STATISTICALLY MODELING FUEL CONSUMPTION WITH HETEROSCEDASTIC DATA

THESIS

L. Elaine Dazzio, GS-12, DAF

AFIT-ENG-MS-17-J-075

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

AFIT-ENG-MS-17-J-075

STATISTICALLY MODELING FUEL CONSUMPTION WITH HETEROSCEDASTIC 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 Science

L. Elaine Dazzio, MSCIS

GS-12, DAF

June 2017

AFIT-ENG-MS-17-J-075

STATISTICALLY MODELING FUEL CONSUMPTION WITH HETEROSCEDASTIC DATA

L. Elaine Dazzio, MSCIS

GS-12, DAF

Committee Membership:

Scott R. Graham, PhD
Chair

Capt Joshuah A. Hess, PhD
Member

Lt Col John M. Pecarina, PhD
Member

AFIT-ENG-MS-17-J-075

Abstract

Aircraft operate in unpredictable environmental conditions. As a result, autopilot design is difficult, as optimal responses cannot be anticipated for all conditions. Consequently, the autopilot might overcorrect for conditions, using more fuel than necessary. By analyzing performance data on a subject aircraft, the relationships between environmental condition variables and fuel consumption using linear regression models have been characterized. These relationships are accurate, even though the data is non-normal and heteroscedastic.

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Dr. Scott Graham, for his guidance and support throughout the course of this thesis effort. The insight and experience was greatly appreciated.

L. Elaine Dazzio

Table of Contents

Page

List of Figures

STATISTICALLY MODELING FUEL CONSUMPTION WITH HETEROSCEDASTIC DATA

I. Introduction

1.1 Background

Knowledge of statistical methods alone is not enough to analyze data. The field of data science includes statistical methods, knowledge of modeling and simulation, programming, and technical writing and presenting. Also needed is familiarity with the data and knowledge within a subject matter area. Though previous efforts have modeled fuel consumption in an aircraft, none have explicitly quantified the relationship between density altitude, turbulence, and fuel consumption.

1.2 Motivation

The environment in which aircraft operate is not predictable. This fact makes autopilot design difficult, as optimal responses cannot be planned for all possible conditions. Only by examining operational data can the actual performance of an aircraft be assessed. With the knowledge gained by examining data, designers can better determine if any re-engineering efforts to improve performance are warranted.

1.3 Research Goals

The goals of this research are to develop accurate statistical models with the desired variables. These variables include indicators for turbulence, density altitude, and

fuel consumption. These models will allow the relationship between predictor and response variables to be studied for the operational aircraft.

1.4 Approach

The data provided from the subject aircraft will first be examined, then cleaned to place it into a format suitable for analysis. Correlations between variables will be calculated, and the variables characterized to assess performance in linear regression models. Once models are developed, their performance is assessed by tests for heteroscedasticity and residual error.

1.5 Assumptions

This research assumes the relationship between predictor and response variables must be defined in an explicit linear fashion. Therefore, non-linear statistical methods will not be used, in order for the relationships to be clearly understood.

1.6 Contributions

This thesis contributes to the body of work in data science related areas within the Department of Defense. Specific contributions include characterization of the data provided for the subject aircraft and development of linear models using the desired variables. As Department of Defense budgets fluctuate, data analysis becomes a necessary activity to gain knowledge from existing data.

1.7 Thesis Overview

This thesis is organized into six chapters. Chapter Two offers background and pertinent related work relating to analyzing data and estimating fuel consumption.

Chapter Three details the methods and procedures used to clean data from the subject aircraft. Chapter Four gives the methods used to analyze the clean data. Chapter Five provides a summary of the results and analysis of the data. Chapter Six concludes the study and suggests recommendations for future work.

II. Background and Literature Review

2.1 Chapter Overview

This chapter provides necessary background relating to the research effort of statistically modeling fuel consumption from heteroscedastic data.

2.2 Background and Related Work

2.2.1 Heteroscedasticity

In linear regression, the values of one or more predictor variables predict the value of a corresponding response variable. However, the prediction does not occur without some error. The assumption is made that the residual error, or the difference between the actual versus predicted values of a response variable, remains constant. In practice, the residual error does not remain constant, and varies for certain values of the response variable. This uneven variance in the residual error is called heteroscedasticity. The presence of heteroscedasticity presents a problem, as extreme residual error values can either inflate or lower the Mean Squared Error (MSE) of a linear regression model. Therefore, when heteroscedasticity is present, the MSE cannot be relied upon to accurately reflect the average error for predicted values.

2.2.2 Data Science

Data science is not a synonym for statistics. Rather, the discipline encompasses statistics as a tool [1], [2]. It also includes product knowledge specific to the area of research, knowledge of programming, modeling and simulation, machine learning, and technical writing and presentation [3]. Data science is the interdisciplinary subject of

obtaining knowledge from data and transforming the knowledge into usable products for the end user.

Acquiring knowledge from data starts with the subject of data cleaning. Data cleaning involves all the necessary steps to prepare data for analysis. Of these steps, ensuring data is tidy is the most important, as it ensures the meaning of the data is supported by its structure [3], [4]. With tidy data, each observation is a separate row, each variable has a separate column, and each table originates from a separate source.

2.2.3 Data Cleaning

The time consuming nature of data cleaning [4] has led to the development of several automated data cleaning tools. The LLUNATIC data cleaning framework by Geerts et al [5], possesses a uniform framework to balance multiple constraints and strategies to clean data. Instead of generating all possible solutions for a given cleaning scenario, which can be computationally expensive, the framework calculates minimal solutions for data repair. Cleaning data in a batch ensures all data is cleaned with the same set of constraints and rules. Alternatively, data can be cleaned as it arrives at a database in a stream, as with the Bleach cleaning system [6]. Data is checked first to see if it violates any rules. If any violations are found, they are then repaired. A dynamic rule based system to detect and repair errors ensures optimal flexibility. Though this system cleans data more quickly, already processed data cannot be cleaned again. The ActiveClean system [7] was developed to support statistical modeling. This system simultaneously cleans data and trains linear regression models. Models are first trained on dirty data, then the data is selectively cleaned to improve errors detected in the previous rounds of model training.

Once data is tidy, the quality of the information can be addressed [8]. More precisely, we are concerned with missing or extreme values that affect accurate data analysis. Missing values can be addressed by broad replacement with a single value or replacement by simulation and modeling of the suspected true value of a variable. Extreme values are harder to fix, as they likely represent valid observations in the data. If extreme values are few in number, the choice can be made to simply discard the observations in which the values reside. Alternatively, the data set can be transformed by mathematical functions such as taking the mean and standard deviation of the values in variable columns.

Freire et al., argue that data cleaning should be an integral part of exploratory data analysis [9]. Only as data characteristics are found through exploratory analysis, can accurate constraints be formulated and applied. They point to the fact that certain characteristics can either be dirt or valid features. Human intervention is needed to decide between the two categories. Freire's method is similar in concept to the ActiveClean system, though human intervention replaces the rule-based constraints.

2.2.4 Applicability

The data science workflow of cleaning, analyzing, then creating data products can be applied to a multitude of subjects. For example, medical data science examines relationships in diverse subjects such as Theory of Mind (ToM) as it applies to Alzheimer's disease [10], feeding of the gut microbiome in infants [11], and identifying risk factors for out of hospital cardiac arrest [12], among other areas within the discipline.

Aeronautical engineering is another subject that continues to benefit from data science techniques [13], especially in the area of estimating fuel consumption. Testing

procedures and flight records yield large amounts of data. Such data can certainly be examined, with the goal of improving aircraft performance.

As early as 1982 Collins developed estimates for aircraft fuel consumption, based on energy balance concepts [14]. While not a statistical technique, it did enable fuel conservation opportunities to be identified. In contrast, Chati and Balakrishnan have modeled aircraft engine fuel flow rate as a statistical system using Gaussian process regression [15]. Their approach can be used for multiple aircraft types, providing parameter values for each specific aircraft are known. Several factors affect aircraft fuel consumption, one of which is the mass of the aircraft. Turgut has developed regression models using aircraft initial mass, air speed, and altitude [16] that are highly significant. They indicate average fuel consumption for a wide body jet is found to be between two and three percent of the non-payload mass per flight hour. By reducing non-payload mass, fuel consumption over the life of the aircraft can be decreased.

2.2.5 Flight Conditions

Two flight conditions that affect fuel consumption of an aircraft are density altitude and turbulence. Density altitude is the air density as it corresponds to a given height above sea level for a given place of observation [17]. Air pressure affects density altitude the most. When compressed, a given mass of air is relatively denser and occupies less space. Conversely, when expanded, the same mass of air occupies more space and is relatively less dense. Environmental conditions affect this compression and expansion. At a constant temperature, if the pressure is reduced, the density is reduced. If pressure is increased, the density increases. Air density also decreases for high temperatures and increases for low temperatures. This inverse relationship holds true

when pressure remains constant. Though both temperature and pressure decrease with altitude, they have opposing effects on air density. However, the drop in air pressure dominates, resulting in air that is relatively less dense at higher altitudes, even if the temperature is colder.

When air is less dense, the performance of the aircraft decreases in the forms of reduced lift and engine efficiency [17]. Manipulation of the aircraft's control surfaces can compensate for reduced lift. However, both reduced lift and engine efficiency typically require an increase in the use of the throttle [17], which increases the rate of fuel consumption or engineering burn rate.

Turbulence is the movement of masses of air, without identifying visual clues. Though turbulence usually occurs at altitudes above 23,000 feet, it can occur at lower altitudes due to a change in temperature over a given direction, or the difference in the relative speed of two adjacent air masses. Both increased temperature and relative speed of an air mass decrease its density. So, turbulence can be defined as unanticipated and extreme changes in the density altitude.

To maintain stability of the aircraft's altitude and heading, a human pilot uses her intuition and knowledge to react appropriately to any conditions that occur. However, more common is to have an autopilot control an aircraft during cruise flight, since conditions needing human intervention are unlikely. When conditions do occur that require deviation from the expected flight plan, an autopilot's actions can lack the finesse of a human pilot, resulting in possible overcorrections to local conditions. If it could be shown that higher density altitude and turbulence conditions result in a higher rate of fuel

consumption than expected, because of autopilot overcorrections, this might justify an effort to re-engineer the autopilot reactions, in order to use less fuel.

III. Data Cleaning

3.1 Chapter Overview

This chapter covers the criteria and rationale for cleaning the raw data from the subject aircraft. Data cleaning will occur in three stages, with the results of one stage building on the previous stage or stages.

3.2 Problem Definition

The raw data provided for the subject aircraft is in the form of aircraft logs in a text file format. Examination of randomly selected files indicates each observation appears in a separate row. However, variable columns in the files do not appear to be distinct, and there appears to be multiple values for each cell. The raw data files also possess multiple formats. These files are unusable for analysis in their original form, as the data read from these files is not formatted for input into analysis functions.

For accurate data analysis, the data must first be tidied [4], [8]. Each observation in the data must be reflected in a separate row. Each variable must have a separate column, and each value must be in a separate cell. Since our research goal is to analyze a common dataset, only the variables common to all raw data files can be included in the data set.

Once an initial tidied data set is produced, it can be examined for the variables it contains. Only the variables related to the research need be examined. If desired variables are not found in the initial data set, alternative variables can be selected. These variables can be used as is, or used to calculate new variables.

A data set can be cleaned further in order to support the desired data analysis. One or more key variables can be examined for their value, or relationships to other variables. The data set rows or observations can then be filtered on selected criteria in order to reduce the size of the data set and further increase the accuracy of the data analysis.

3.3 Goals

The objective of the cleaning phase is to produce a data set that facilitates accurate data analysis. The proposed process includes cleaning the data in three stages. The first goal of the cleaning phase is to produce a data set that is tidied, and has the variables common to all originating raw data files. The next goal is to refine the initial data set, by selecting only the variables needed to support the research. The final goal is to select the rows or observations in the data set that contain the desired variable values or characteristics.

3.4 Approach

This section provides a high level view of the planned data cleaning process for the subject data files. The environment for data cleaning will consist of the R statistical language installed on a personal computer with at least 64 gigabytes of random access memory (RAM). This large amount of memory will enable data files to be read and manipulated. The interface for the environment will consist of the R Studio integrated development environment (IDE).

3.4.1 First Round

The following tasks will be performed using suitable R functions. First, the individual log text files will be read into system memory line by line. The data will then be spilt into individual columns. Any rows that contain all null or zero values will be discarded. These files will be written into the designated directory as comma separated value (CSV) files. Three cleaned files will then be chosen at random, and their variable lists compared. Only columns common to these individual files will be chosen to build a list of common columns. The cleaning script of R functions will be amended to select only the common columns before the individual files are written out. The expected output of this round is one cleaned CSV file corresponding to each log text input file.

3.4.2 Second Round

The common column list will be examined and variables pertinent to the research selected from it. Pertinent variables will include, but are not limited to those that quantify engineering burn rate (rate of fuel consumption), turbulence, and density altitude. The files from the first round of data cleaning will be filtered to select only pertinent variables. If desired variables are not present in the data, then alternative variables will be selected. These alternative variables will be used as they are, or used to calculate new variables. The expected output from this round of cleaning is one CSV file corresponding to each first round output file.

3.4.3 Third Round

The goals of this thesis are to define the relationships between engineering burn rate and density altitude, and engineering burn rate and turbulence. To do this, we must examine data where the density altitude is consistent. Additionally, we must select data

where turbulence indicators fluctuate in value. Both of these conditions occur during cruise flight, when the subject aircraft is attempting to maintain a desired course and heading. These conditions do not occur during deliberate ascent or descent. Furthermore, continuous time segments must be selected from the data of at least three minutes in duration.

