of contiguous samples. The window size is the number of samples considered. If the standard deviation in the window is above a given threshold, the samples in the window is omitted. This algorithm can be applied to airspeed, fuel flow, or both. Window size and the corresponding standard deviation thresholds are tuning parameters. Qualitative tuning was attempted with some success. Data that appeared to be deliberate adjustments to flight were omitted and only some data that appeared to be either adjustments or exceptionally turbulent flight were omitted. Though this approach still shows promise, it was not pursued for two reasons.

First, data from turbulent flight, potentially omitted by this technique, may be important data to preserve. The core of this effort is to predict SR for real-world flight. Turbulence may have an important effect on SR. The effect from turbulent flight cannot be captured by a model where turbulent data is omitted. Turbulent

flight may have an impact on resulting FF and TAS approximations.

Second, data omitted by rules of standard deviation may cause bias in the data due to circumstances that result large standard deviations. True airspeed is a result of a data reduction equation. Errors in the calculation of this value likely propagate more noise into this value at some ranges of operation than at others. For example, the amplitude of the oscillations in true airspeed may be greater at higher temperatures than low due to a combination of sensor error and how its propagates. Eliminating data by this method would unknowingly and disproportionately remove data from higher temperatures. Elimination of this data may introduce a bias that reduces the predictive power of a model for higher temperatures.

Omit Data for Which Auto Thrust is Not Engaged.

A similar option is to omit data where the auto thrust function on the autopilot was not engaged. This is a logical extension of processing already applied. All of the data used has altitude hold engaged. Therefore, it could be said of the data used in this research: "The control surfaces were being altered by the control system to maintain an equilibrium between lift and weight." If auto thrust is likewise required, an additional statement could be added: "The throttle inputs were being altered to maintain an equilibrium between thrust and drag." If the control system is effective, this well describes Peckham's description of cruise flight. This was not pursued for several reasons.

First, FF oscillates even when auto thrust is engaged. There is no reason to consider such data to be more theoretically ideal than FF under human control. Indeed, the converse may be true. Data where a human is in control of the aircraft may provide useful information in predicting SR that is lost when airspeed hold is engaged. It is reasonable to consider the frequency of oscillation under autopilot

control to be high compared to an aircraft under manual control. An argument may be made that flight from human input better suits theory. This is because fuel flow is often adjusted more gradually and less frequently by a human.

Second, if this constraint were made on the data, it would narrow the scope of the models predictions. Any confidence in a predictive model's accuracy would only apply to flight segments where auto thrust was engaged.

Apply a Moving Average (MA) Filter on Fuel Flow (Selected).

A moving average (MA) filter was considered most likely to reduce the error from oscillating FF values. The filtered FF data is sure to be a better approximation of that which would result in conditions where thrust is equal to drag in still air. Applied sparingly, using this approximation does not risk additional bias to the extent of the previous options. It is prudent to define a MA filter. The definition used is from Kamen. [6, p.32] Let the input-output relationship of an N-point moving average filter be defined by the equation:

$$y[n] = \frac{1}{N}(x[n] + x[n-1] + x[n-2] + ... + x[n-N+1]) \tag{12}$$

The values of $y[n]$ where $n < N$ are not defined. The output of such a moving average filter has N fewer data points than the original data. Additionally, the resulting information is shifted across the samples. That is, $y[n]$ is the average of the values $x[N-n+1...n]$. The value of $y[n]$ contains no information for values of $x > n$. This results in a phase shift. One way to mitigate that phase shift is by using the moving average filter twice: once forward, and once in reverse. The result of the pair of filters cancels the phase shift. [19]

There are trade offs to consider when applying a MA filter. A parameter for the filter is the value of N. The greater the number, N, the greater the reduction in

noise. However, the data that results from each iteration of a moving average filter is N data points shorter. If the output, y, improves with increasing values of N, this is paid for in the loss of the number of data points in y. Also, as N becomes large, the quality of data surely decreases as N approaches the length of X. In the single pass filter case where L, the length of X equals N, the result is a single value for y which is simply the average of X.

There was no obvious or simple way to quantify the results of different values of N. With the tradeoffs in mind, a conservative value of $N = 30$ was selected. The effects of such a filter can be seen in Figure 9. A majority of the raw fuel flow oscillation is still captured. This means two applications of the $N = 30$ MA filter results in shorter processed flight segments. The filtered segments are the length of the original segment minus $2N = 60$. Flight segments shorter than 60 samples (60 seconds) would have no value. To prevent significant losses of samples and include segments between 30 and 60 samples, a padding method was used. Padding means putting additional values at the beginning or the end of a set of contiguous samples. Padded samples are often zero, the average value of the sample set, or another suitable value based on the application. In this case the padding consisted of adding 30 samples to the beginning and end of each segment. Each set of padded data was the reverse order of the samples from the original set to which it was adjacent. In this case, the padding was applied for $N = 30$ at the beginning and end of each segment. For a flight segment S of length L with samples indexes $m..k$, the processed segment P has this relationship.

$$P = S_{N..m} \cup S_{m..k} \cup S_{k..L-N} \tag{13}$$

Segments with less than 30 samples were omitted. Results greater than 30 samples had the same number of samples as the original data. The results of the filter applied

to an example flight segment can be seen in Figures 9 and 11.

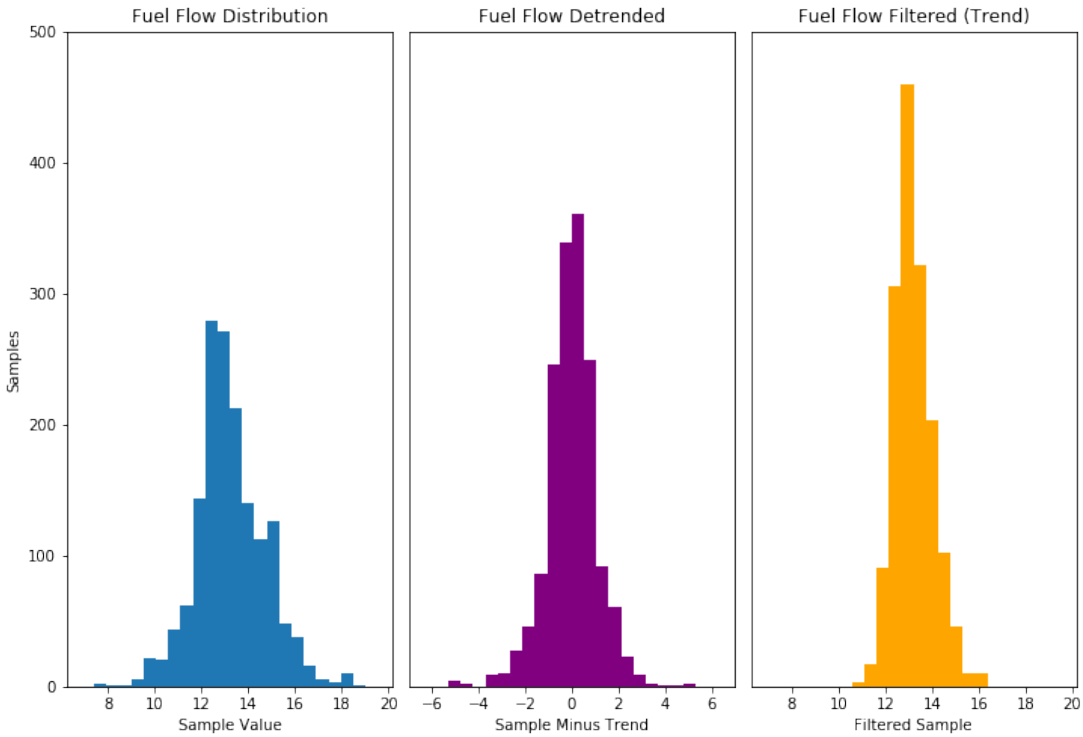

Figure 11. The distribution of fuel flow samples of the same flight segment used in Figure 9. As in that figure, the blue is the raw fuel flow data, the purple is the detrended data, the orange is the filtered data. The x and y ranges are the same on each graph for comparison. Conceptually, removing the variance represented in the purple distribution from the blue distribution results in orange distribution. The orange samples tend toward the median and have a smaller range.

The following exponentially weighted moving average was considered: [6, pp.45-47]

$$y[n] = \sum_{i=0}^{N-1} w_i X[n-i], \tag{14}$$

where

$$w_i = ab^i, i = 0, 1, 2, ..., N-1, \tag{15}$$

$$0 < b < 1, \tag{16}$$

and a is a normalization constant.

An exponentially weighted filter has a quicker response to time variations com-

pared to a moving average filter for a given N. In general, the moving average filter does a better job of removing noise. [6, p.46] A moving average filter was preferred as noise reduction was preferred and the ability to respond quickly to time variations was undesirable.

3.3 Procedure (Data Preparation)

The actions taken for preprocessing can be summarized in the following sequential process.

1. Filter Data
2. Add Calculated Values to Dataset
3. Add Standardized Input Values to Dataset
4. Make 10 K-Fold Cross-Validation Indexes

Filter Data.

Section 3.2 describes the method used to filter FF. The parameters of true airspeed, temperature, gross weight, and altitude were filtered for each segment, though not for the same reasons nor to the same extent as FF. These latter measurements were discretized which creates noise. The noise results from the error made from rounding. Filtering removes noise. The filtering on the flight segments is not expected to have a significant impact on modeling, but it is sure to improve the accuracy of the measurements by a small amount. A MA filter with a length of ten was applied twice. This filter was applied in forward and in reverse, once each, to avoid a phase shift.

Add Calculated Values to Dataset.

Instantaneous calculated values for TASR were calculated for each data point. TASR was calculated from SR according to the C-17 manual [2] where:

$$SR = \frac{TAS}{FF} \tag{17}$$

Adjustments to this value were made for the TASR based on the altitude. These adjustments were made based on the standard temperature. The value for standard temperature, ST, was calculated from the samples' altitude. If the altitude was greater than 36,000 feet, ST was calculated by the equation:

$$ST = 15 - (36 \cdot 1.98) \tag{18}$$

In the case where the altitude was less than 36,000 feet, the standard temperature was calculated using the equation:

$$ST = 15 - (Alt \cdot 1.98) \tag{19}$$

where

$$Alt = \text{Altitude}$$

The adjustment ΔT was made using the standard temperature and the sample's air temperature T:

$$\Delta T = T - ST \tag{20}$$

The final adjustment calculation resulting in TASR was:

$$TASR = \dot{SR}(1 + (.001 \cdot \Delta T)) \tag{21}$$

Add Standardized Input Values to Dataset.

For some of the modeling techniques, it was prudent to standardize the input features. If a model is said to have standardized inputs, the inputs were scaled to have a variance of one and a mean of zero. [12] This is a common and recommended practice when building KNN and ANN models. Standardized values for gross weight and altitude were calculated from the training set. The parameters were recorded to ensure consistent standardization on the test set.

10 K-Fold Cross-Validation Indexes.

Typical K-Fold Cross Validation will randomly select one of ten folds for each data point. This is an acceptable method when each data point is said to be independent and identically distributed. [11, p.108] This is not the case for this data. Data points that are adjacent in time are almost identical. Data points in a given segment are correlated. To decrease the impact of correlated data in validation estimates, consecutive data points were put into the same partition when selected for a fold. These resulting clumps are then randomly selected for a fold. Two minutes of flight, 120 data points, was the size of each partition.

Partitions from the same flight are still correlated, though less correlated than would be if the samples were not clumped. The assumption is that this will be sufficient decorrelation to consider the validation accuracy and estimate the Goldilocks hyperparameters of each modeling technique. This clumping is not done on a sortie by sortie basis. Instead is is directly applied to the ordered data set. The ordered set of data contains consecutive sorties. Each sortie contains consecutive segments. The flights are in no particular order. This means a clump may contain data from the end of one sortie and the beginning of another. The validation folds are identical for all validated modeling techniques and hyperparameters.

3.4 Procedure (Modeling)

The actions taken for model building and selection can be summarized in this sequence:

1. Make Naïve Model
2. Make Linear Regression Model and Record Validation Results
3. Make K-Nearest Neighbors Models and Record Validation Results
4. Make Bagged Tree and Random Forest Models and Record Validation Results
5. Make Artificial Neural Network Models and Record Validation Results
6. Down-Select Models
7. Evaluate Selected Models on Training Set

To compare models, there are many metrics of accuracy that may be applied. Root mean squared error was selected as the default value for comparison.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{22}$$

In this equation MSE is the mean squared error, n is the number of samples in a prediction set, y_i is the sample's value to be predicted, x_i is the prediction feature vector, and $\hat{f}(g)$ is the prediction function. The $RMSE$ is the square root of MSE.

$$RMSE = \sqrt{MSE} \tag{23}$$

A good quality of $RMSE$ is that it is in the same units as y_i which makes the results easy to interpret.