In this third cleaning round, an R script will be developed that reads in the output files from the second round of cleaning, filters out cruise flight time segments of at least three minutes in duration, and merges these time segments into one CSV output file.

IV. Methodology

4.1 Goals

This research focuses on analyzing log data from a subject aircraft in order to define the relationships between density altitude and engineering burn rate, and turbulence and engineering burn rate. The following questions are addressed:

1. Can accurate linear regression models be developed to predict engineering burn rate for the subject aircraft using both altitude density and turbulence indicators?
2. Can accurate significant linear regression models be developed to predict engineering burn rate for the subject aircraft using only altitude density?
3. Can accurate significant linear regression models be developed to predict engineering burn rate for the subject aircraft using only turbulence indicators?

4.2 Approach

The goal of linear regression is to model a continuous response variable Y as a function of one or more X predictor variables. This model is then used to predict Y when only X is known. The relationship is defined mathematically by the following equation:

$$Y = \beta_1 + \beta_2 X + \varepsilon \quad (1)$$

In this equation, β_1 is the intercept, with β_2 representing the slope. They are the regression coefficients. The error term is represented by ε, and cannot be explained by the regression model.

In order to develop an accurate regression model, we must select predictor variables that possess a strong linear relationship to the response variable. For this research, predictor variables will include those that adequately represent density altitude

and turbulence. Ideally, these predictor variables should have a strong relationship to engineering burn rate. To evaluate the relationship between the predictor variables and engineering burn rate, a correlation matrix will be constructed in the analysis system environment. This correlation matrix will provide the strength of the linear relationship between the variables in the input data set on a scale of zero to one, or negative one to zero. Zero indicates no relationship, while one or negative one indicates the strongest relationships. Only those predictor variables that possess a relatively strong relationship with engineering burn rate, and adequately represent the desired attributes, will be evaluated further. The strength of the linear relationship will be calculated by Pearson's Correlation Coefficient [18], defined by the formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_{i-}\overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (2)$$

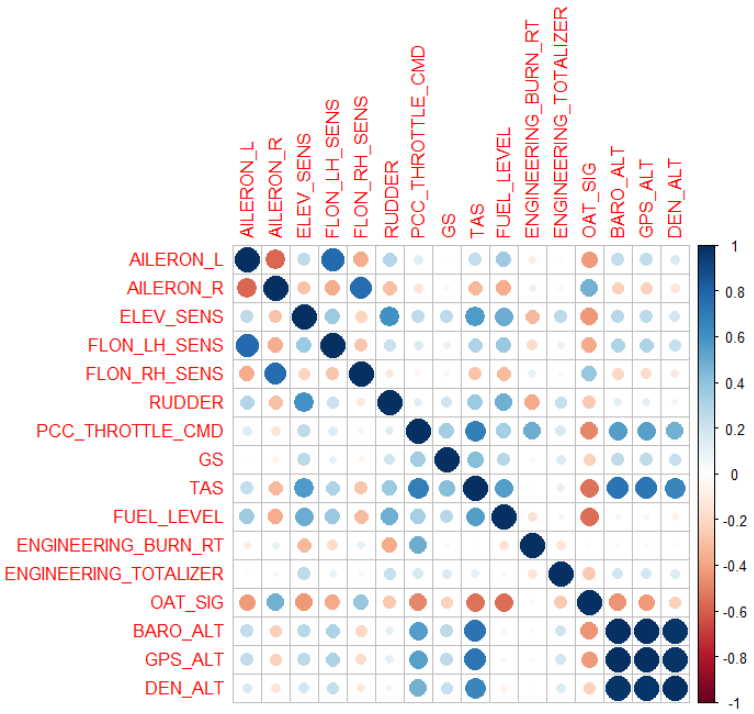

Figure 1: Correlation Matrix

Once predictor and response variables are selected, they are examined for outlying values using box plots. This plot visualizes the interquartile range (IQR) as a box, where the majority of the variable values reside in the distribution. The median value divides the box by a heavy solid line. Dotted lines drawn from the box to a solid line represent one and a half times the IQR. Any values that fall between the box and the solid line are suspected outlying values. Any variable values that fall outside of the solid line are confirmed outlying values.

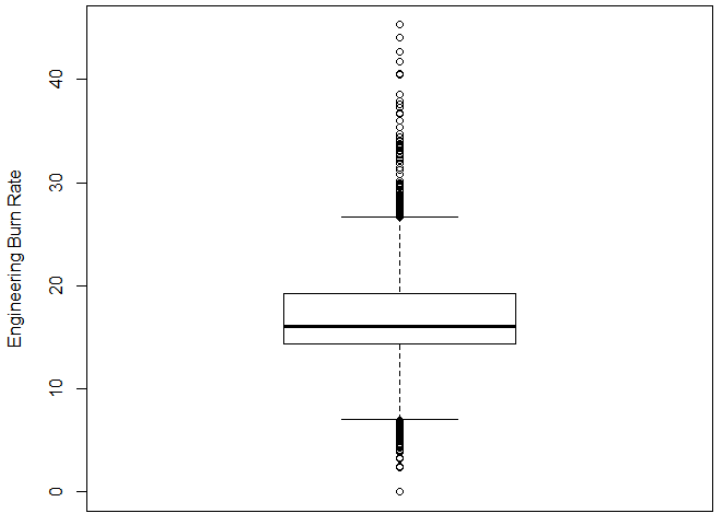

Figure 2: Inter Quartile Range Plot

To evaluate the normality of the selected predictor variables, a density plot of each variable will be constructed. This plot will show the shape of the variable value distribution, how smooth the distribution is, and if the distribution is skewed positively or negatively. It is possible to fit accurate linear regression models with non-normal data. However, heteroscedasticity may be present and should be tested for in the fitted models.

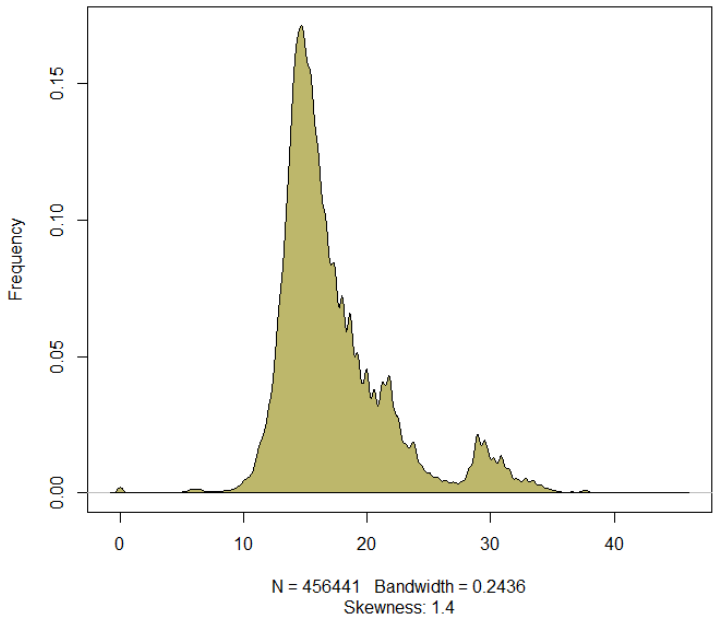

Figure 3: Density Plot

The input data set will be partitioned to support the development of the regression models. Using the same data to develop, then later test, regression models will result in overly optimistic models with high error rates. Accordingly, seventy percent of the data set will be designated for model development, with the remaining thirty percent being used to test the predictive performance of the model.

Linear regression models will be fit by two different methods. Simple linear regression using the ordinary least squares (OLS) method relies on the linear relationship between the predictor variables and the response variable. In contrast, a neural network uses a nonlinear process to compute a linear regression model [19], [20]. This model is a better choice to develop a predictive linear regression model, when predictor variables do not have a strong linear relationship with the response variable. Furthermore, the neural network models will only have one neuron in the hidden layer to avoid overfitting [21].

Linear regression models will be fit in three separate cases. For case one, the linear regression models will be fit with all of the predictor variables, with engineering burn rate being the response variable. For case two, only density altitude or its indicators will be used to predict engineering burn rate. For case three, only turbulence indicators will be used to predict engineering burn rate. The rationale for these cases is to better isolate density altitude and turbulence as variables affecting engineering burn rate, and quantify their relationships.

Performance of the fitted linear regression models will be measured in two ways. Heteroscedasticity in the fitted OLS linear regression models will be measured by the Breusch-Pagan test [22]. This test is unavailable for the fitted neural network linear regression models, as it is computed as a function of the OLS linear regression model residuals. The presence of heteroscedasticity affects the accuracy of OLS regression models as the residual errors are not evenly distributed as explanatory variable values increase [23]. If heteroscedasticity is not present, performance of both linear regression models can be compared by mean squared error (MSE), which is calculated after testing the predictive performance of the OLS linear regression model. This metric measures the average error between predicted and actual values. If heteroscedasticity is detected, the MSE will not be taken, as it will be assumed to be inaccurate for both models.

If the presence of heteroscedasticity in the initial regression models is detected, or if the models have poor performance as measured by MSE, a secondary round of analysis will be performed to evaluate the data set based on mean values for the variables of interest from each cruise flight time segment. Taking the mean value of the predictor and response variables instead of the individual values will better describe the central

tendency, or value, of the data. Additionally, taking the mean values in this way allows for better comparison between missions, rather than between individual observations in the missions.

A correlation matrix will be constructed from mean value data to check for improved relationships between predictor and response variables. Choice of predictor variables will be either confirmed or refuted based on the values of the new correlations. Analysis checks for outlying values, and normality of the mean predictor and response values will be repeated. If these checks demonstrate fewer outlying values, and improved normality of the data, then regression models based on the mean values will be developed, and their performance evaluated using the Breusch-Pagan test and MSE.

The accuracy of regression models will be improved by sorting the data in two different ways, resulting in four distinct partitions. If a secondary round of analysis is performed, then the mean value data set will be used, rather than the original analysis data set. First, the data will be sorted according to the value of the standard deviation of the turbulence indicators into more and less turbulent partitions. The standard deviation is used, as it is a measure of how much the average value of the turbulence indicators varies. More turbulent time segments will have higher standard deviations for turbulence indicators. Less turbulent time segments will have lower standard deviations for turbulence indicators. Second, the data will also be sorted according to the mean value of the density altitude indicators into higher and lower density altitude partitions. Sorting in these ways will narrow the data value ranges, which will assist in addressing heteroscedasticity, if it is present. All four partitions will undergo variable analysis, regression model fitting, and regression model performance analysis.

Regression models will not be cross validated. Cross validation using the same data set, but differing random draws for model development and predictive testing was considered, but ultimately dismissed. Even with multiple random draws, using the same data set for model training and prediction testing would result in overly optimistic models.

4.3 Analysis System

The analysis system will consist of the most recent version of the R statistical language installed on a personal computer (PC) with at least sixty-four gigabytes of random access memory (RAM). The large amount of RAM will facilitate the manipulation of the data. The interface will consist of the most recent release of the R Studio integrated development environment (IDE).

4.4 Analysis Tasks

4.4.1 Correlation Matrix Development

Develop a script in the analysis system to accept the analysis data set, and calculate the linear correlations between variables. The results will be presented in the form of a correlation matrix graph, with larger and darker symbols representing stronger linear correlations.

4.4.2 Variable Graphical Analysis

Develop a script in the analysis system to accept the analysis data set, and evaluate the normality of the selected predictor and response variables. For each variable of interest, a boxplot to check for outlying values, and a density plot to show the shape,

bandwidth or smoothness, and skew of the distribution of the variable data values will be produced.

4.4.3 Linear Regression Model (LRM) Development and Performance

Develop a script in the analysis system to accept the analysis data set, randomly partition the data set according to a seventy/thirty split, and calculate linear regression models for all three cases by OLS and neural network methods. The models should be developed using the majority of the partitioned data. The script should also include functions to test the predictive performance of the models, using the smaller amount of the partitioned data set. The output of this step will include plots to graphically show the predictive performance of the models compared to the model regression line. The Breusch-Pagan test will be run to assess heteroscedasticity. The MSE of both models will also be calculated.

4.4.4 Calculating Mean and Standard Deviations

If additional analysis is warranted, a script will be developed to accept the analysis data set, and calculate the means and standard deviations for each time segment for the variables of interest. A new CSV file containing the mean data will be written out.

4.4.5 Correlation Matrix Development for Means Data

Develop a script in the analysis system to accept the mean data set, and calculate the linear correlations between variables. The results will be presented in the form of a correlation graph, with larger and darker symbols representing stronger linear correlations.

4.4.6 Graphical Analysis for Means Variables

Develop a script in the analysis system to accept the mean data set, and evaluate the normality of the selected predictor and response variables. For each variable of interest, a boxplot to check for outlying values, and a density plot to show the shape, bandwidth or smoothness, and skew of the distribution of the variable data values will be produced.

4.4.7 LRM Development and Performance for Means

Develop a script in the analysis system to accept the mean data set, randomly partition the data set according to a seventy/thirty split, and calculate linear regression models for all three cases by OLS and neural network methods. The models should be developed using the majority of the partitioned data. The script should also include functions to test the predictive performance of the models, using the smaller amount of the partitioned data set. The output of this step will include plots to graphically show the predictive performance of the models compared to the model regression line. The Breusch-Pagan test will be run to assess heteroscedasticity. The MSE of both models will also be calculated.