The coefficient of determination or R^2 is often used to evaluate models. This is often used in hypothesis testing and is usually said to describe how much of the variance in the data a modeling technique can account for. For the validation portion, this tool may be misleading. This is because processing is done on both the training and test sets for model comparison. The nature of this processing can directly impact the R^2 value. A different technique to summarize the performance on the test set is described in the subsection 3.4.

All models made have relevant parameters settings included here. Parameters not stated are those that are the default values for the respective libraries used.

Make Naïve Model.

As a baseline, it is useful to compare how modeling techniques compare to a very simple (naïve) estimation. [4] In this case, the root mean square error (RMSE) is evaluated for a prediction of the average specific range of the training set without any consideration of gross weight or altitude.

Linear Regression.

This model was not expected to have the best performance, but allows for comparison with Havko's results. Altitude, gross weight, altitude squared, gross weight squared, altitude times gross weight, and an intercept term were used to train the linear regression model. The average RMSE of the ten folds was recorded. No hyperparameters were considered for this technique.

K-Nearest Neighbors Regression.

There were two hyperparameters considered in for this model. The first was uniform weighting of the k neighbors versus Euclidian distance weighting. The second

hyperparameter was the value for k-neighbors. Exploring the application of this model indicated the best value of k was large. The recorded values of k ranged from 2,500 to 11,000 in increments of 500. Standardized values for gross weight and altitude were used as inputs.

Random Forest Regression.

One hyperparameter was considered for the forest regression modeling technique. The hyperparameter considers the way the value-feature split is generated for each bagged tree. The first option was to split the bagged trees on the best feature-value pair. The alternative was to split on one of the two features at random.

If one model were made with each hyper parameter, a technique with the lowest RMSE may result from chance. To better estimate which may be the Goldilocks hyperparameter, three models for each setting were made. The average of ten fold validation RMSE was recorded. Six models were made. Building many models may result in an increased probability of a model that will overestimate its accuracy. This is referred to as an "information leak." [4, p.97]

Scikit Learn default parameters were used with one exception. The minimum data points permitted in a given split is two. With the large number of samples used in model training, a more reasonable value of 100 was selected.

The criteria for the best split was MSE. No maximum tree depth was used. Max features was the parameter used to differentiate tree approaches. The default number of ten trees was used. Bootstrap samples had a default setting of "True."

Artificial Neural Network.

To estimate the quality of a model made from an ANN, an architecture with a large capacity for the numbers of features was made. This architecture consisted of

four layers. An important note is that the word "layer" may be understood differently in different contexts. For clarity, consider "layer" to have the same definition as the documentation for the software library Keras. [17] Layers are described by software components that have inputs and outputs. Layer is taken to mean a software element that gives output values for a set of input values. The Keras software application of this term is used because it is a commonly used software API and was used to implement the network construction, fitting, validation, and testing in this effort. A dense input layer, two dense hidden layers, and an output layer were used. The width of each hidden layer was eight perceptrons. Eight was estimated to be a relatively large capacity for the model. Input values were standardized. The output values were not standardized. Activation functions within the network were tanh. This function limits the output of each perceptron from negative one to one as shown in Figure 12. The output layer had a linear activation function. Mean squared error was used as a loss function. The Adam optimizer was used. The size of a batch was 128 samples.

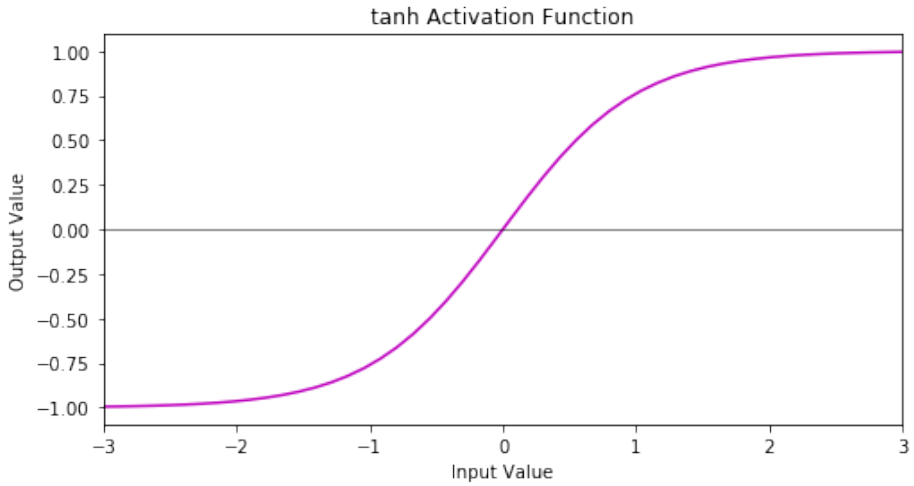

Figure 12. The tanh Activation Function

Dropout was added to this model for regularization. In Keras this is implemented by adding dropout layers. The layers add a masking function to a subsequent layer. The dropout layers randomly select perceptrons outputs and set them to zero. A 50%

dropout rate, indicating half of the perceptrons outputs were dropped for a given mini-batch, was used. Dropout layers act as a mask only during training, not during prediction on a trained model. Weight adjustments for training dropout networks and applying the network to make predictions were handled by Keras. The layers used are enumerated here and illustrated in Figure 13:

1. Dense Input Layer
2. Dense Hidden Layer, Width = 8
3. Dropout Layer
4. Dense Hidden Layer, Width = 8
5. Dropout Layer
6. Dense Output Layer

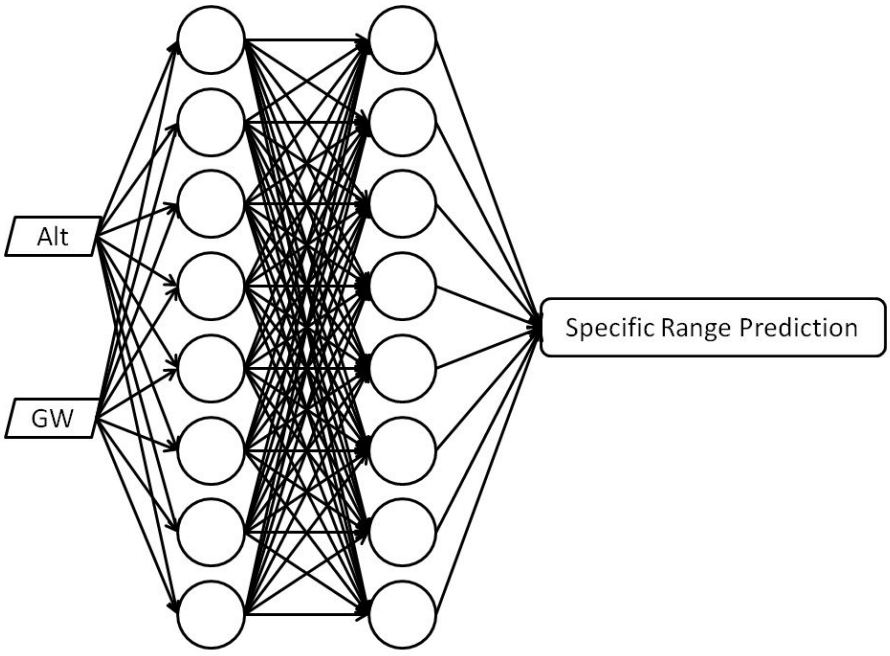

Figure 13. Dense neural network with two inputs, one output, two hidden layers, and a width of eight perceptrons.

The ten models, one for each fold, were trained for 35 epochs on each fold. Performance on the training and validation sets for each epoch were evaluated to estimate the minimum number of epochs the final model should be trained. The number of epochs used must give a high confidence that the network is sufficiently trained. The average of the ten models' training and validation RMSE was recorded. The ability of the large network to overfit the data was assumed.

Model Selection and Evaluation.

Straightforward model selection might consist of a simple decision process.

1. For each modeling technique, select the hyperparameters that result in the best validation accuracy.

2. From the best of each modeling technique, select the one with the best validation accuracy.

3. Train the selected modeling technique and hyperparameters on the entire training set.

4. Evaluate the accuracy of that model on the test set.

For model evaluation, the model accuracy of each technique is similar as shown in Chapter IV, Table 4. It cannot be said with certainty based on the recorded RMSEs that one technique dominates the others in terms of accuracy. RMSE is one estimate of accuracy, and the goal is not to have a small RMSE but to estimate the range of an aircraft for a given fuel load. For this reason, rather than select one model, other considerations were used to down-select. The modeling techniques that remained after down-selecting were trained on all data in the training set. Evaluation using the test set was accomplished. The test set RMSEs were recorded. Visualizations of

performance of each model based on different ranges of parameters were generated to allow thorough evaluation.

It is key to note that MSE and RMSE serve only to compare one set of models to others using the test set. The preprocessing techniques that were applied to both the training and test sets should directly impact these estimates.

An evaluation must give information on how the model is likely to perform in its intended use. A predicted distance traveled must be compared to the actual distance traveled. The segments vary significantly in length. This results in a difficulty in evaluation as performance is expected to vary with time. A way to visualize this performance was developed. For the distance each segment covered, a prediction of the range would be made for the weight of the fuel used in the segment. This would be graphed along the ranges each segment covered on the x axis. The y axis would contain the percentage of error of the estimated range.

To illustrate the accuracy of the model, each model was used to predict the distance covered for each segment in the test set. The models were used to predict a change in air distance for the change in aircraft weight over the segment. The range can be described as the output of the predictive model function $M(a, gw)$ that takes the current altitude a, current gross weight, gw, and returns a distance, D.

$$R_{est} = \sum_{i=0}^{1000} M(a, gw_i) \tag{24}$$

where

$$R_{est} = \sum_{i=0}^{1000} M(a, gw_i) \tag{25}$$

$$gw_0 = GW_{max} \tag{26}$$

where GW_{max} is the maximum gross weight of the segment and

$$gw_i = gw_{i-1} - \Delta gw \tag{27}$$

and

$$\Delta gw = \frac{GW_{max} - GW_{min}}{1000} \tag{28}$$

where GW_{min} is the minimum gross weight of the segment.

The distance covered by a flight was the air distance. This was calculated from the true airspeed. The truth value for the each segment's distance covered to which R was compared was computed from the equation:

$$R_{true} = \sum_{i=1}^{n} \frac{TAS}{60^2} \tag{29}$$

where n is the number of samples in the segment. Recall one sample represents one second of flight. This converts from units of NM per hour to NM.

Visualizations for each model and one visualization with the combined results from models were made. The visualization shows the percentage error in the predictions on the y axis. This was calculated with the formula

$$y = 100\frac{R_{est} - R_{true}}{R_{true}} \tag{30}$$

This visualization conveys the accuracy and relative goodness of each of the models.

Finally, the net error on the test set was summed. The error in predictions for the training set were recorded as well as the dividend of net error and total distance flown.

IV. Analysis of Results

This chapter presents the findings from the procedure described in Chapter III. The outline is similar, starting with data understanding, model validation and selection, and concludes with evaluation.

4.1 Understanding Flight Recorder Data (Data Understanding)

The data had been used for previous research and was already segmented into cruise sections. Each sample accounted for one second of flight. All together the training and test set of data accounted for 1,512 hours of flight time. Other details can be seen in Table 2.

Table 2. Data Details

	Training	**Test**
Total Sorties	543	268
Total Segments Included	1,412	668
Total Samples Included	3,685,190	1,759,590
Segments Excluded	438	188
Samples Excluded	595,019	270,318
Data Samples Excluded (%)	14%	13%

Data Occurrence.

The segment lengths are not uniform. A majority of segments in both the training and test sets are under 1,000 samples in length. Figures 14 and 15 show the distributions of the training and test sets. This is after segments shorter than 30 seconds in length were omitted.

Distribution of data across the parameters of altitude and gross weight is not uniform or random. Samples are more frequent where aircraft tend to fly. Flight planners attempt to maximize specific range which is represented approximately by

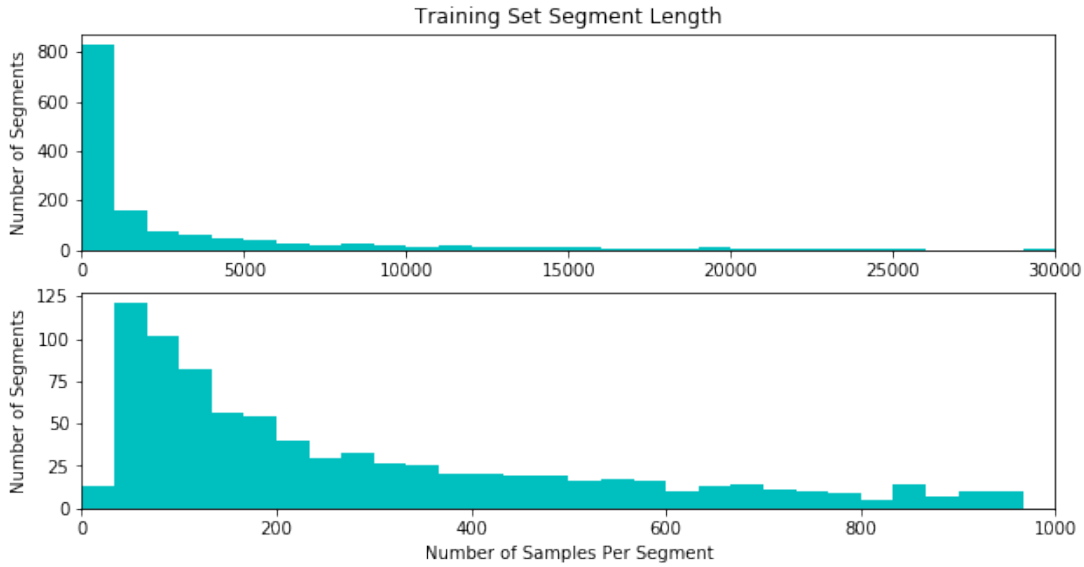

Figure 14. The top histogram shows the number of segments that fit into each range of the training set. The bottom histogram shows how the length of segments is distributed for segments under 1,000 samples in length. The bottom histogram is a detailed view of the first bar of the top histogram.

the green band in Figure 7. Figures 16 and 17 show that a large majority of data exists where flight is thought to be more efficient. The difference in data available between these cruise parameters is very significant. The densest areas contain nearly 190,000 samples. Compare this to areas where there are 200 to 3,200 samples. The stark contrast in the density of samples in these areas can be considered qualitatively as the purple and black areas. Purple would refer to the areas of purple tint and the colors with increasing density.

Aside from a concentration of data at best expected specific range, there are noteworthy areas where there is no significant data. There are several reasons that might explain the absence of data in these areas. Consider the gross weight of the aircraft, which is the sum of three weights:

1. Empty Aircraft Weight

2. Cargo Weight

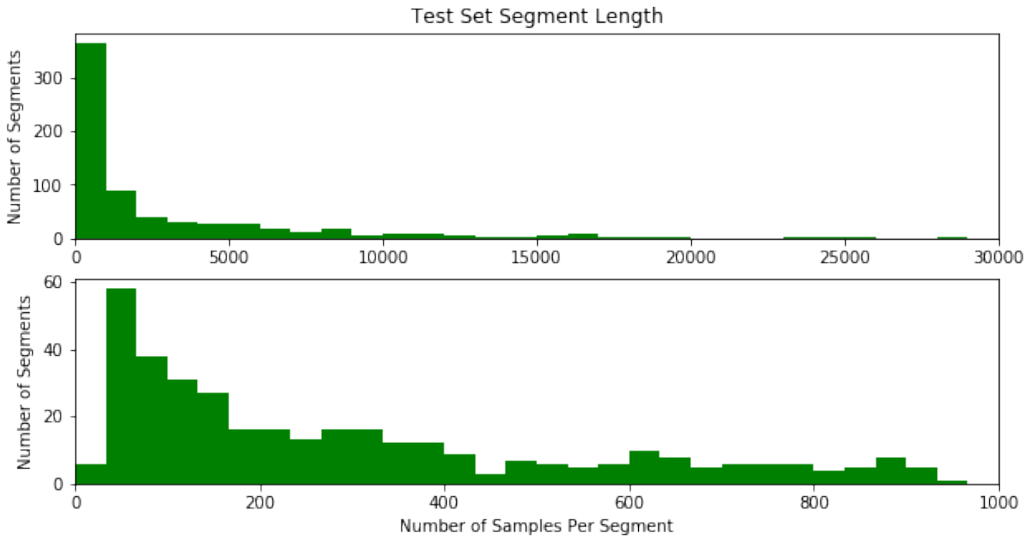

Figure 15. The top histogram shows the number of segments that fit into each range of the test set. The bottom histogram shows how the length of segments is distributed for segments under 1,000 samples in length. The bottom histogram is a detailed view of the first bar of the top histogram.

3. Fuel Weight

Moving things by air is cost intensive. This is true for the overhead cost of completing a single sortie, and the increased cost of a single sortie due to the weight of the cargo. There is a preference not to fly more cargo than necessary. Likewise, mission planners add weight through additional fuel sparingly. This may be why samples that have a large gross weight are rare. Additionally, the aircraft weights decrease over the course of a sortie through fuel usage. Aircraft heavy with a large fuel load for a long trip will tend to have the same altitude and gross weight parameters later in flight as shorter flights with similar cargo loads.

The range where the altitude is high and the aircraft weight is heavy, in the bottom right of Figures 16 and 17, is likely due to a combination of two things. The first is the less desirable fuel efficiency at higher altitudes for a given weight. The other may be the physical limits of the aircraft in terms of flight ceiling for a given gross weight.

Relatively low altitudes (below 25,000 feet) occur less frequently in general. They

occur more often when the gross weight is light (420,000 pounds or less). This corresponds to the top left portion of the figures. This can be explained by the flight requirements in the different phases of flight. When an aircraft takes off and is relatively heavy with fuel, it climbs directly to cruise. When an aircraft gets close to its destination, it must follow regulations and other requirements to prepare for approach. This means that if a sortie has a low altitude cruise portion, it is likely at the end of its flight. The aircraft has burned much if its fuel by this time in the sortie which causes the aircraft to be lighter.

It is also useful to consider how a given flight may be represented by this data. Consider a sortie similar to the example flight in Chapter I. The flight in Figure 7 may be mentally overlayed onto Figures 16 and 17 and tracked in Figure 1. The aircraft begins with 496,000 pounds gross including 160,000 in fuel. It takes off and climbs to its first cruise altitude. The figures show approximately 10,000 pounds of fuel was used in its climb and the first samples of its cruise section occur at 30,000 feet and 486,000 pounds. There are around 50,000 samples that correspond to similar flight parameters. The fuel weight of the aircraft declines to about 110,000 pounds without changing altitude. The aircraft weight is approximately 446,000 pounds. The samples contributed to the graph for this section of flight move up. The example flight has two step climbs in succession, burning slightly more fuel but moving to the right. At its new altitude of about 34,000 feet, the aircraft continues until it is time to descend. It makes a contribution to the data, moving left, at about 25,000 feet, descends again to close to 10,000 feet, and the remainder of the flight's contribution to the sample distribution end by the 10,000 feet data cutoff.

Comparing the two graphs in Figures 16 and 17, it is very clear that there is little training data represented by the bottom left regions of the graphs. These infrequent occurrences are also not uniform. This is a set of features for which any model is

liable to have poor performance.

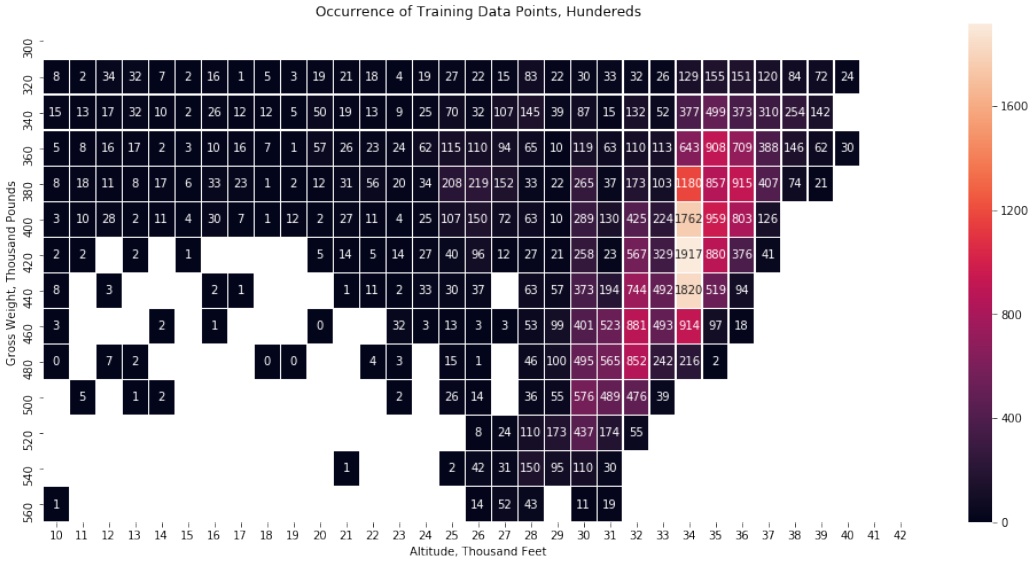

Figure 16. How many data points, in hundreds are closest to a given gross weight and altitude in the training data used. Areas where there were no data points are white. Areas that show 0 are areas where there are more than 30 data samples but less than 100.

4.2 Procedure Results (Data Processing)

Standardizing the data requires the mathematical adjustments to be saved. The mean and standard deviation of the training set is unlikely the be the same as the test set. Standardized values, S, of the training set were calculated from the original values, O, by using the mean, M, and standard deviation σ of the samples for a given parameter in the filtered training set. It was necessary to record these values of M and σ to perform the same adjustments on the test set. These numbers were used to standardize both the training and test sets.

$$S = \frac{(O - M)}{\sigma} \tag{31}$$

The values used are listed in Table 3.

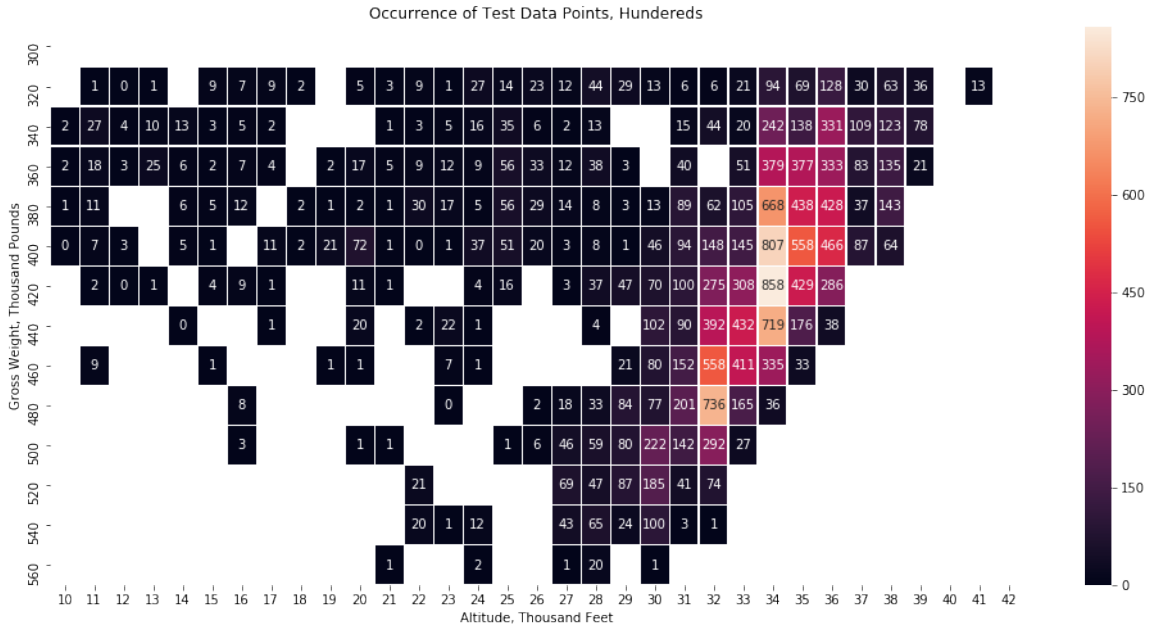

Figure 17. How many data points, in hundreds are closest to a given gross weight and altitude in the test data used. Areas where there were no data points were are white. Areas that show 0 are areas where there are more than 30 data samples but less than 100.