4.4.8 Sort Means Data

Develop a script in the analysis system to sort either the original analysis data set, or the means data set, by ascending values for the standard deviations of the turbulence variables. The data set will then be divided into a top half and a bottom half. These partitions correspond to less and more turbulence, and each partition will be written out into as a new CSV file. An additional a script will be developed in the analysis system to sort either the original analysis data set, or the means data set, by ascending values for the

mean of the density altitude variable. The data set will then be divided into a top half and a bottom half. These partitions correspond to lower and higher density altitude, and each partition will be written out into as a new CSV file.

4.4.9 Variable Graphical Analysis for Less Turbulent

Develop a script in the analysis system to accept the relatively less turbulent data set, and evaluate the normality of the selected predictor and response variables. For each variable of interest, a boxplot to check for outlying values, and a density plot to show the shape, bandwidth or smoothness, and skew of the distribution of the variable data values will be produced.

4.4.10 LRM Development and Performance for Less Turbulent

Develop a script in the analysis system to accept the relatively less turbulent data set, randomly partition the data set according to a seventy/thirty split, and calculate linear regression models for all three cases by OLS and neural network methods. The models should be developed using the majority of the partitioned data. The script should also include functions to test the predictive performance of the models, using the smaller amount of the partitioned data set. The output of this step will include plots to graphically show the predictive performance of the models compared to the model regression line. The Breusch-Pagan test will be run to assess heteroscedasticity. The MSE of both models will also be calculated.

4.4.11 Variable Graphical Analysis for More Turbulent

Develop a script in the analysis system to accept the relatively more turbulent data set, and evaluate the normality of the selected predictor and response variables. For each variable of interest, a boxplot to check for outlying values, and a density plot to show the

shape, bandwidth or smoothness, and skew of the distribution of the variable data values will be produced.

4.4.12 LRM Development and Performance for More Turbulent

Develop a script in the analysis system to accept the relatively more turbulent data set, randomly partition the data set according to a seventy/thirty split, and calculate linear regression models for all three cases by OLS and neural network methods. The models should be developed using the majority of the partitioned data. The script should also include functions to test the predictive performance of the models, using the smaller amount of the partitioned data set. The output of this step will include plots to graphically show the predictive performance of the models compared to the model regression line. The Breusch-Pagan test will be run to assess heteroscedasticity. The MSE of both models will also be calculated.

4.4.13 Variable Graphical Analysis for Higher Density Altitude

Develop a script in the analysis system to accept the relatively higher density altitude data set, and evaluate the normality of the selected predictor and response variables. For each variable of interest, a boxplot to check for outlying values, and a density plot to show the shape, bandwidth or smoothness, and skew of the distribution of the variable data values will be produced.

4.4.14 LRM Development and Performance for Higher Density Altitude

Develop a script in the analysis system to accept the relatively higher density altitude data set, randomly partition the data set according to a seventy/thirty split, and calculate linear regression models for all three cases by OLS and neural network methods. The models should be developed using the majority of the partitioned data.

The script should also include functions to test the predictive performance of the models, using the smaller amount of the partitioned data set. The output of this step will include plots to graphically show the predictive performance of the models compared to the model regression line. The Breusch-Pagan test will be run to assess heteroscedasticity. The MSE of both models will also be calculated.

4.4.15 Variable Graphical Analysis for Lower Density Altitude

Develop a script in the analysis system to accept the relatively lower density altitude data set, and evaluate the normality of the selected predictor and response variables. For each variable of interest, a boxplot to check for outlying values, and a density plot to show the shape, bandwidth or smoothness, and skew of the distribution of the variable data values will be produced.

4.4.16 LRM Development and Performance for Lower Density Altitude

Develop a script in the analysis system to accept the relatively lower density altitude data set, randomly partition the data set according to a seventy/thirty split, and calculate linear regression models for all three cases by OLS and neural network methods. The models should be developed using the majority of the partitioned data. The script should also include functions to test the predictive performance of the models, using the smaller amount of the partitioned data set. The output of this step will include plots to graphically show the predictive performance of the models compared to the model regression line. The Breusch-Pagan test will be run to assess heteroscedasticity. The MSE of both models will also be calculated.

4.5 Summary

This chapter detailed the approach to develop accurate linear predictor models based on selected predictor and response variables. Variables are first tested for the strength of their relationships with other variables, and then selected based on those relationships, along with desired attributes. The data is then fitted to regression models and the results analyzed. Successive cycles of data revision allow for better characterization of the data, but also compensate for outlying values and heteroscedasticity.

V. Analysis and Results

5.1 Chapter Overview

This chapter outlines the results of data cleaning and linear regression analysis on the raw log data of the subject aircraft.

5.2 Data Cleaning Results

Preliminary examination of randomly selected raw input files revealed the existence of four distinct file formats. The raw files were separated into directories by format, and cleaning scripts developed for each format. These scripts were developed incrementally. First, the files were tidied with R functions to split each observation into a separate row, ensure each column possessed a separate variable, and ensure each cell contained only one value. This enabled the columns in each format to be examined.

One randomly selected file from each format was read into memory. The intersection of the names of the columns from each file was taken to produce a list of common columns, which was written out as a CSV file. This file was later read into memory by each cleaning script, and the list of columns used to select the columns in each file.

Additionally, two columns were added to facilitate data analysis. The first column added was “MISSION_ID, which assigned a unique number to each mission. Though indexing data existed in the “INFO” column, it was only necessary to distinguish missions, not uniquely identify them. The second column added was “DEN_ALT” for density altitude. Density altitude was calculated for each row based on the values for

"BARO_ALT" or barometric altitude, and "OAT_SIG" or outside air temperature according to the following formula [24]:

$$DEN_{ALT} = BARO_{ALT} + 118.8([BARO_{ALT}/500] + OAT_{SIG} - 15) \quad (3)$$

Therefore, the first cleaning round for each raw input file consists of tidying the data, selecting the common columns, and adding the two new columns. The results were written out as CSV files, with one first round cleaning file corresponding to each raw input file.

One file was examined from this first round of cleaning to determine which column variables were pertinent to the research. No variables were found that directly measured turbulence. However, column variables containing the values for the settings of the control surfaces of the subject aircraft were found – these were selected as alternate turbulence indicators. A column variable directly corresponding to engineering burn rate was found, along with two other indicators of fuel consumption. Other variables selected include those for mission identification, mission information, ground speed, total air speed, outside air temperature, barometric altitude, GPS altitude, and finally density altitude.

The list of desired columns was entered into a CSV file. For the second round of cleaning, the file of desired columns was read into memory by the cleaning scripts, and the list used to select only the columns pertinent to the research from each file cleaned in the first round. The results were written out as CSV files, with one second round cleaning file corresponding to each first round cleaning file.

The third round of cleaning consisted of finding cruise flight time segments of at least three minutes in duration in the second cleaning round files, and writing out the

information for all time segments to one CSV file. The cruise flight time segments were found by an incremental process. First, a script was developed that plotted the GPS altitude for each second round cleaning file. In analyzing the files, cruise flight time segments usually appeared above 4500 feet. By eliminating rows that had GPS altitude values below 4500 feet, rows describing deliberate ascent and decent were largely eliminated.

Next, cruise flight was abstracted to be a horizontal line with respect to GPS altitude. Minor fluctuations in GPS altitude do occur during cruise flight, but do not represent significant deviations from the horizontal line. Any deviations from the horizontal line can be quantified by finding the slope between points that comprise the line. Here, points are represented by GPS altitude values and the row indexes where the GPS altitude values originated. To find the slope, the current point and previous point are used as parameters to an ordinary least squares (OLS) regression function that computes a straight regression line. The slope of the line is obtained, and if the value of the slope is between negative three and positive three, cruise flight is assumed and a time segment indicator assigned. If subsequent pairs of points possess slopes that fall within the desired range, they are added to the same time segment. If subsequent pairs of points possess slope values that fall outside of the desired range, a new time segment indicator is assigned.

For each time segment, the number of rows is counted. One observation or row was recorded every second in the data, so it is assumed the number of rows in a time segment corresponds to the duration of the time segment in seconds. To obtain time segments of at least three minutes in duration, only the time segments with at least one

hundred and eighty rows are selected. These time segments are written out to one CSV file, resulting in the product of the third round of cleaning – a data set suitable for analysis.

5.3 Analysis Results

This research has demonstrated reasonably accurate linear models can be developed for non-normal, heteroscedastic data under certain conditions. First, correlation coefficients alone should not be the sole determinates of the predictor variables used, as the coefficients are affected by non-normal data distributions. Second, taking the mean values better showed the central tendency in the variable data. This compensated for the majority of the outlying values in the data. Third, sorting and then partitioning the data reduced the range of the variable values. This mechanism improved the shape of the variable value distributions in the data sets, which resulted in more accurate linear regression models developed for the partitioned data.

Analysis of the original data set began with constructing a correlation matrix for the variables of interest in the analysis system environment, which is shown in Appendix A. Examination of this plot revealed the “ELEV_SENS”, “RUDDER”, and “PCC_THROTTLE_CMD” variables had the strongest correlations with engineering burn rate. Density altitude did not have a strong relationship with engineering burn rate, contrary to expectations. Higher density altitudes reduce aircraft performance, and necessitate higher rates of fuel consumption [17]. Despite the lack of a strong relationship, density altitude was still selected as one of the final predictor variables, along with the elevator sensor, rudder, and throttle.

Examination of the characteristics of the variable values showed every variable (predictors and response) had a significant number of outlying values, accompanied by an obviously non-normal distribution. These results are in Appendix B. Recall that outlying values in a data set affect the mean of a variable, by either increasing or lowering its true value [25]. Since mean values are used in the calculation of the Pearson correlation coefficients, outlying values will affect the strength of the linear relationships between variables [26]–[28]. Therefore, correlations between the selected variables do not reflect their true relationships [29]. While accurate linear regression models can still be developed with non-normal data sets [30], the shapes of the distributions must be similar in order for heteroscedasticity to be absent. Since there are no variable distributions with similar shapes, heteroscedasticity in the fitted OLS linear regression model is suspected. Furthermore, the linear regression models are expected to fit poorly for all three cases.

When using the original analysis data set, linear regression model development was unsuccessful for all three cases. Plots of the actual versus predicted values of the engineering burn rate did not show clustering around the regression line. These are shown in Appendix C. Instead, the plot points were evenly spread across the regression line, indicating both the OLS and neural network linear regression models were not accurate. The Breusch-Pagan test for all three cases produced a very small p-value, which indicated strong evidence for heteroscedasticity. Since heteroscedasticity was present, the MSE figures for all cases were not calculated. The figures would not reflect the majority of the data.

Since linear regression model development was unsuccessful using the original analysis data set, the decision was made to calculate the mean and standard deviation

values for each variable in each time segment. Taking the mean values would show the central tendency of the majority of the data, and would reduce the influence of outlying values. Calculating the standard deviations would enable sorting on turbulence indicators.

A correlation matrix constructed for the mean and standard deviation data showed improved relationships for the chosen predictor variable means and the engineering burn rate mean (Appendix E). However, the correlation values are still suspect.

An examination of the mean variable characteristics (Appendix F) indicated fewer outlying values, though the distributions were still non-normal. However, the shapes of the distributions were less random and relatively more normal than the distributions for the original analysis data set. The shapes of the mean variable distributions were still dissimilar to each other, indicating poor performance of the linear models.

In general, the OLS and neural network linear regression models showed improved performance using the mean variable values due to better value distributions for the predictor and response variables. These plots are in Appendix G. While some actual versus predicted engineering burn rate mean plot points seemed to cluster around the regression line, there were still a notable amount of plot points that were far from the line. Taking the means did not seem to eliminate the influence of all the outlying variable values. The Breuch-Pagan test in all three cases indicated the presence of heteroscedasticity, though the p-values were higher than for the models fitted with the original data.

To narrow the variable value ranges, the means data was first sorted based upon the standard deviations for the throttle variable. Of the turbulence indicators, the throttle

variable seemed to have the strongest correlation with the engineering burn rate, so it was used exclusively to sort the data into less and more turbulent partitions. Lower values for the throttle standard deviation indicated a lower amount of turbulence, with the reverse also being true. The data was sorted in ascending order and divided into two partitions.

The variable characteristics for the less turbulent mean data set showed narrowed ranges, and fewer outlying values. The plots are shown in Appendix H. The shapes of the distributions were also improved compared to the mean data set, with the shapes of the distributions being closer to each other, except for density altitude. This distribution was very different from the distribution shape for the engineering burn rate. The distribution for the throttle mean variable showed the most similarity to the distribution for the engineering burn rate mean variable.

OLS and neural network linear regression models showed improved performance using the less turbulent mean data set. The plots reside in Appendix I. The plot points of the actual versus predicted engineering burn rate mean for all three cases show clustering around the regression line, with fewer outlying values. The Breusch-Pagan test indicated the absence of heteroscedasticity in case one (all predictors), as well as case three (only turbulence predictors). Homoscedasticity was present in case two (density altitude), though the p-value was higher for this case than the p-value for the same case for the entire mean data set. Using all predictors, the MSE figures for the resulting linear regression models was 6.53 for the OLS model and 5.75 for the neural network model. For only the turbulence predictors, the MSE figures were 6.71 for the OLS model and 5.04 for the neural network model. The figures seem to indicate density altitude is a necessary predictor for an accurate statistical model predicting engineering burn rate.

However, density altitude does not have a strong enough relationship with the engineering burn rate to serve as the only predictor variable.

The variable characteristics for the more turbulent mean data set still display outlying values for all but one variable. The plots are in Appendix J. The shapes of the value distributions are also dissimilar to one another, except for those for the throttle mean and engineering burn rate mean. These distribution shapes possess weak similarity.

For the relatively more turbulent mean data set, only case one demonstrated acceptable performance of the linear models. The analysis plots reside in Appendix K. The actual versus predicted engineering burn rate mean plot points cluster well around the regression line, with very few outlying plot points. Since the Breusch-Pagan test indicated the absence of heteroscedasticity, the MSE was calculated for both linear regression models. The MSE for the OLS model was 1.05, while the MSE for the neural network model was 1.18. Both linear regression models have excellent performance with very low error rates. The actual versus predicted plots for case two indicate poor predictive performance of the models and with heteroscedasticity being present according to the Breusch-Pagan test. Case three actual versus predicted plots showed uneven clustering around the regression line. The Breusch-Pagan test indicated the presence of heteroscedasticity.

The variable value ranges for the mean data set were also narrowed by sorting on the mean value for the density altitude variable for each time segment, in ascending order. The data set was divided into two even partitions, corresponding to relatively higher and lower density altitude.

The variable characteristics for the relatively higher density altitude mean data set did not show narrowed value ranges for all variables. Additionally, the box plots still showed some outlying values. The plots produced are in Appendix L. However, the shapes of the value distributions do show increased similarity to each other, though the shapes are not exact matches.

Linear regression model development was successful for the relatively higher density altitude mean data for cases one and three. For both cases, the plots demonstrated acceptable clustering around the regression line, though the values were concentrated at the lower end of the line. These plots are in Appendix M. Both cases possessed a lack of heteroscedasticity as determined by the Breusch-Pagan test. For case one, the MSE for the OLS model was 6.65, while the MSE for the neural network model was 5.66. For case three, the MSE for the OLS model was 6.73, while the MSE for the neural network model was 6.23.

The variable characteristics for the relatively lower density altitude mean data demonstrated narrowed value ranges for the engineering burn rate mean, rudder mean, and throttle mean. However, these variables did show some outlying values. The distribution shapes were most similar between the rudder mean and the elevator sensor mean, and the throttle mean and the engineering burn rate mean. The plots reside in Appendix N. The shape of the density altitude distribution did not resemble that for any other variable.

Regression model development was successful for the relatively lower density altitude mean data for all three cases. For all cases, the actual versus predicted plot points of the engineering burn rate mean were more evenly distributed along the regression line,

though outlying points were still present. The points were also concentrated toward a specific range on the line. These plots are in Appendix O. Heteroscedasticity was absent in all three cases. For case one, the MSE for the OLS model was 4.18. The MSE for the neural network model was 5.09. For case two, the MSE for the OLS model was 6.10. The MSE for the neural network model was 8.04. For case three, the MSE for the OLS model was 4.27. The MSE for the neural network model was 4.65. These results are congruent with previous results that indicated the density altitude variable is necessary for a more accurate linear regression model, but the variable is not the most accurate predictor of engineering burn rate when used alone.

5.4 Summary

Using correlation alone to assess potential predictor variables is unwise. Unless it is known that all variables follow a normal distribution, the correlation coefficient should be assumed to be inaccurate. In particular, outlying variable values will cause the coefficient to be either too low or too high, depending on where the outlying values reside on the value distribution. Human intervention is needed to determine if the correlations make sense.

Taking the means of the variables improves linear regression model development results. However, this strategy does not compensate for all outlying values, and does not automatically narrow variable value ranges. Additional sorting is needed to narrow variable value ranges, and lower the number of outlying values, though this strategy is not always successful. Taking the means also improves the shape of the variable value distributions, but cannot improve the shape alone.

Additionally, the research has shown reasonably accurate linear regression models can be developed using density altitude and turbulence variables to predict engineering burn rate, using only density altitude to predict engineering burn rate, and using turbulence indicators to predict engineering burn rate. However, the accuracy of the linear regression models relies upon variables that possess similar variable value ranges, as well as similar value distribution shapes.

VI. Conclusions and Recommendations

6.1 Chapter Overview

This chapter presents a summary of the research conclusions, impact, and future work. Section 6.2 provides conclusions from the data analysis results. Section 6.3 discusses the impact of the research. Section 6.4 details ideas and recommendations for future work and related research.

6.2 Conclusions of Research

This research successfully developed linear regression models using non-normal, and heteroscedastic data. Each research goal from section 4.1 is discussed below. Each goal is satisfied successfully through the conducted analysis.

6.2.1 Development of an accurate statistical model using all predictors

Development of accurate statistical models using all chosen predictors was the most successful. However, the data used to develop the models did need to be transformed so that the predictor variables possessed similar ranges and distribution shapes to the engineering burn rate response variable. All predictors were necessary to more accurately predict engineering burn rate, but were not sufficient individually to do so. Additionally, predictor variables that possessed the greatest similarity in value range and distribution shape to engineering burn rate had the most influence in the model.

6.2.2 Development of an accurate statistical model using density altitude

Development of accurate statistical models using density altitude alone as a predictor was ultimately successful, but only when the data was altered, so that the density altitude variable possessed an acceptably similar value range and distribution

shape to the engineering burn rate response variable. While the data was transformed to improve model performance, the characteristics of the density altitude predictor variable were consistently the most dissimilar to those of the engineering burn rate variable. This model was developed to quantify the influence of density altitude on engineering burn rate by itself, but it should be noted this case consistently had the poorest performance due to the dissimilar variable characteristics.

6.2.3 Development of an accurate statistical model using turbulence

Development of accurate statistical models using turbulence indicators alone was successful when the data was altered so the predictor variables possessed similar value ranges and distribution shapes to the engineering burn rate response variable. Since the throttle variable characteristics were the most similar to the characteristics of the engineering burn rate variable, the throttle variable had the most influence in this model.

6.3 Significance of Research

The contributions this research makes are the characterization of the provided variable data and its use in developing accurate linear regression models. Based upon this knowledge, data acquisition practices can be altered to reduce heteroscedasticity. More accurate linear regression models can also be developed with additional information.

6.4 Recommendations for Action and Future Research

6.4.1 Altering data acquisition procedures

Altering data acquisition procedures will result in less heteroscedasticity in the data. This can be done in two different ways. First, procedures that generate the data can

be altered so that the full value ranges of the individual variables are better expressed, and form more normal distributions.

Two, a more accurate algorithm can be developed to determine cruise flight as well as correlations between time segments. Using data correlated in specific patterns will result in more accurate linear regression models [31]. Nikolic indicates the length of time segments matters when measuring variance of, and correlation between variables. Relatively long time segments tend to suppress the variance of a variable over time, while short time segments emphasize it. To better capture variance, time segments should be reduced in duration from three minutes. When a suitable length for time segments is selected, the standard deviation of the GPS altitude can be calculated for all segments. Cruise flight time segments will have standard deviations for GPS altitude at or near zero. Once cruise flight time segments are determined, correlations between variables within time segments can be calculated. Time segments with specific correlation patterns can then be separated into separate data sets. For example, time segments with a high correlation between density altitude and engineering burn rate, but little correlation with turbulence indicators and engineering burn rate, can be exclusively used to develop linear regression models to detail the relationship between density altitude and engineering burn rate.

6.4.2 Using more accurate linear regression models

The goals of the research related to developing linear regression models that explicitly showed the relationships between desired predictor variables and the engineering burn rate response variable. However, using these variables alone did not fully characterize all the factors that affect the engineering burn rate. Chati's work [15]

indicates using additional variables such as dynamic pressure, wing reference area, and takeoff mass of the aircraft result yields better results, provided the data used has a Gaussian distribution. As shown here, the use of normal data with a Gaussian distribution is not a necessary requirement for accurate linear regression model development. Assuming heteroscedasticity has been reduced by better data acquisition procedures, new linear regression models based upon Chati's work will be more accurate.

Appendix A - Experiment 1: Correlation Plot for Original Data

An R script was developed to output a correlation matrix plot of the candidate variables. Here, larger and darker symbols indicate stronger linear relationships. It should be noted that there were no direct indicators of turbulence found in the raw data. Therefore, subject aircraft control surface settings were selected as alternate turbulence indicators.

According to the plot, the designated response variable engineering burn rate has the strongest relationships with the candidate predictor variables elevator sensor reading, rudder setting, and throttle setting. Other candidate predictor variables show little to no correlation with the engineering burn rate, including density altitude. However, since we wish to establish the relationship between density altitude and engineering burn rate, density altitude will be included as a predictor variable. The variables of interest are therefore engineering burn rate, rudder, throttle, elevator sensor, and density altitude. A new analysis data set was created, comprised of only the variables of interest.

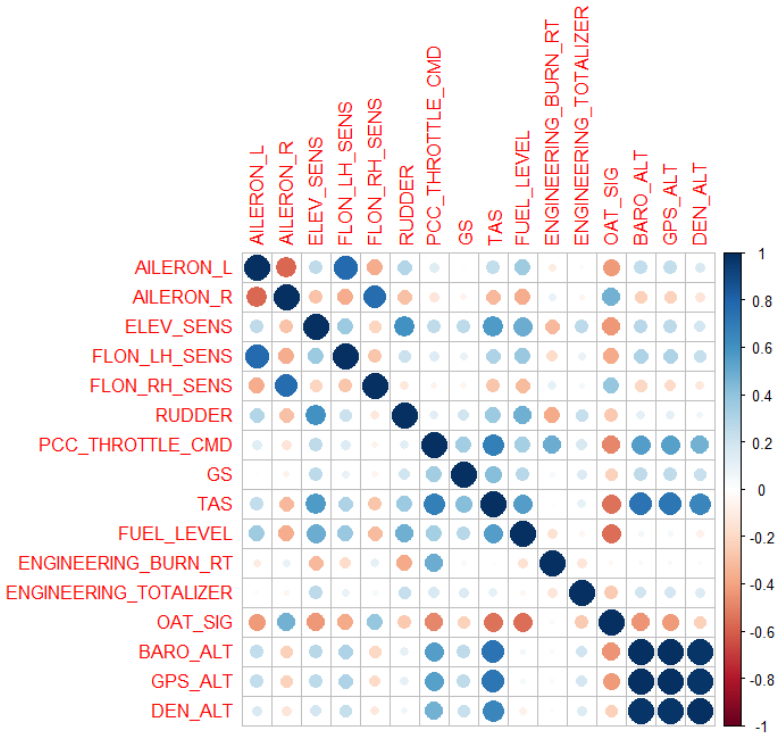

Figure 4: Correlation Matrix for Original Data

Appendix B – Experiment 2: Variable Analysis

For each of the variables of interest, a box plot showing outlying values, and a density plot showing the shape of the variable distribution along with skewness and bandwidth were constructed.

B.1 Engineering Burn Rate

a. Box Plot

The box plot of the values from the data for the engineering burn rate shows a significant number of outlying values, along with a relatively narrow IQR. This suggests the distribution for this variable might not be normal.

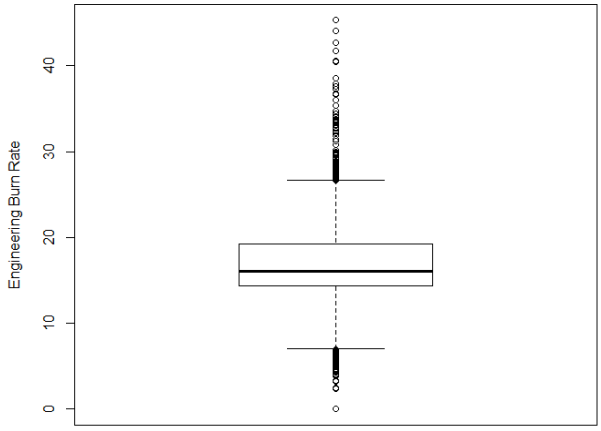

Figure 5: Engineering Burn Rate Box Plot

b. Density Plot

The density plot for engineering burn rate shows a less than normal distribution for the values of the variable. Most notable is the fact that the distribution shows a secondary peak, which is caused by the high number of outlying values, first found in the

box plot. This secondary peak also causes the data to be skewed positively. The bandwidth indicates the underlying distribution for the engineering burn rate variable was difficult to estimate. As a result, the density plot shows a lack of smoothness for higher values.

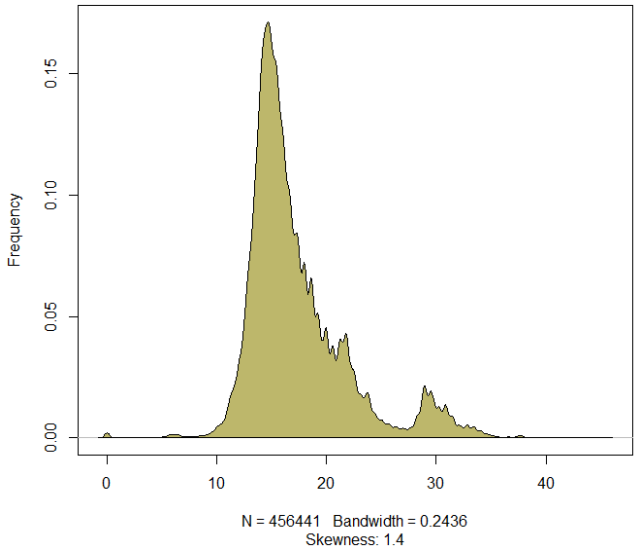

Figure 6: Engineering Burn Rate Density Plot

B.2 Density Altitude

a. Box Plot

The box plot for the values of the density altitude variable indicates a high number of outlying values in the data, relative to a narrow IQR. The underlying value distribution for density altitude will not be normal.