Table 3. Mean and Standard Deviation of Parameters in Training Set

	Mean	σ
Gross Weight	414.5947	51.9384
Altitude	32.4400	4.0164
TASR	28.3737	4.0564

4.3 Calculating Specific Range

Figures 18 and 19 show the average values of the calculated specific range over the training and test set. Where data is sparse, likely from few flights, variance from the expected average increases. Areas that deviate significantly from the color and numeric trend represent areas that are likely to result in poor performance by a model.

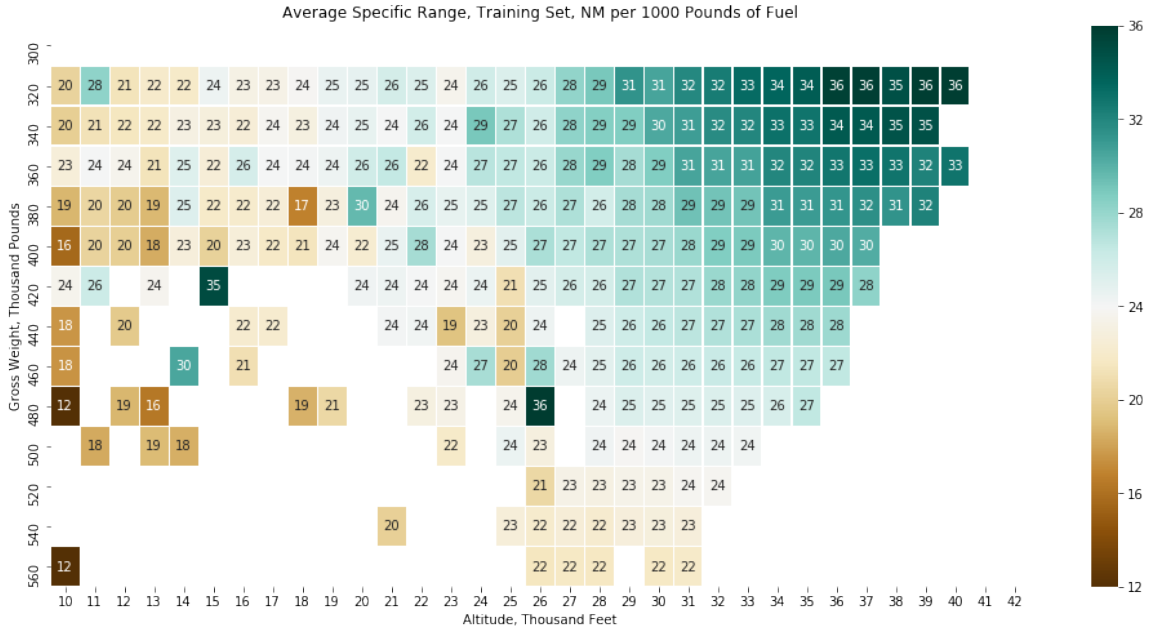

Figure 18. The average specific range for the training data.

4.4 Modeling

The training and test performance of the different modeling techniques can be seen in Table 4. As noted in Chapter III, the modeling techniques were down-selected. K-nearest neighbors was not evaluated on the test set. This is because it showed similar performance on the validation set as other techniques but had relatively high computation time to generate predictions. It is very unlikely any benefit from this technique over the others will outweigh the downside of its computation time.

Table 4. Validation and Test RMSE

	Val RMSE	Test RMSE
Naïve	4.0564	3.7395
Linear Regression	2.6227	1.9737
Forest Regression Random Split	2.5718	2.1436
Forest Regression Best Split	2.6384	NA
KNN Uniform Weight (Best k = 8,500)	2.5963	NA
KNN Distance Weight (Best k = 10,000)	2.6552	NA
ANN	2.6067	2.0976

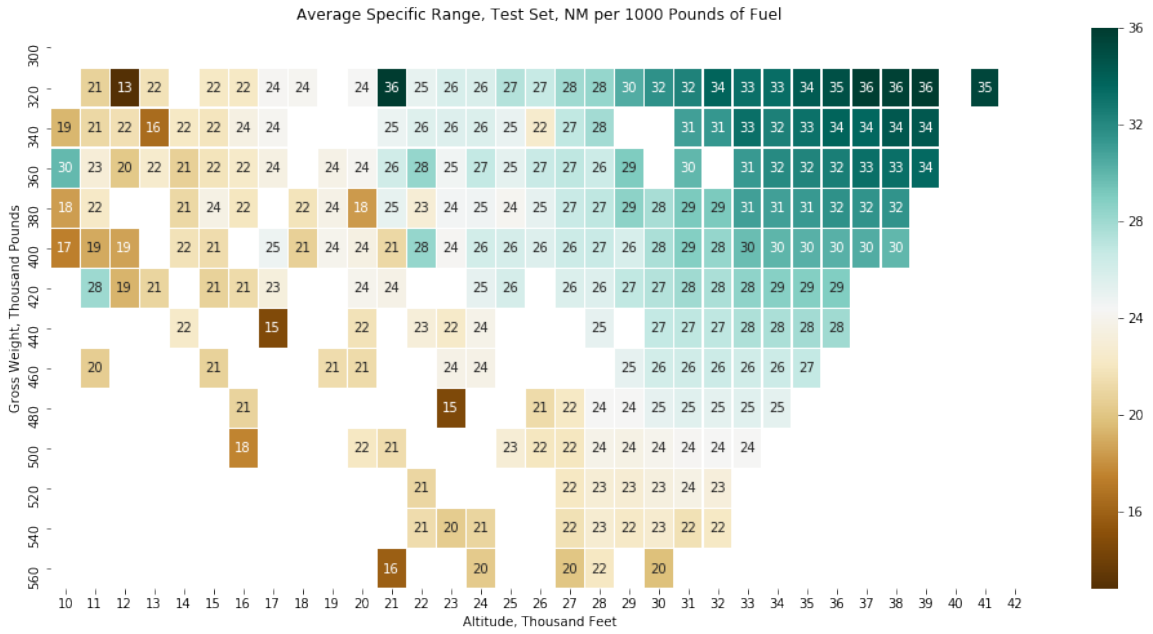

Figure 19. The average specific range for the test data.

K-Nearest Neighbors Regression.

The validation performance for the values of k tested can be seen in Figure 20.

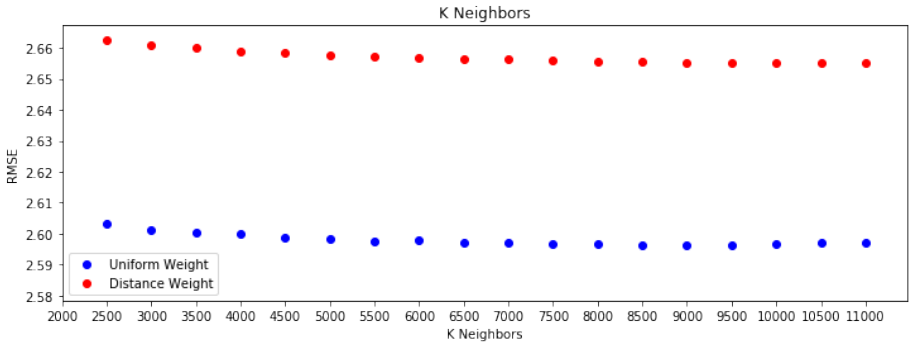

Figure 20. Average validation results of the ten folds. The lowest RMSE is the uniform weighting at k = 8,500.

Linear Regression.

The coefficients for linear regression are listed in table 5. The coefficients are the β values from Equation 5.

Table 5. Coefficients of linear regression model from training set data. These are comparable to the coefficients found by Havko [8]. In the regression accomplished by Havko, the coefficients were found using a random sample of 100 flight segments. Processing described in Chapter III was not done before finding coefficients in Havko's regression.

		Coefficient	Havko
Intercept	β_0	8.9156	5.952
Altitude	β_1	1.232	1.132
Gross Weight	β_3	$1.0643 \cdot 10^{-2}$	$4.7 \cdot 10^{-2}$
Altitude^2	β_4	$1.2766 \cdot 10^{-3}$	$6 \cdot 10^{-3}$
Gross Weight^2	β_5	$2.061 \cdot 10^{-5}$	$-1.05 \cdot 10^{-5}$
Altitude $\cdot$ Gross Weight	β_6	$-2.2274 \cdot 10^{-3}$	$-2.8 \cdot 10^{-3}$

Random Forest Regression.

The validation RMSE for forest regression can be seen in Table 6 and plotted in Figure 21. The modeling technique in which the trees selected a split from a random variable performed better than each of the three models that split on the best feature-value pair.

Table 6. Random Forest Validation RMSE

Random Split	Best Split
2.575	2.641
2.569	2.635
2.572	2.639

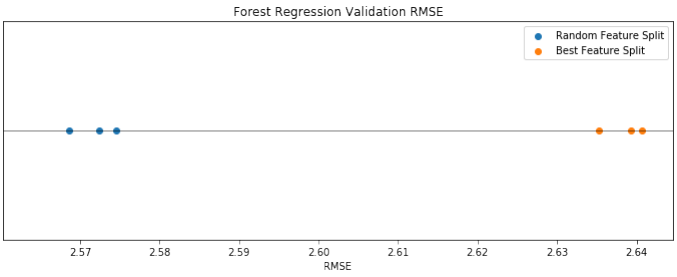

Figure 21. The validation RMSE of six random forests. The models were all similar in performance, but it is very likely a random feature split will be more accurate than a best feature split technique.

Artificial Neural Network.

Figure 22 illustrates the validation RMSE for the 10-fold cross validation. It is clear the network stops improving relatively early. Ten epochs were considered sufficient to train the final model. The figure shows the RMSE for the entire training set as the model was trained for ten epochs on all of the test data and the RMSE for the final model.

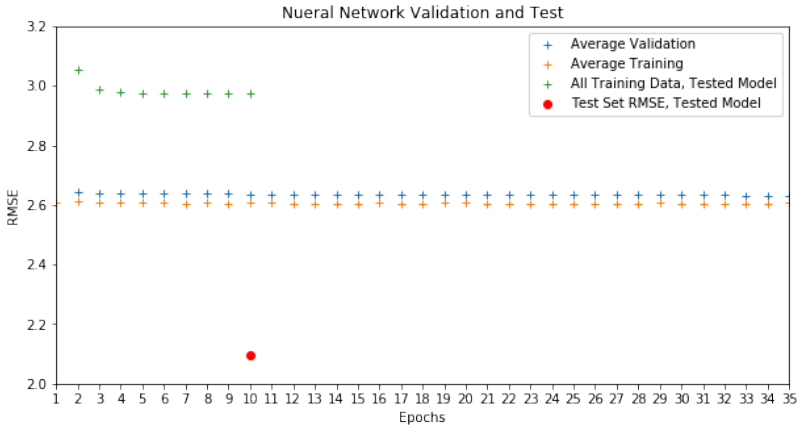

Figure 22. This figure clips the values of the first epoch to focus on the convergence of the RMSE over training iterations. The final model was trained on the entire training set for ten epochs. The resulting RMSE on the test set is shown in red.

Evaluation.

The MSE for each of the models for varying gross weights and altitude can be seen in Figures 23, 24, and 25. To see the difference more clearly, the absolute differences in model accuracy in terms of MSE between the three models is show in Figures 26, 27, and 28. MSE was used instead of RMSE for visual purposes. Distinguishing differences in RMSE requires the use of decimals which would clutter the figure. The differences in MSE values may be differentiated clearly. Note that the ANN and LR differences are much smaller and its color scale is different from related figures.

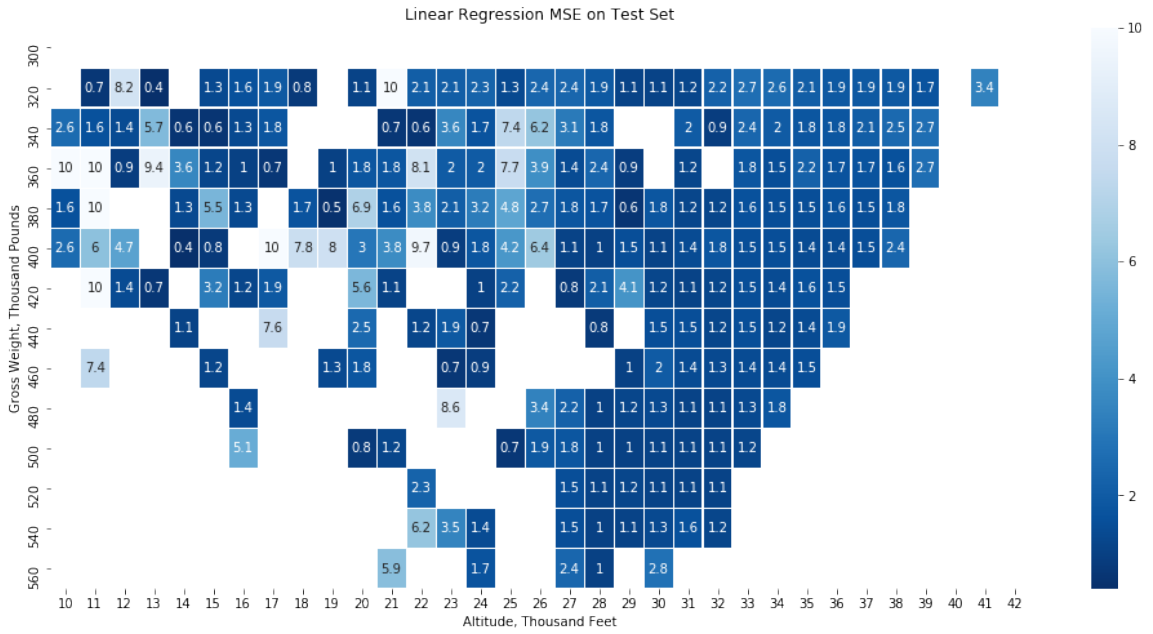

Figure 23. Linear Regression Test MSE

The models had their best and notably similar performance where samples were dense. The models had varying performance for the low altitude segments where data samples were less frequent.

Random forest regression seemed to perform poorly compared to the other models for intervals where there was little training data. This is likely due, in part, to unbalanced data. The trees are made from bagged (sampled with replacement) data. There may have been trees with almost no representative samples where data samples are infrequent. The ANN and linear model tended to make better predictions over this range.

ANN and linear regression result in very similar predictions. Forest regression differs from ANN and Linear Regression. The locations where ANN and linear regression show significant differences are in the low altitude data where the density is sparse. Much of the data in these areas may come from one or two flights, making prediction in this range very sensitive to the variance from aircraft to aircraft. It is likely that the ANN's relative flexibility to accommodate those variations may have

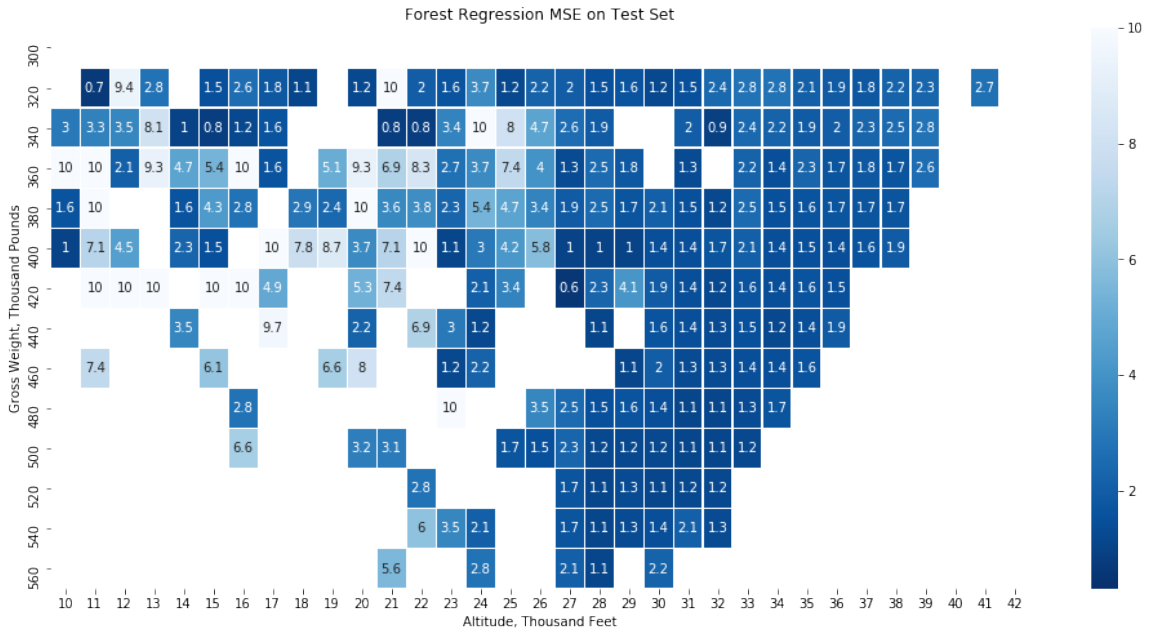

Figure 24. Forest Regression Test MSE

been the cause of its RMSE, which is high compared to the linear regression model.

4.5 Predicting Range

Figures 29 through 31 show the percentage error of range predictions for the fuel burned on the test set. Figure 32 shows the error for all models on the same graph for comparison. The top of each of these graphs shows the range of results in terms of percentage error. The bottom graphs in each figure exclude data points but allow the results to be seen in more detail. The models make similar predictions. Where there are apparent erroneous estimations, the models made similar errors. For example, a single 25 NM prediction was nearly 600% in error and this error was approximately consistent across the models.

Where the predicted segments are short, a higher variance on the graph is expected. This is because variations that are slight make a significant difference compared to the total range. There are two significant trends in these figures:

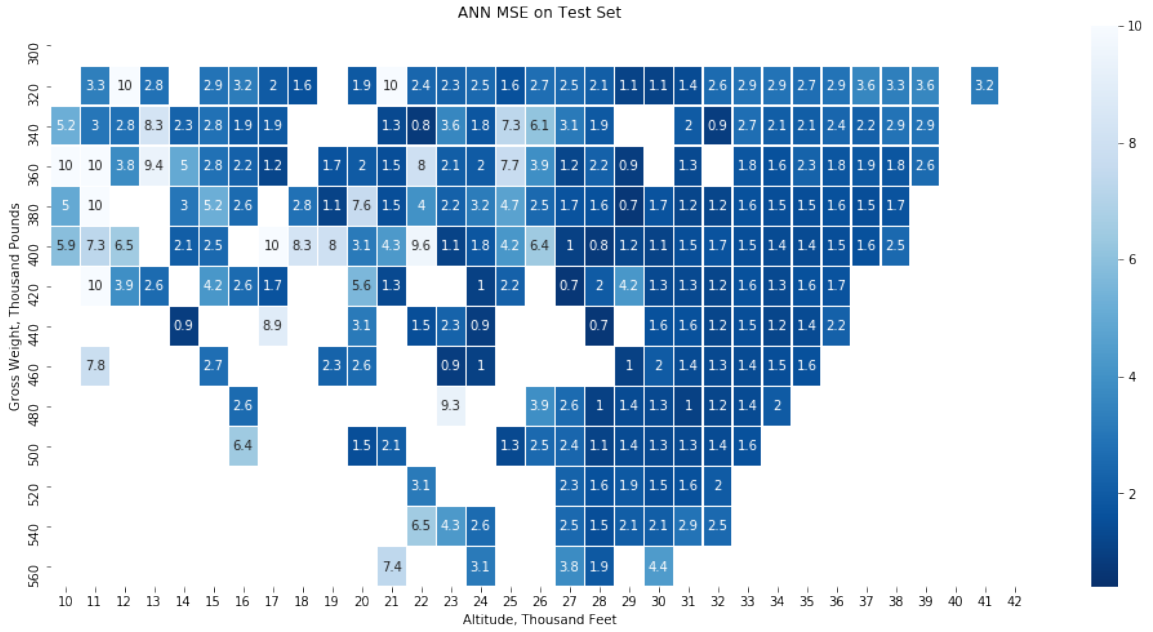

Figure 25. ANN Test MSE

1. When the models err predicting short segments, it is disproportionately positive. That is, errors on short segments tend to overestimate the range that can be traveled for the fuel used.

2. When the models error predicting long segments, it tends to be negative. That is, errors on long segments tend to underestimate the range that can be traveled for the fuel used.

The test set of data had a cumulative range 225,262 nautical miles flown. The weight of the fuel used to fly this was 8.5 million pounds. The cumulative error in this prediction is shown in Table 7. The net estimate of aircraft range for the fuel used was high for all models. Forest Regression resulted in the smallest amount of error.

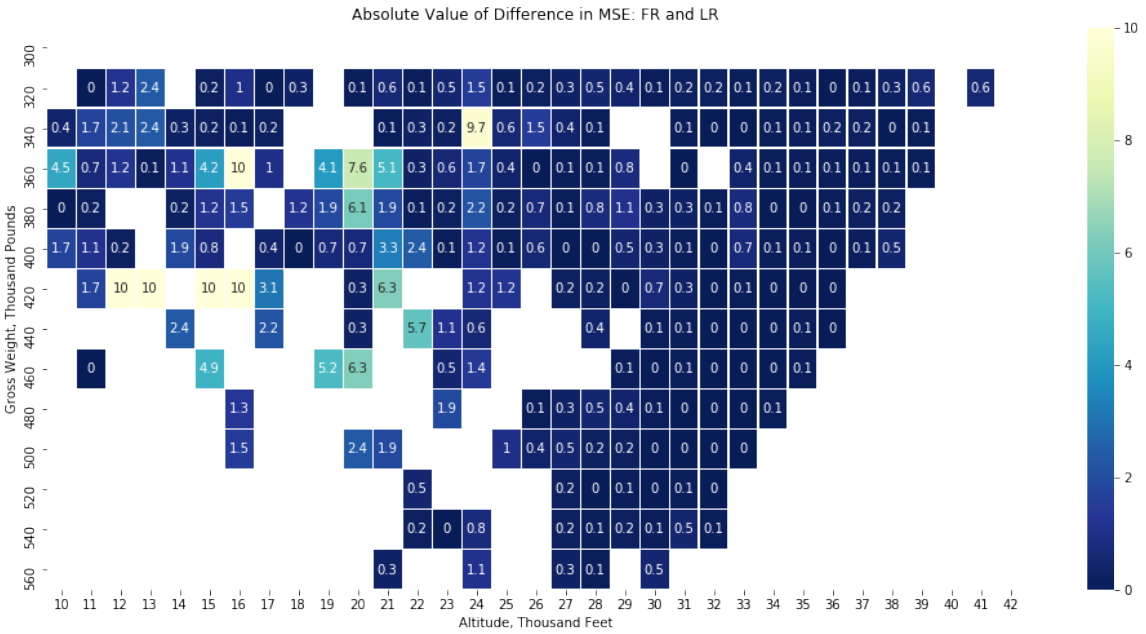

Figure 26. The absolute difference in performance on the test between forest regression regression and linear regression. A value of ten indicates a MSE difference of ten or greater.

Table 7. Cumulative Error

Model	Linear Regression	Forest Regression	ANN
Cumulative Error (nm)	6068	5000	8726
Predicted/Actual (%)	2.69	2.22	3.87

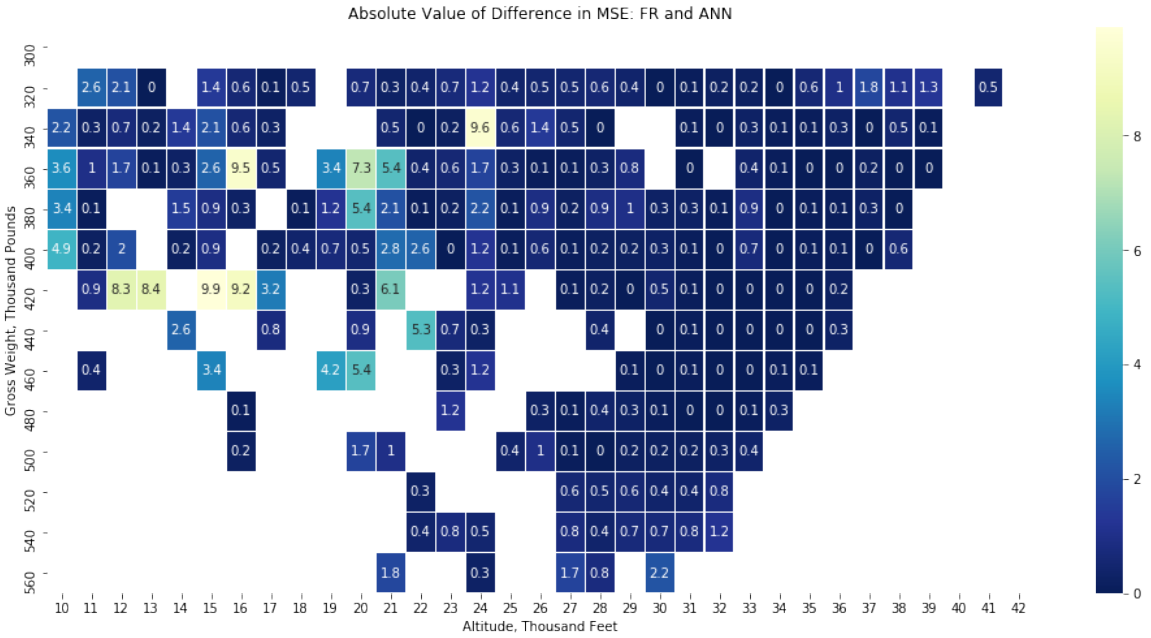

Figure 27. The absolute difference in performance on the test between forest regression regression and ANN. A value of ten indicates a MSE difference of ten or greater.