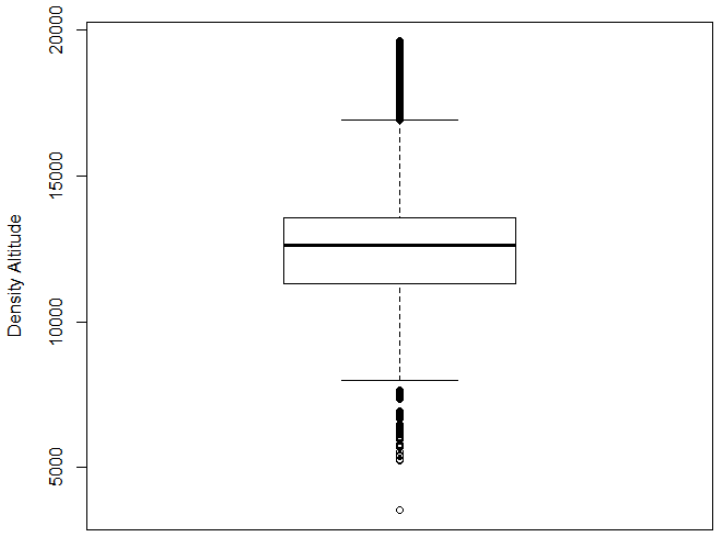

Figure 7: Density Altitude Box Plot

b. Density Plot

The density plot shows a less than normal distribution for the values of density altitude. More outlying values exist on the left hand tail of the graph, skewing the distribution negatively. The high figure for bandwidth indicates a severe lack of smoothness when estimating the underlying value distribution.

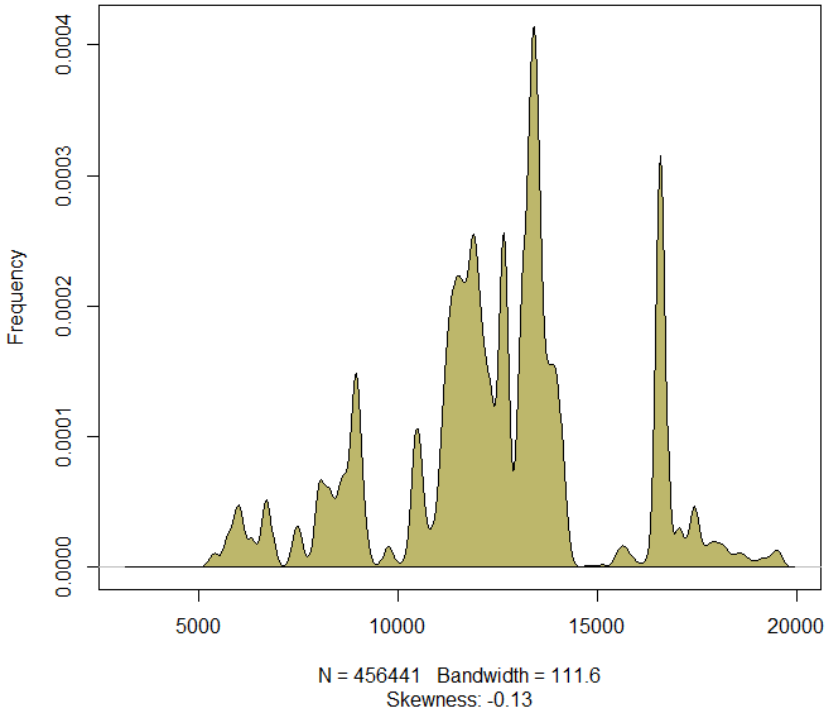

Figure 8: Density Altitude Density Plot

B.3 Throttle

a. Box Plot

The box plot for the variable values of the throttle indicates a high number of outlying values, particularly for the lower end of the range. Since the IQR is relatively narrow, the density plot should show the underlying value distribution is not normal and is skewed.

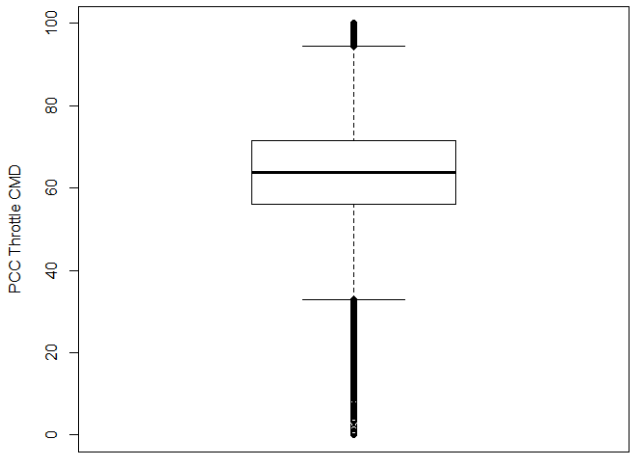

Figure 9: Throttle Box Plot

b. Density Plot

The density plot for the values of the throttle variable indicates a non-normal distribution that is skewed positively, due to the high number of values at the maximum end of the range for this variable. Bandwidth indicates poor estimation of the underlying value distribution.

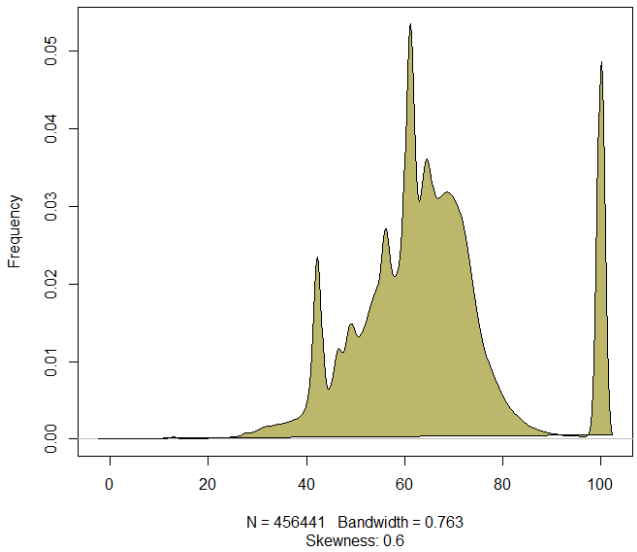

Figure 10: Throttle Density Plot

B.4 Rudder

a. Box Plot

The box plot for the variable values of rudder indicates a high number of outlying values for a very narrow IQR. The underlying distribution will not be normal.

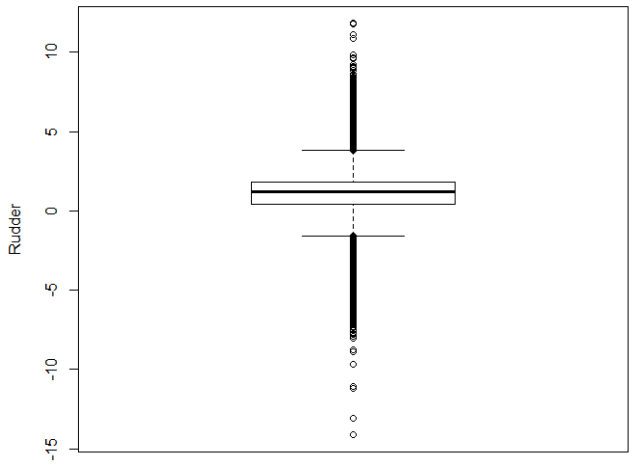

Figure 11: Rudder Box Plot

b. Density Plot

The density plot for the values of rudder show a non-normal distribution, with a negative skew due to the higher number of values in the lower part of the value range. Bandwidth indicates a relatively good fit to the underlying value distribution.

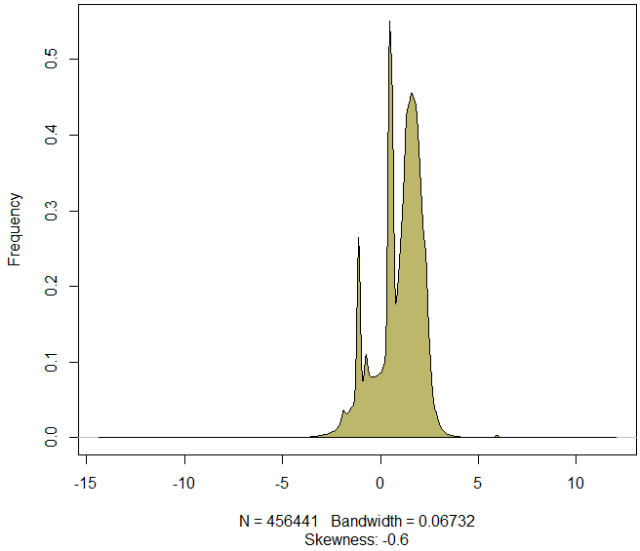

Figure 12: Rudder Density Plot

B.5 Elevator

a. Box Plot

The box plot for the elevator sensor indicates a high number of outlying values relative to a narrow IQR. It appears more outlying values reside in the lower range. The density plot should show a non-normal distribution that is skewed negatively.

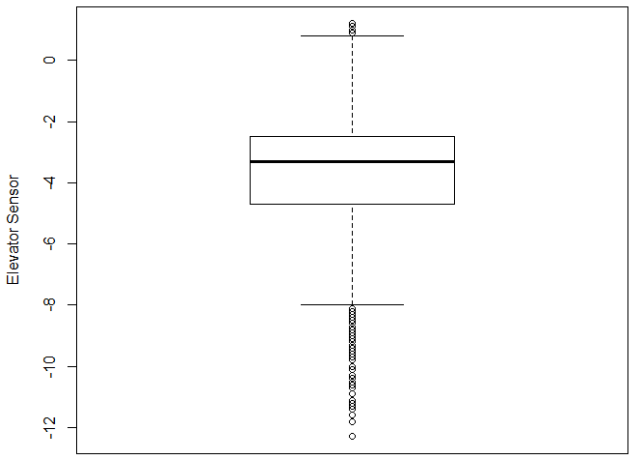

Figure 13: Elevator Box Plot

b. Density Plot

The density plot for the elevator variable shows a non-normal distribution that is skewed negatively. Bandwidth indicates a good fit for the underlying value distribution, even though it is non-normal.

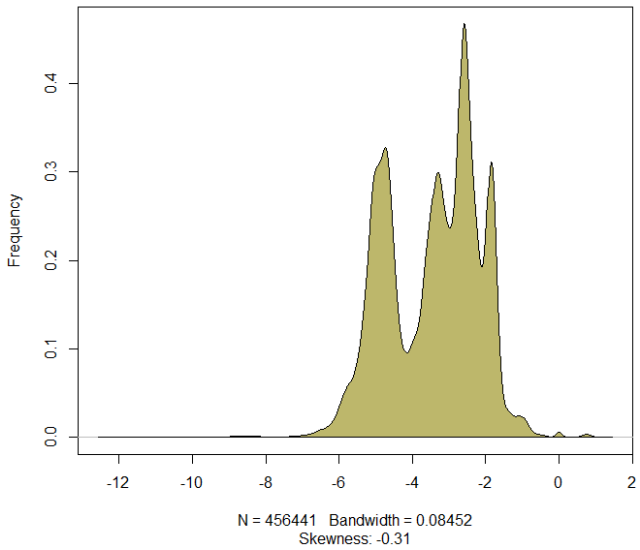

Figure 14: Elevator Density Plot

Appendix C – Experiment 3: Linear Regression Original Data

Regression models using the variables of interest were developed for three distinct cases. These cases are intended to quantify the relationship between turbulence indicators, density altitude, and engineering burn rate, between density altitude and engineering burn rate, and between turbulence indicators alone and engineering burn rate.

C.1 Turbulence Indicators and Density Altitude

A script was developed in the analysis environment to accept the analysis data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, all four predictor variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line and possess a significant amount of spread.

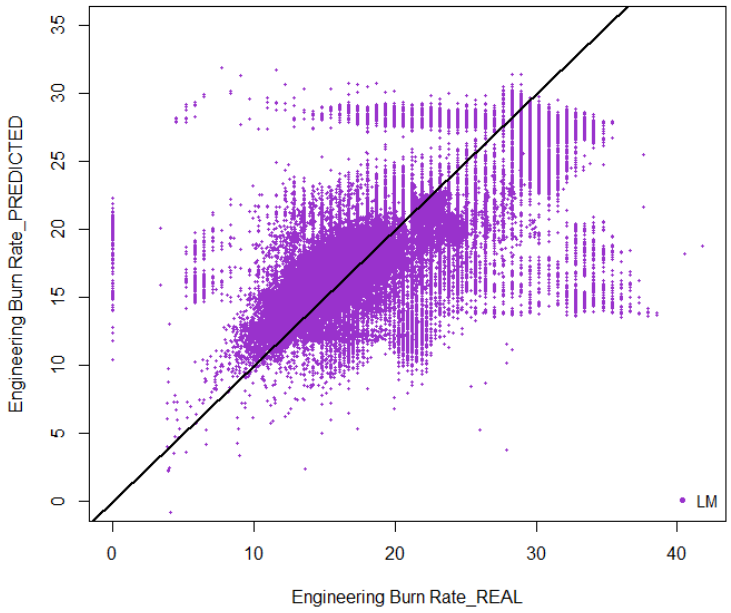

Figure 15: OLS Model – All Predictors

b. Neural Network Regression

Neural network regression shows a significant spread in the predicted values for engineering burn rate. However, the range of predicted values is narrower than that of the OLS model.

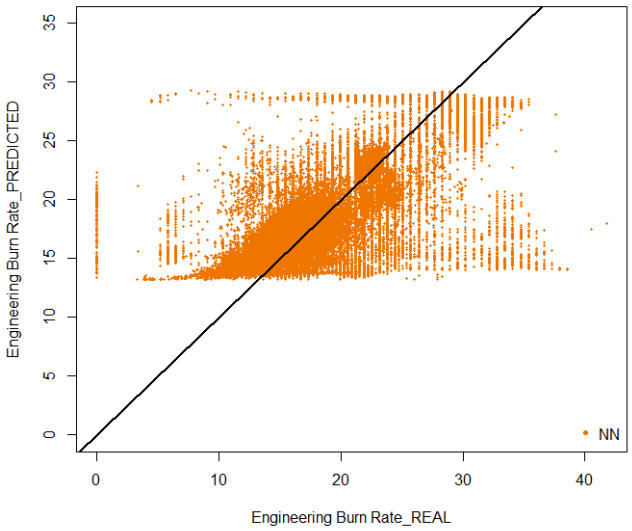

Figure 16: NN Model – All Predictors

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of less than 0.00000000000000022204. The very small p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

C.2 Density Altitude

A script was developed in the analysis environment to accept the analysis data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the density altitude predictor variable was used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line and possess a significant amount of spread, though the range is narrow.

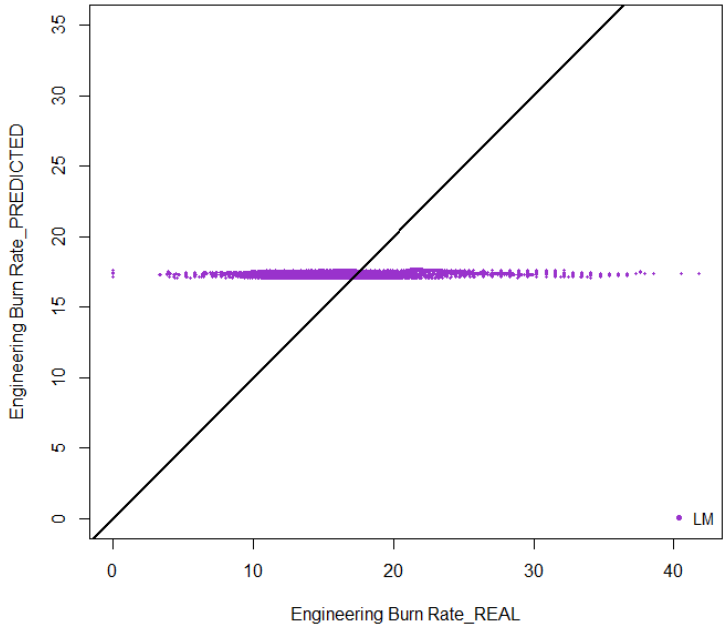

Figure 17: OLS Model - Density Altitude Predictors

b. Neural Network Regression

The neural network model did not converge.

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of less than 0.00000000000000022204. The very small p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural

Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

C.3 Turbulence Indicators (Throttle, Rudder, Elevator)

A script was developed in the analysis environment to accept the analysis data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the turbulence indicator variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line and possess a significant amount of spread.

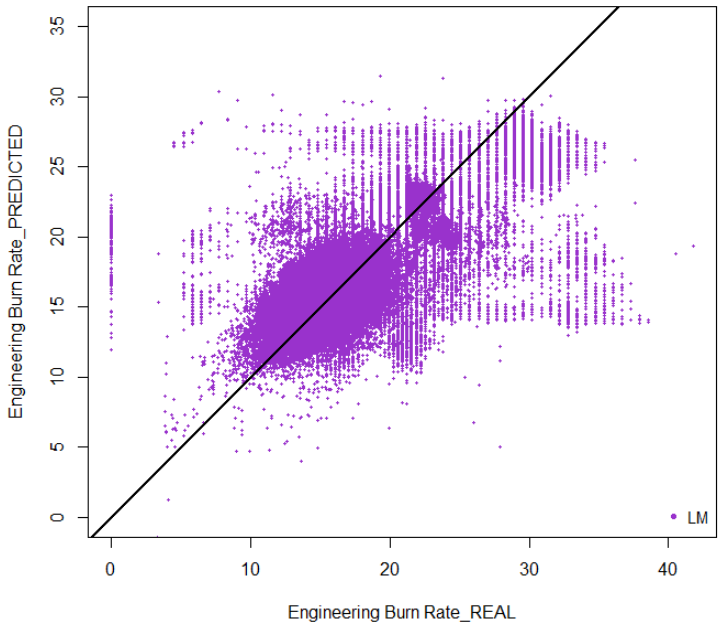

Figure 18: OLS Model - Turbulence Predictors

b. Neural Network Regression

Neural network regression shows a significant spread in the predicted values for engineering burn rate. However, the range of predicted values is narrower than that of the OLS model.

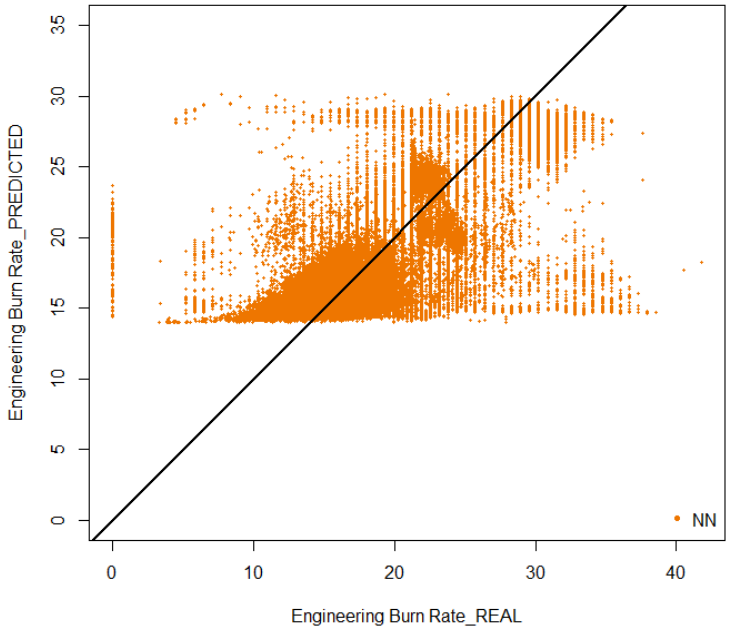

Figure 19: NN - Turbulence Predictors

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of less than 0.00000000000000022204. The very small p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

Appendix D – Experiment 4: Calculating Mean and Standard Deviation Values

Additional analysis was warranted due to the heteroscedasticity present in the data. A script was developed to calculate the means and standard deviations of the variable values for each time segment in the data. A new CSV file containing the information was written out.

Appendix E – Experiment 5: Correlation Matrix Means Data

An R script was developed to output a correlation matrix plot of the means data. Here, larger and darker symbols indicate stronger linear relationships.

According to the plot, the response variable engineering burn rate mean has the strongest relationships with the predictor variables density altitude mean, rudder mean, and throttle mean. Elevator mean shows a very weak correlation to the engineering burn rate mean. Other candidate predictor variables show little to no correlation with the engineering burn rate. Since we are examining only the means data, the correlations with variable standard deviations will be ignored.

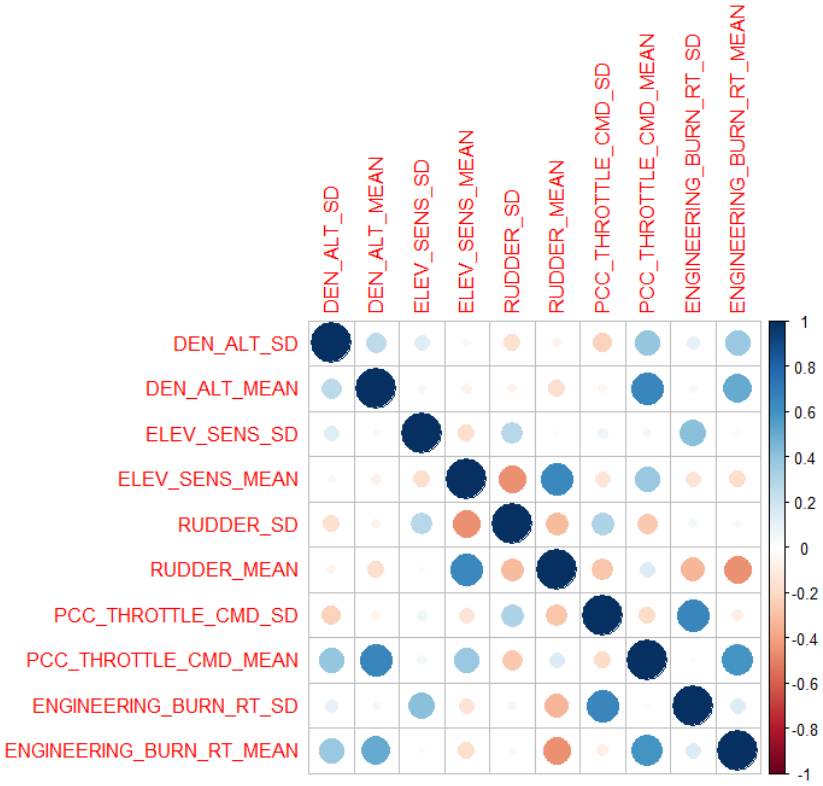

Figure 20: Correlation Matrix - Means Data

Appendix F – Experiment 6: Variable Analysis Means Data

For each of the variables of interest from the means data, a box plot showing outlying values, and a density plot showing the shape of the variable distribution along with skewness and bandwidth were constructed.

F.1 Engineering Burn Rate

a. Box Plot

The box plot for the engineering burn rate mean values indicates a lower number of outlying values compared to the original analysis data. Since outlying values are still present, the distribution is suspected to be non-normal.

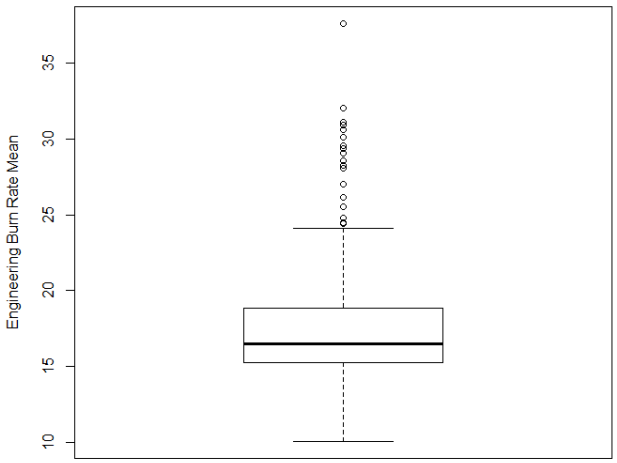

Figure 21: Engineering Burn Rate – Box Plot Means Values

b. Density Plot

The density plot for the engineering burn rate means variable shows a non-normal distribution that is skewed positively. Bandwidth indicates a relatively good fit for the underlying value distribution, even though the distribution is non-normal.

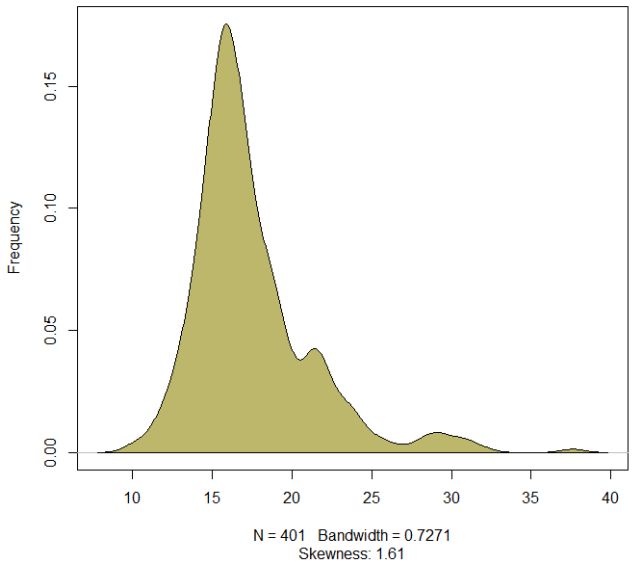

Figure 22: Engineering Burn Rate - Density Plot Means Values

F.2 Density Altitude

a. Box Plot

The box plot for the density altitude mean values indicates a lower number of outlying values compared to the original analysis data. Since outlying values are still present, the distribution is suspected to be non-normal.

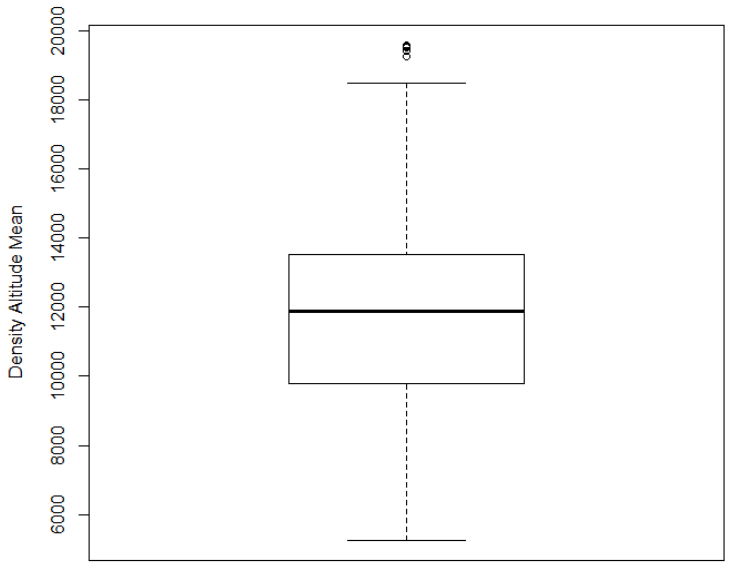

Figure 23: Density Altitude - Box Plot Means Values

b. Density Plot

The density plot for the density altitude means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates a poor fit for the underlying non-normal value distribution.