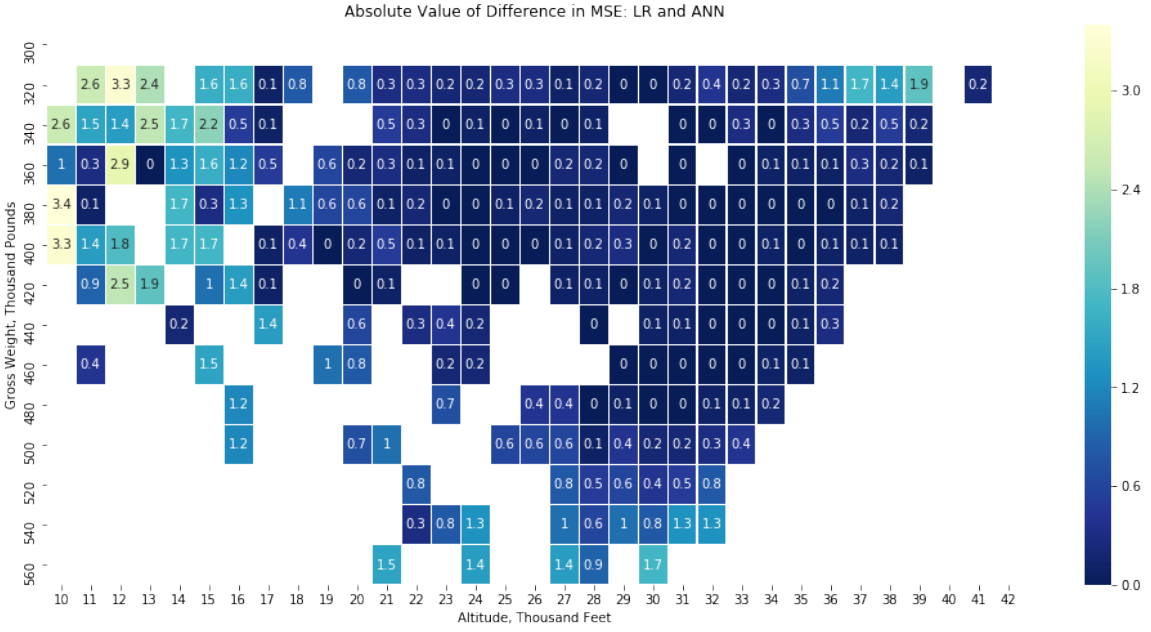

Figure 28. The absolute difference in performance on the test between linear regression and ANN. Note this color scale is different from the differences that include forest regression. The predictions of ANN and linear regression are similar.

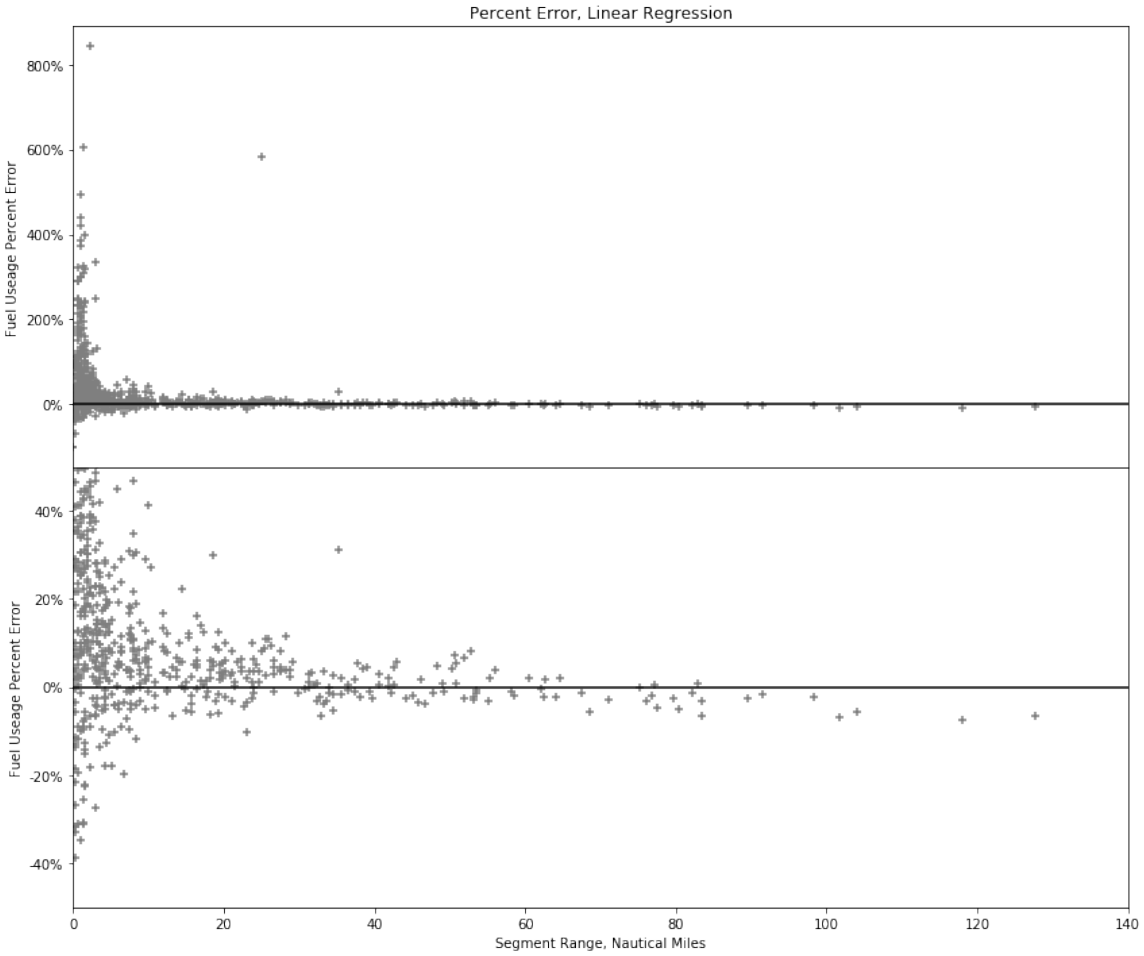

Figure 29. Percent Error, Linear Regression

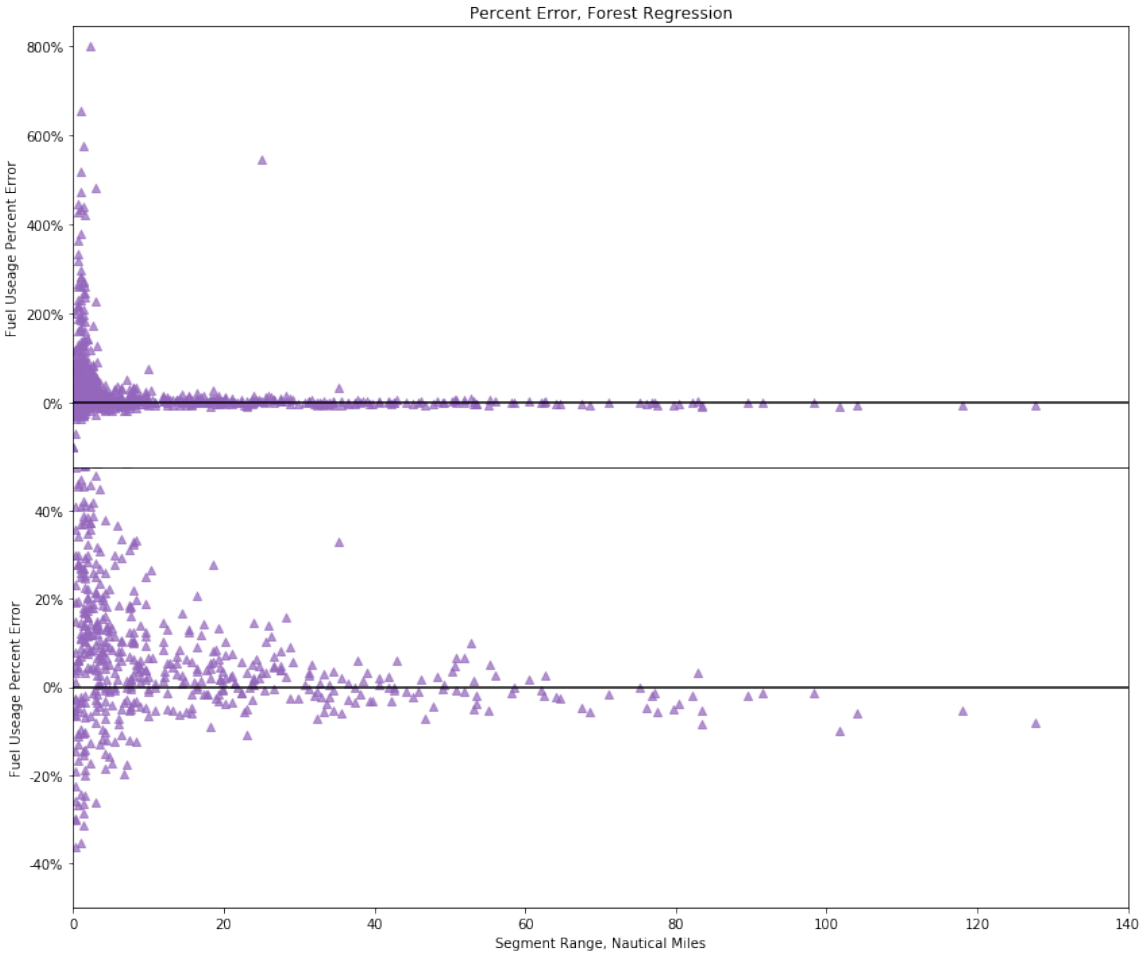

Figure 30. Percent Error, Forest Regression

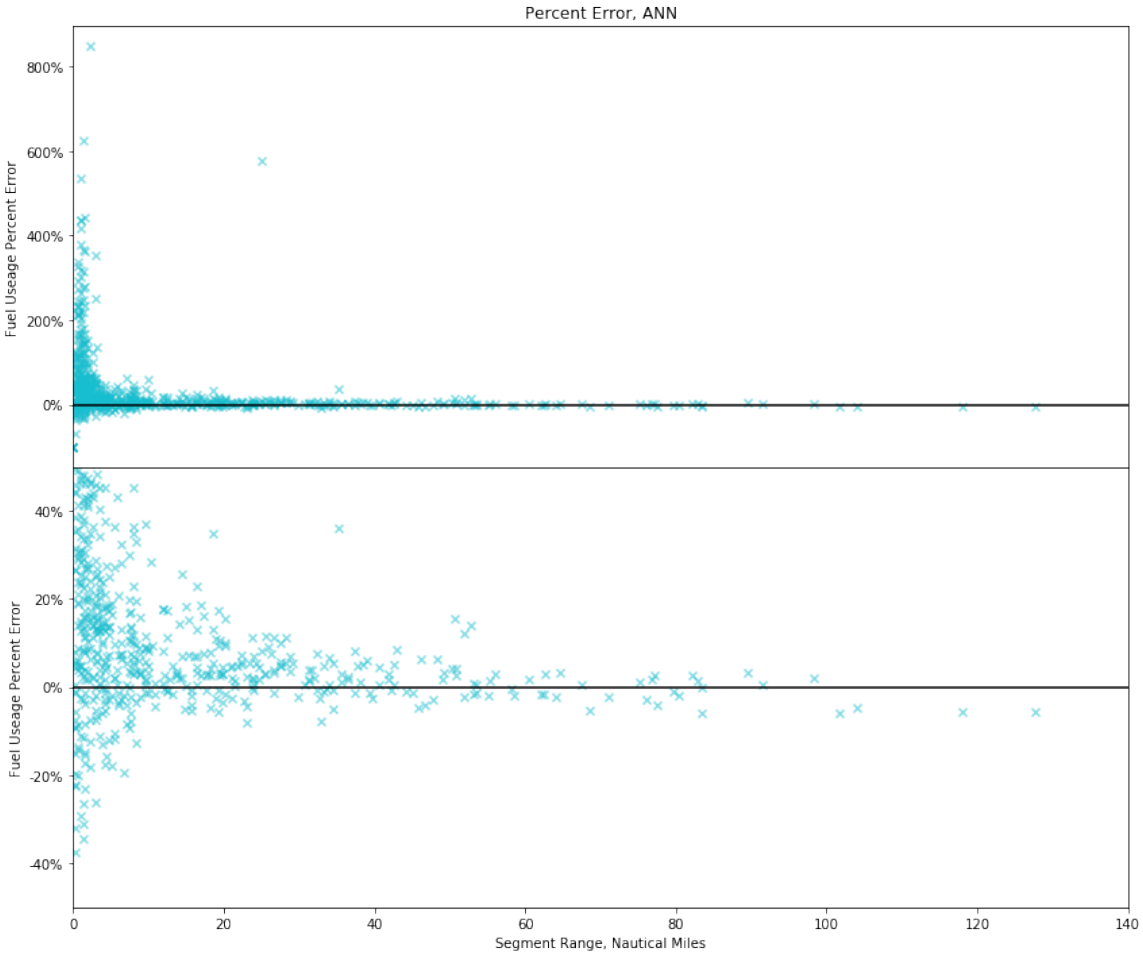

Figure 31. Percent Error, ANN

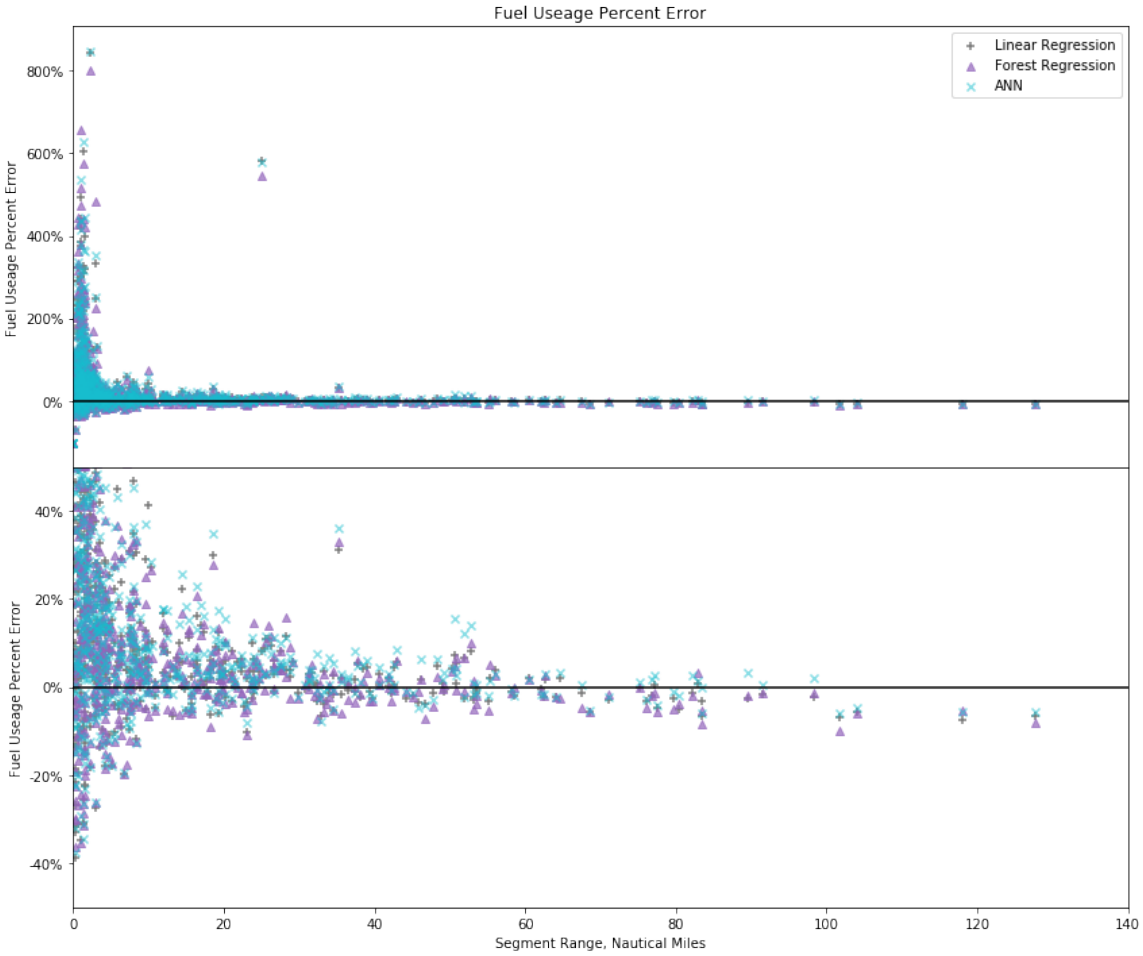

Figure 32. Percent Error, All Models

V. Conclusion and Future Work

5.1 Overview

This chapter summarizes the findings, offers conclusions, and recommends future work.

5.2 Findings

The linear regression technique used by Havko [8] yielded the model with the best general accuracy in terms of RMSE at 1.97 nautical miles per thousand pounds of fuel. ANN and Linear Regression techniques had similar performance. They varied in accuracy in regions where data was sparse. This is likely due to the higher capacity of the ANN. The higher capacity allowed an increased ability of the model to fit erroneous data. In prediction of total range flown on the test set, Forest Regression performed the best with a 2.2% overestimation of the range flown. In terms of total error in range prediction on the test set, all models predicted a greater range than was flown.

Consider the two hypotheses that were stated by Chollet [4, p 111]:

1. The outputs can be predicted given your inputs.

2. The available data is sufficiently informative to learn the relationships between the inputs and outputs.

The performance of the machine learning models on the test set indicate both of these hypotheses are true. It is clear that the outputs can indeed be predicted for the given inputs. For the second hypothesis, the informative nature of the data is not considered to be exhausted. It is expected additional processing techniques, such as those suggested in Section 5.3 will lead to model improvement.

All models performed similarly well in areas where sample density was high. This indicates there is sufficient data in this range for most machine learning algorithms to capture the probability distribution in that region.

The forest regression model had more inaccurate areas where the MSE was high than ANN or linear regression where the data was sparse. The trees in the forest used bagged data sets, that is, sampled with replacement. The relative difference in data density may have resulted in bagged data samples that had few or no samples from these regions. Forest regression is also more likely to have bad estimators in ranges outside existing data.

There is evidence that the forest regression model contains a hyperparameter that was not initially considered. The minimum samples per leaf were adjusted due to the high density of the data. The value was set to two rather than the default value of 100. In Figures 16 and 17, Two hypothesis are suggested for minimum leaf size values between two and 100:

1. The smaller the minimum samples in leaf size, the better the accuracy in ranges where data samples are sparse.

2. The higher the minimum samples in leaf size, the better the model may perform where data samples are dense.

The nature of the final predictions of total range reveal the most about the models' performance. The nature of the error seems to be highly dependant on the length of the segment predicted. This was visualized in Figures 30 through 32. Where a segment's traveled range is short, the models tend to estimate high. As the range traveled increases, the prediction tends to be low. This is the case for all models tested. Figure 32 reveals error trends are similar for each model. The clusters of predictions that seem to be prediction errors on the same segment indicate that for at some errors, the source is the data, not the model.

5.3 Conclusions and Future Work

This research showed nonlinear predictive models may be made to calculate the range an aircraft can travel given a starting gross weight, altitude, and fuel used. Data analysis revealed data could be improved by filtering elements of data reduction equation. Filtering fuel flow with a moving average filter resulted in an improved estimate for specific range. The complexity of reconciling theoretical techniques with data has not been resolved. The harm to model accuracy posed by stochastic system inputs and the resulting transient response has not been entirely mitigated.

It is likely the cumulative error predictions may cancel over the course of a sortie. If this is the case, these models show promise in their ability to take mission planning parameters and predict cruise fuel consumption.

It is expected that these are the two causes of most of the error illustrated in the figures. How to proceed with the effort of accurately predicting range for cruise sections is evaluated in terms of the data mining process.

One phase of the CRISP-DM [3] process is Evaluation where the next steps of the process determined. The results of the project are evaluated to see if it is ready for deployment. These models are not ready for deployment. Significant improvements can be made in this model by revisiting the other phases of the data mining process.

Business Understanding.

Future work must be tied in with the end goal. Planners must make accurate predictions and less fuel must be loaded into a C-17. Organization of the effort is key to estimating the most valuable steps toward fuel conservation.

A tentative structure for realizing cost savings is presented in Figure 33. The figure shows tentative relationships in a process that may frame research and modeling efforts while capitalizing on fuel savings soon after the system is constructed.

Additional cost savings will be realized over time with improvements to the predictive models and systematically eliminating or accounting for sources of variance in predictions.

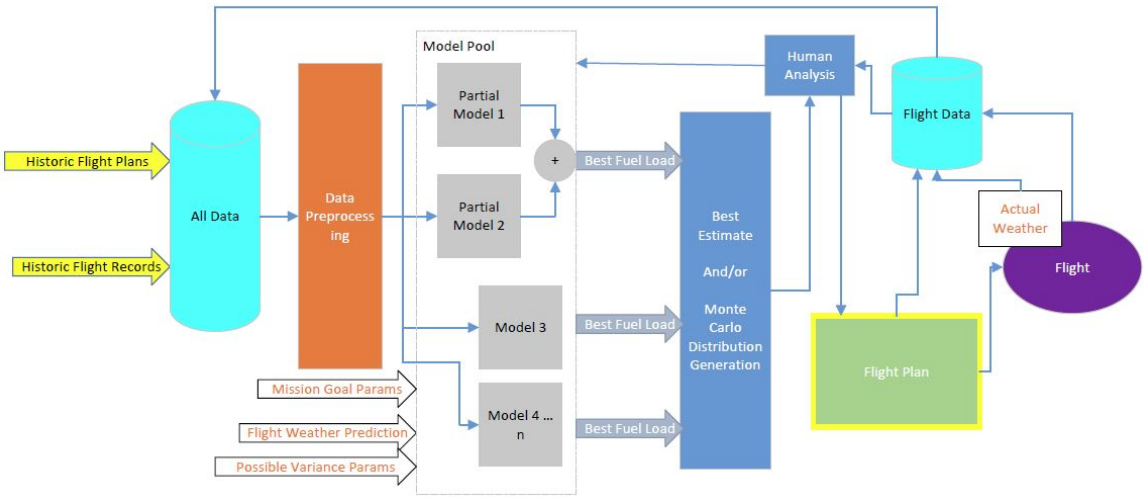

Figure 33. An initial view of structuring future work to realize reduction in fuel use through the application of predictive models. The elements will be described from left to right. "All Data" initially begins with data that has been recovered from flight recorders and other sources. This data will be the basis for data-centric models. Data preprocessing accounts for the means which data is standardized and made available to the model pool. The model pool are the specific modeling techniques that make estimations from mission parameters. The result of the model pool will be several varying estimates that result in best prediction and confidence intervals. Humans analyze both the predictions and result, changing the models and flight plans for details which computational models cannot account. The conclusion of each flight should result in data being gathered and analyzed for differences. Data collection should continue, increasing the data pool.

The flow of information generally goes from left to right. The "All Data" and "Data Processing" efforts are significant in scope. More data from more sources must be collected. A consistent standard for storing this data must be made. Data from varying domains must be integrated in a usable way. This data will initially consist of flight records and other data similar to what was available to this effort. The data that is required for precise and confident predictions must begin to be gathered. This data will consist of the predictive information before a sortie and the resulting "truth" data. At a minimum, before (prediction) and after (truth) data must be collected for

each sortie flown. Minimum appropriate information will include:

1. Flight Plan
2. Weather Prediction
3. Flight Record
4. Actual Weather Measurements

Once such data is collected, estimates may be made of best case and likely fuel savings based on what is physically possible. There is a physical minimum amount of fuel that can be estimated to complete each sortie given hindsight. Human consideration may be all that is needed for initial cost savings. Bringing data together may reveal "low hanging fruit" easily picked by adjustments in flight planning policy. Additional savings would be realized through increasingly accurate predictive models.