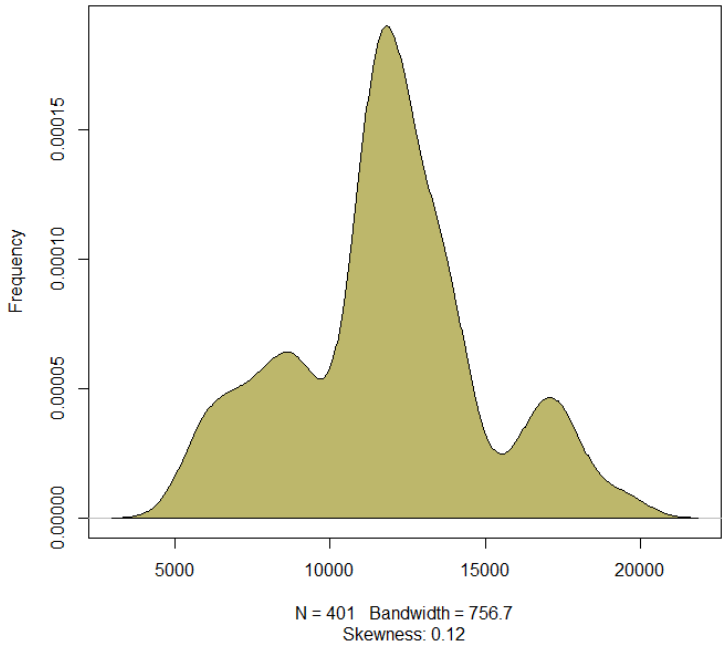

Figure 24: Density Altitude - Density Plot Means Values

F.3 Throttle

a. Box Plot

The box plot for the throttle mean values indicates a lower number of outlying values compared to the original analysis data. Since outlying values are still present, the distribution is suspected to be non-normal.

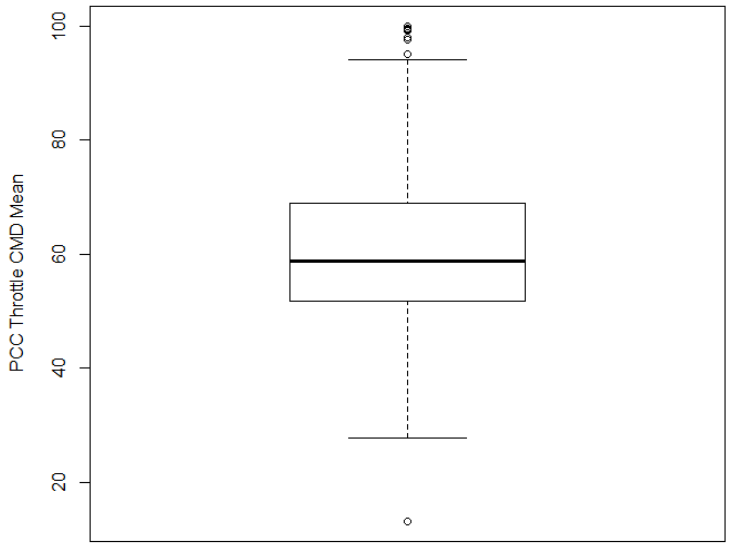

Figure 25: Throttle - Box Plot Means Values

b. Density Plot

The density plot for the throttle means variable shows a non-normal distribution that has a positive skew. Bandwidth indicates a relatively good fit for the underlying non-normal value distribution.

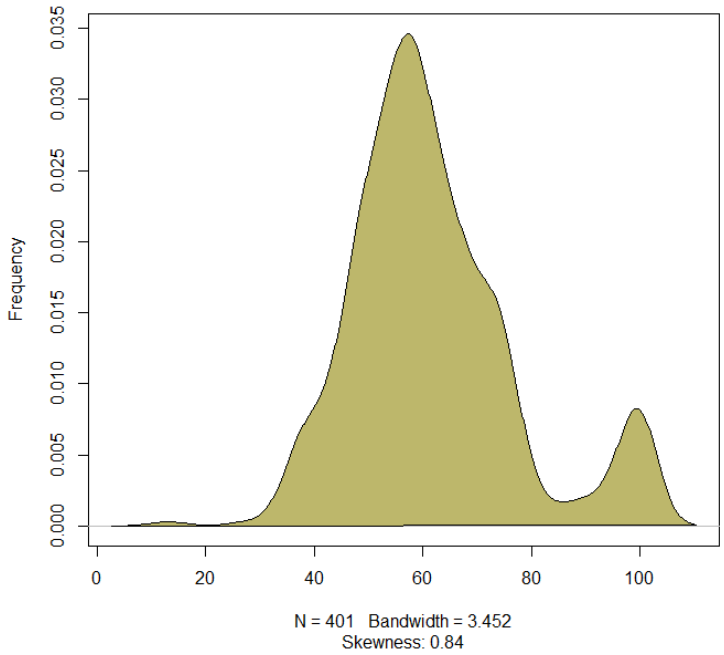

Figure 26: Density Altitude - Density Plot Means Values

F.4 Rudder

a. Box Plot

The box plot for the rudder means variable shows very few outlying values. Non-normal distribution of the variable values is not suspected, but should be diagnosed by the density plot. Bandwidth indicates a good fit to the underlying value distribution.

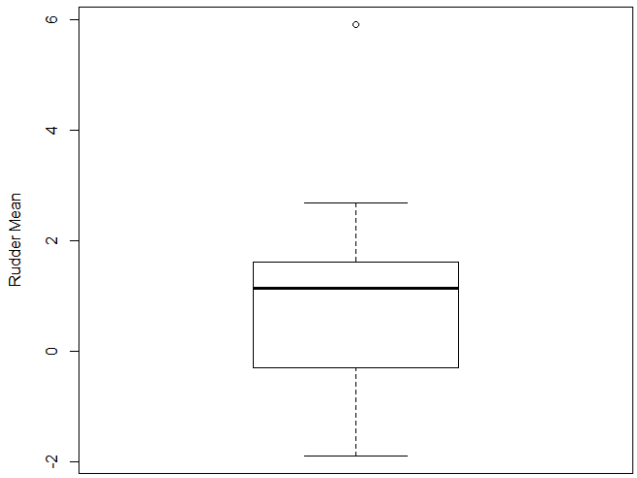

Figure 27: Rudder - Box Plot Means Values

b. Density Plot

The density plot of the rudder means variable shows a non-normal distribution with little negative skew. The double peaks suggest the values are clustered around separate values. Bandwidth indicates a good fit to the underlying value distribution.

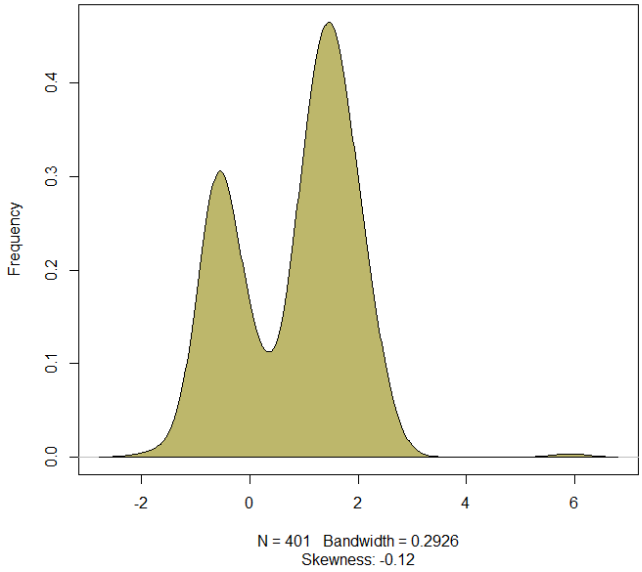

Figure 28: Rudder - Density Plot Means Values

F.5 Elevator Sensor

a. Box Plot

The box plot of the elevator sensor means variable indicates a lower number of outlying values compared to the original data set. The density plot should diagnose normality.

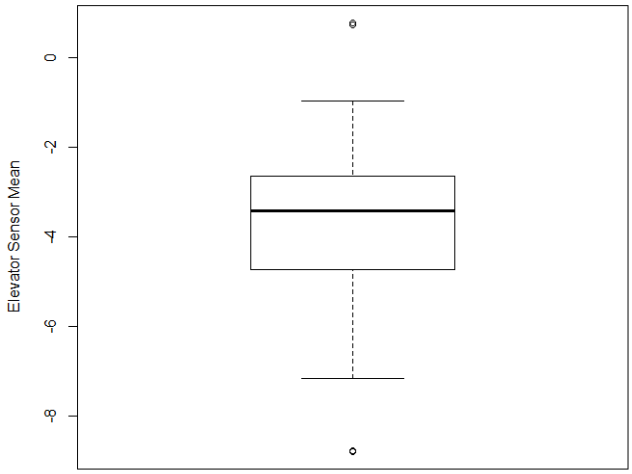

Figure 29: Elevator Sensor - Box Plot Means Values

b. Density Plot

The density plot for the elevator sensor means variable shows a non-normal distribution with two distinct peaks in the data values. Bandwidth indicates a good fit to the underlying variable value distribution.

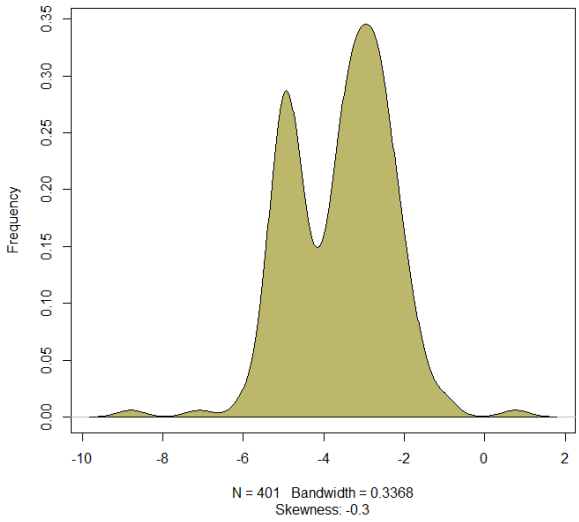

Figure 30: Elevator Sensor - Density Plot Means Values

Appendix G – Experiment 7: Linear Regression Means Data

Regression models using the variables of interest from the means data were developed for three distinct cases. These cases are intended to quantify the relationship between turbulence indicators, density altitude, and engineering burn rate, between density altitude and engineering burn rate, and between turbulence indicators alone and engineering burn rate.

G.1 Turbulence Indicators and Density Altitude

A script was developed in the analysis environment to accept the means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, all four predictor variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate cluster close to the regression line and do not possess a significant amount of spread.

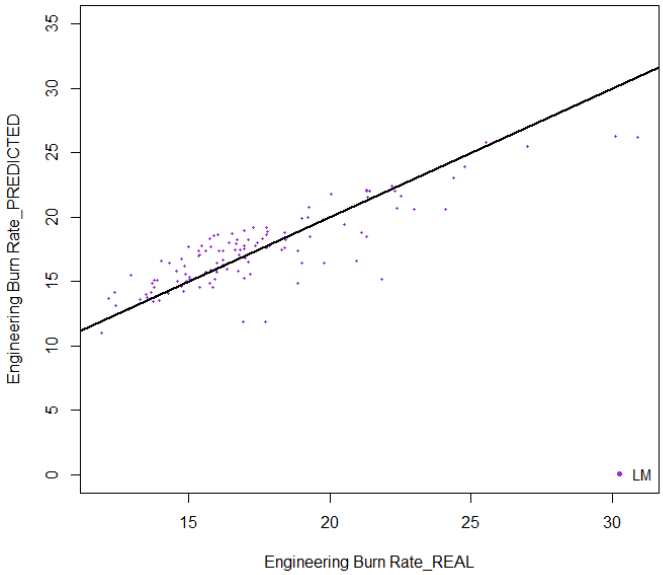

Figure 31: OLS Model – All Predictors Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate cluster close to the regression line and do not possess a significant amount of spread.

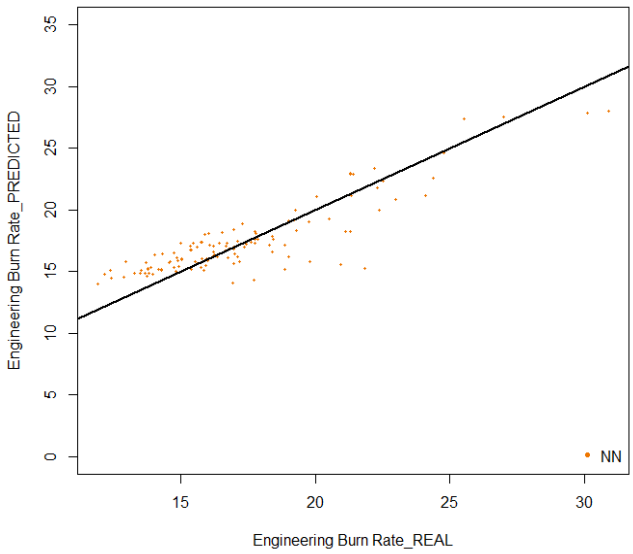

Figure 32: NN Model – All Predictors Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.001814642. The very small p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

G.2 Density Altitude

A script was developed in the analysis environment to accept the analysis data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the density altitude predictor variable was used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line to a great degree, and possess a significant amount of spread, though the range is narrow.