Before a model can be relied on for predictions, it must be demonstrated that fuel consumption can be accurately predicted in the planning phase. Two sources of information exist and should be used. These sources are the flight plans themselves and the flight recorder data. Acquisition of paired data (the flight plan and the flight record) will be suitable to both build and evaluate a model's predictive power. Continued collection of paired data must be used to track and update models. Policy or procedure changes are only two possible causes that may result in a decrease in predictive accuracy from flight plans to real flights. Changes in predictive accuracy must be adapted into models and flight planning policy to ensure prolonged fuel savings.

The model pool may consist of models that are based on physics, data, hybrids of both, or whatever can be conceptualized by human ingenuity and supported by reason. Models may make predictions of a segment of a flight or work in tandem with

other predictive models. This research effort considered air distance to eliminate the effects of weather and air movement instead of the ground distance covered by the aircraft. In the figure, models 1 and 2 show how models may be combined to account for different sources of variance to make a prediction. Model 1 may estimate air range and Model 2 may estimate how the air moves relative to the surface of the earth given weather estimates. The estimated fuel usage for a flight plan combines the two to make a prediction.

There are many approaches that may be researched to develop a model in the model pool. For example, a model may be made for a common route that is taken between two airports. If there is a weekly C-17 flight from airport A to airport B, an accurate model of the fuel usage for future flights from A to B may be made from data limited to sorties from A to B. Another model may consider tail-number specific estimates of fuel efficiency and take advantage of historic maintenance data. Recent engine overhauls may result in a confident decrease in the required fuel load compared to the baseline for sister aircraft. The collection of models may come from a mix of research disciplines from data mining, to meteorological, to aeronautical, or something else.

The model pool will result in predictions from a variety of sources of information. The pool is basically an ensemble of predictive models. The pool may grow or change as more knowledge is acquired. Significant differences in the output of one model compared to another must be accounted for by humans. How the predictive information is aggregated to a person depends on the models in the pool and the humans making the decision. This selection effort may be simulations from composite models that result in a variety of possible flight strategies and fuel loads. This can be achieved through algorithms that can help identify the best or most likely result from a variety of predictions. Humans will finalize the flight plan. After the flight,

the prediction must be evaluated against reality. There may be variables from a given flight that are unaccounted for in the models. This "feedback loop" is essential in the fuel saving system. The goal is to load as little fuel as possible. The confidence that an aircraft concludes a sortie in a given range of fuel is a necessity. Only humans can identify and research the root causes of deviations. Deviations will occur and must be accounted for in future predictions.

The bottom right of Figure 33 shows that the Flight Plan, Flight Data, a Flight and the recorded weather experienced should all be collected. After it is analyzed for deviations and insight, it should be added to the data set for use in data-driven models.

Data Understanding, Data Preparation, and Modeling.

Work that is more technical is represented in the middle phases of the CRISP-DM [3] process. The technical effort should be focused on understanding how to predict fuel consumption over the course of a flight. Some of this can be accomplished using nothing more than flight recorder data. The use of flight recorder data may refine how theory fits to historical data. If plans are not compared with flight records, there will always be uncertainty that is not accounted for in a predictive model. Flight plan data must be collected with the resulting data from the flight recorder.

This research only considered cruise flight. To supplement and improve the predictive accuracy of the techniques explored, cruise flight data may be further segmented. It may be separated into sections according to confidence in predictive ability. As noted there were ranges of gross weight and altitude that had a tremendous amount of data. Others had little. Different modeling techniques are better suited where data is sparse. In the range where data is abundant, additional features may be incorporated into models. Exploring modeling improvements can be done by adjusting the

following variables:

1. Into how many ranges the data may be segmented and the process for segmentation.

2. Which techniques may be applied to each of the segments.

3. Which additional features to incorporate into the prediction, if any.

For example, a rough hypothesis that may be tested is this: A linear regression model trained on all data will be the best of several techniques when data is sparse. A forest regression model with an angle of attack feature will perform best when trained and used where data samples are frequent. If such a hypothesis were shown to be true, it may be that a linear regression model would be trained on all data but only used as the primary estimator over some of the ranges that contributed to the training data. A different model may be applied exclusively to the range where data samples are frequent.

Prioritizing efforts may be considered in terms of potential cost savings. Not all improvements in accuracy result in equal cost savings. The range with the most data points in Figure 16 has 191,700 samples. Compare this with a square that has 2,500 samples. The sample sizes in the regions vary because the frequency at which a given C-17 can be found in cruise over that range is far more common. Additional predictive accuracy in either range may result in fuel and cost savings. The resulting cost savings for an increase in accuracy is relative to the frequency of aircraft travel in that range.

For example, consider an improvement in prediction that results in a model that increases the accuracy in the range with 2,500 samples that results in 1,000 pounds of fuel decreased per 1,000 seconds of flight for that range. Now consider increased accuracy in the most data-dense range. A decrease of 100 pounds per 1000 seconds

of flight in this range. The savings may be shown to relate to one another:

$$\frac{191,700}{2500} \approx 77 \tag{32}$$

Taking the resulting ratio and applying the frequency to the amount of fuel saved results in

$$77 \cdot 100 = 7,700 > 1,000 \tag{33}$$

In this example, it would be more beneficial to increase the accuracy of the model when predicting common flight parameters slightly. Much more fuel may inevitably be saved with small gains in accuracy over ranges where aircraft frequently fly. Causes of variance in this range may help identify means to improve this prediction, even if only slightly.

Increased accuracy is dependent on statistically significant amounts of data. A precise boundary is something to be defined in future work, but here consider the black and purple distinction used to articulate the relative difference in Chapter IV Figures 16 and 17. There is a significant ability to dissect and improve predictions in the purple range. Models with lower confidence can be made and tailored to make the best use of the data available in the black region. In either case, more data could prove to be profitable.

Measurement Improvement.

Models are only as good as the data from which they are made. Systematic uncertainty analysis that considers sources of error from sensor measurements and analysis of error propagation may yield ways to improve the data itself. Alternately, models may be created that predict directly on measured parameters and minimize the use of data reduction equations. A model built from sensor inputs themselves

are likely to propagate less error into a final model. For example, in this effort true airspeed, a data reduction equation, was used to calculate specific range through another data reduction equation. Perhaps a better estimate could be made using indicated airspeed.

There may be ways to use data the sensors collect over time to create improved estimates. For example, finding a more exact way to estimate the gross weight of the aircraft from fuel quantity may result in more consistent samples. This could be done by using an appropriate way to reconcile the fuel approximations in Figure 4.

Data Inclusion Standards.

Some of the samples clearly represent atypical situations. In this effort, atypical data suppression was done through filtering. There may be a rule for data exclusion that would result in a model with more accurate predictions.

Time Domain Analysis.

There may be many ways to use the time domain relationships with fuel flow and other recorded flight parameters that could lead to better understanding of how fuel is consumed. A model that estimates specific range that considers additional features such as the angle of attack, trim settings, autopilot settings, and throttle settings may help identify parameters that impact fuel efficiency. Though many of these parameters are not known at mission planning, Monte Carlo techniques could be used to get both predictions and a range of predictions.

An example of this is to relate airspeed to fuel flow. The instantaneous airspeed of an aircraft is not a result of the fuel flow in that instant. The airspeed is the result the cumulative fuel flow that occurred in the past and adjusted by the air friction. A machine learning technique may be used to estimate a relationship between these

and be the groundwork for building a more accurate predictive model.

Tail Number Estimation.

The fuel efficiency varies from one aircraft to another. Additionally, each aircraft's efficiency varies with time. Tail number is a key feature to include in future modeling efforts that promises to show significant gains in model accuracy.

Non Cruise Sections of Flight.

It was noted that specific range is an efficiency measure that pertains to the cruise section of flight. Making predictive models of each of the remaining flight segments is likely to yield significant gains in properly predicting fuel consumption over the course of a sortie.

Apply Techniques to Other Airframes.

What works for analyzing the C-17 fleet should have a great deal in common with what may work for other cargo aircraft. Work done here and future work could be replicated with relatively low effort for the C-130 Hercules fleet and the C-5 Galaxy fleet. The added expense of tailoring fuel savings efforts to other fleets should be low relative to the savings in fuel consumption and enhancements to operational energy.

Organizational Culture.

The technical aspect of fuel consumption prediction is only one part of this effort. Perfect predictions of fuel consumption for a sortie will not result in fuel savings if mission planners have no incentive or desire to decrease fuel load based on those predictions. "Fuel efficiency should be incorporated into leadership communications to employees... A committee should be established ... to discuss strategic fuel efficiency

opportunities." [26, p. 20] How modeling efforts fit into strategic planning cannot be overlooked. The strategic and cultural efforts are liable to be as difficult or more difficult than the technical effort.

5.4 Final Words

This effort was intended to perform model selection and evaluation on cruise sections of flight for C-17 Globemasters using flight recorder data. Benefits and detriments of techniques for the given data were compared. More must be done in terms of processing data and quantifying results before any models may be deployed. A recommended path forward that includes comparing the planning data with the resulting flight data was proposed.

Bibliography

1. C. Poland, "How the Air Force got Smart About its Aviation Fuel use in 2018," Online: Air Force Operational Energy, url: https://www.safie.hq.af.mil/News/Article-Display/Article/1711024/how-the-air-force-got-smarter-about-its-aviation-fuel-use-in-2018, Dec 2018, [Jan. 13,2019].

2. *Technical Order 1C-17A-1-1*, Department of Defense, 2013.

3. C. Pete, C. Julian, K. Randy, K. Thomas, R. Thomas, S. Colin, and R. Wirth, "CRISP-DM 1.0," *CRISP-DM Consortium*, 2000.

4. F. Chollet, *Deep Learning in Python*. Manning, 2018.

5. D. Peckham, "Range Performance in Cruising Flight," Royal Aircraft Establishment, Tech. Rep., 1974.

6. E. W. Kamen and B. S. Heck, *Fundamentals of Signals and Systems: Using the Web and MATLAB*. Prentice Hall, 2000.

7. R. C. Dorf and R. H. Bishop, *Modern control systems*. Pearson, 2011.

8. A. C. Havko, "Military Flight Operations Quality Assurance (MFOQA) Derived Fuel Modeling for the C-17," Graduate Research Paper, Air Force Institute of Technology, 2018.

9. M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.

10. A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. " O'Reilly Media, Inc.", 2017.

11. I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT Press Cambridge, 2016.

12. J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer, 2001.

13. H. W. Coleman and W. G. Steele, *Experimentation, Validation, and Uncertainty Analysis for Engineers*. John Wiley & Sons, 2018.

14. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

16. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

17. F. Chollet, "Keras," 2015, Internet: http://www.keras.io 2015 [Jan. 10,2019].

18. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a Simple way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

19. E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open Source Scientific Tools for Python," 2001, Internet: http://www.scipy.org 2001 [Jan. 10,2019].

20. J. Hunter, "Matplotlib History," 2008, Internet: https://matplotlib.org/users/history.html [Jan. 13,2019].

21. M. Waskom, "Seaborn: Statistical Data Visualization," 2012, Internet: https://seaborn.pydata.org [Jan. 13,2019].

22. T. Baklacioglu, "Modeling the Fuel Flow-Rate of Transport Aircraft During Flight Phases Using Genetic Algorithm-Optimized Neural Networks," *Aerospace Science and Technology*, vol. 49, pp. 52–62, 2016.

23. J. Chen, Y.-p. Zhang, and J. Hu, "Analysis of Flight Fuel Consumption Based on Nonlinear Regression," no. Cst, pp. 670–683, 2017.

24. J. Chen, Y.-p. Zhang, and L. Li, "Fuel Consumption Estimation of Cruise Phase Based on Fuzzy Control," no. Cst, pp. 684–692, 2017.

25. M. Mariotti, "C-5M Fuel Efficiency Through MFOQA Data Analysis," Thesis, Air Force Institute of Technology, 2015.

26. A. D. Reiman, "Enterprise Analysis of Strategic Airlift to Obtain Competitive Advantage through Fuel Efficiency," Ph.D. dissertation, 2014.