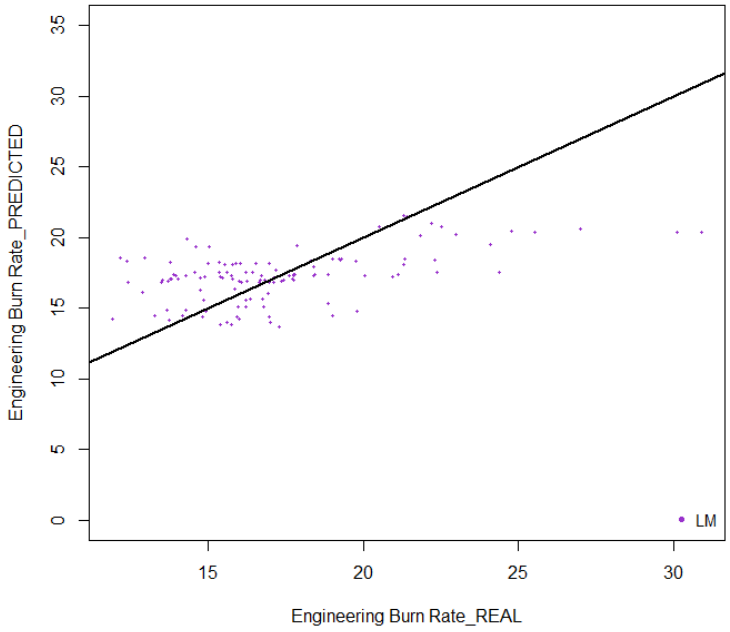

Figure 33: OLS Model – Density Altitude Means

b. Neural Network Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line to a great degree, and possess a significant amount of spread, though the range is narrow.

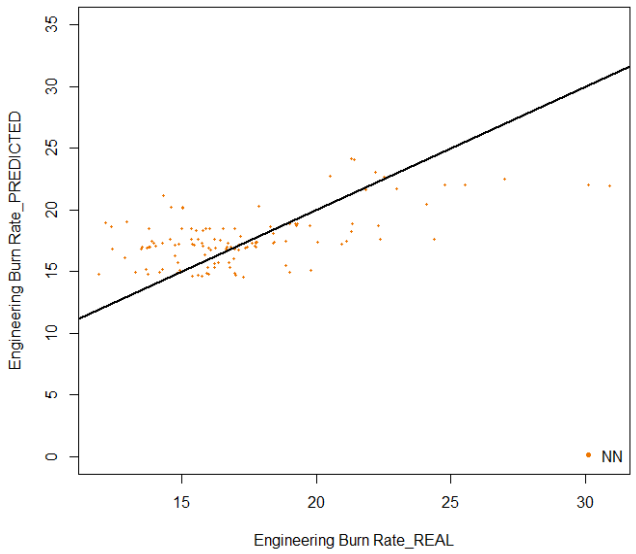

Figure 34: NN Model - Density Altitude Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.0003748491. The small p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

G.3 Turbulence Indicators (Throttle, Rudder, Elevator)

A script was developed in the analysis environment to accept the analysis data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the turbulence indicator variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate cluster near the regression line and do not possess a significant amount of spread.

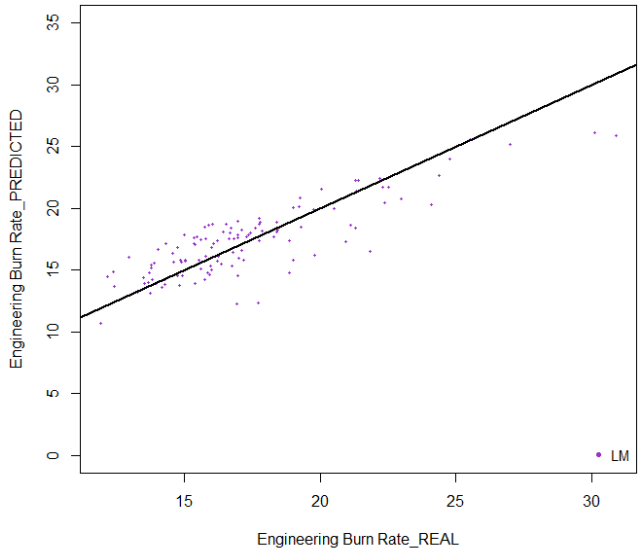

Figure 35: OLS Model – Turbulence Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate cluster near the regression line and do not possess a significant amount of spread.

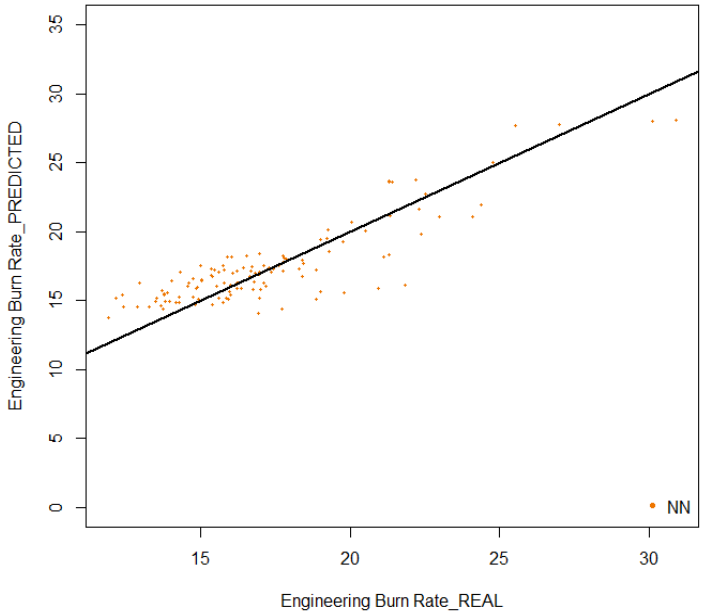

Figure 36: NN Model – Turbulence Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.00600855. The small p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

Appendix H – Experiment 9: Variable Analysis Less Turbulent Data

For each of the variables of interest from the less turbulent means data, a box plot showing outlying values, and a density plot showing the shape of the variable distribution along with skewness and bandwidth were constructed.

H.1 Engineering Burn Rate

a. Box Plot

The box plot for the engineering burn rate mean values indicates a lower number of outlying values compared to the original analysis data. Since outlying values are still present, the distribution is suspected to be non-normal.

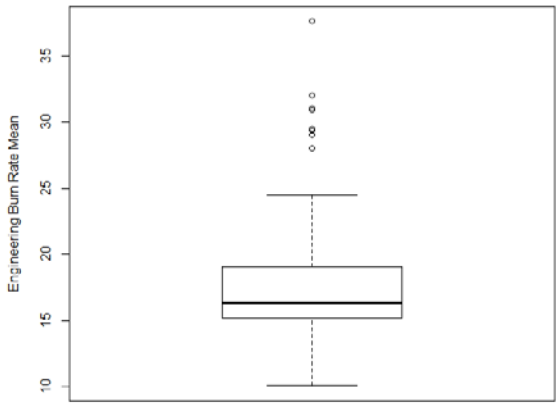

Figure 37: Engineering Burn Rate – Box Plot Less Turbulent Means

b. Density Plot

The density plot for the engineering burn rate means variable shows a non-normal distribution that is skewed positively. Bandwidth indicates a relatively good fit for the underlying value distribution, even though the distribution is non-normal.

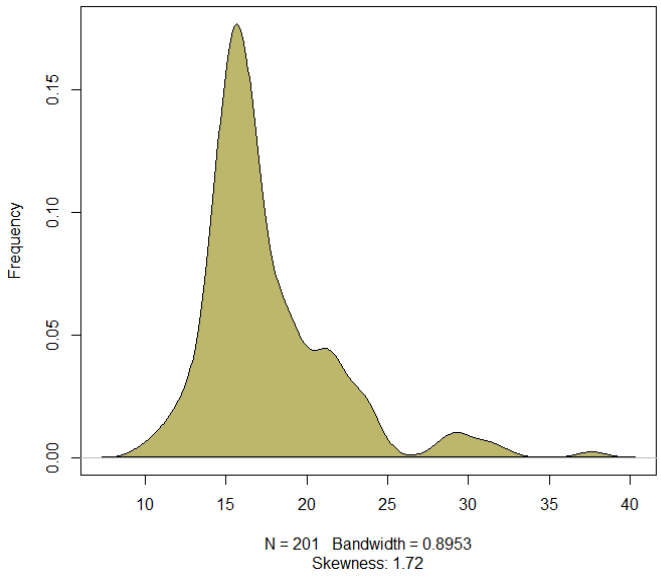

Figure 38: Engineering Burn Rate – Density Plot Less Turbulent Means

H.2 Density Altitude

a. Box Plot

The box plot for the density altitude mean values indicates no outlying values compared to the original analysis data. The density plot should assess normality of data.

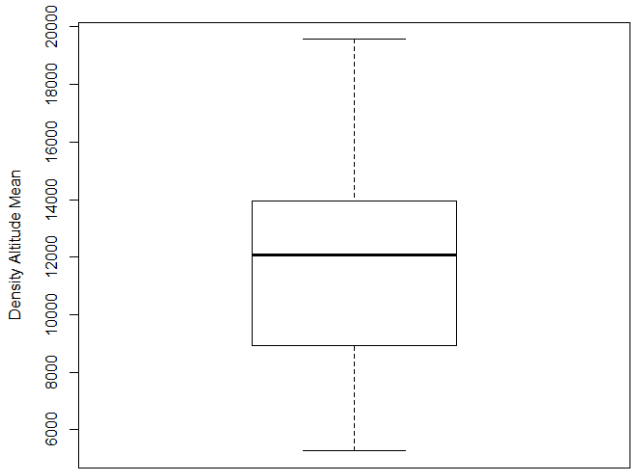

Figure 39: Density Altitude – Box Plot Less Turbulent Means

b. Density Plot

The density plot for the density altitude means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates a poor fit for the underlying non-normal value distribution.

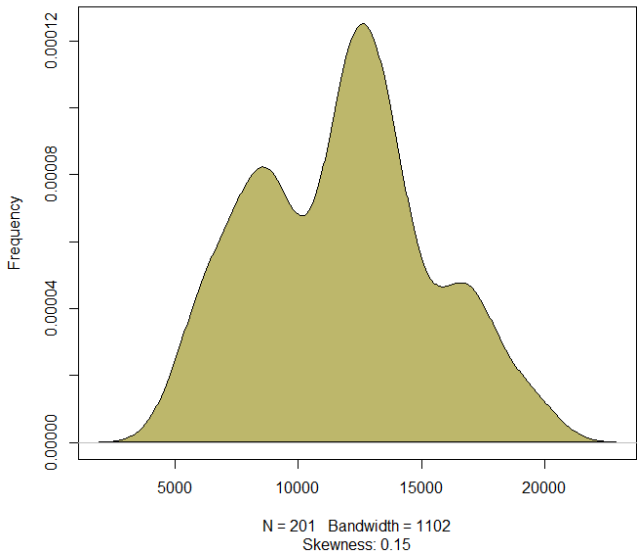

Figure 40: Density Altitude – Density Plot Less Turbulent Means

H.3 Throttle

a. Box Plot

The box plot for the throttle mean values indicates a similar number of outlying values compared to the original means data. Since outlying values are present, the distribution is suspected to be non-normal.

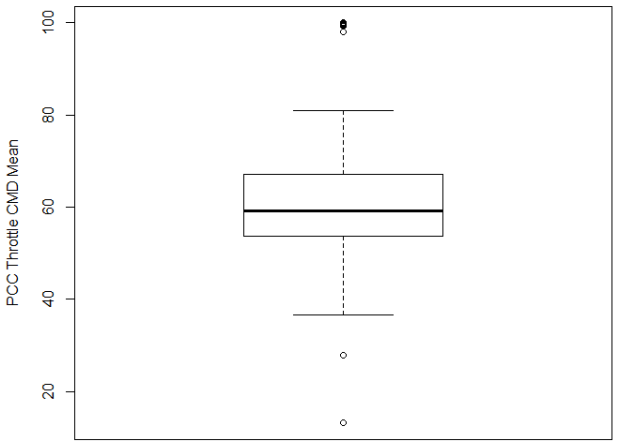

Figure 41: Throttle – Box Plot Less Turbulent Means

b. Density Plot

The density plot for the throttle means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates an acceptable fit for the underlying non-normal value distribution.

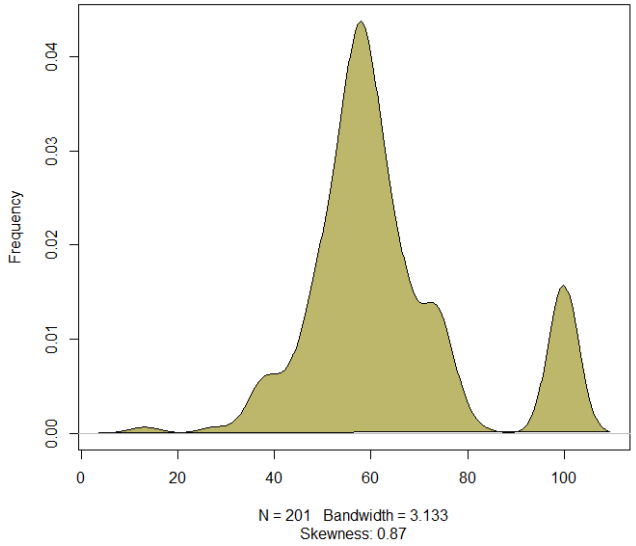

Figure 42: Throttle – Density Plot Less Turbulent Means

H.4 Rudder

a. Box Plot

The box plot for the rudder means variable shows few outlying values. Non-normal distribution of the variable values is suspected, but should be diagnosed by the density plot.

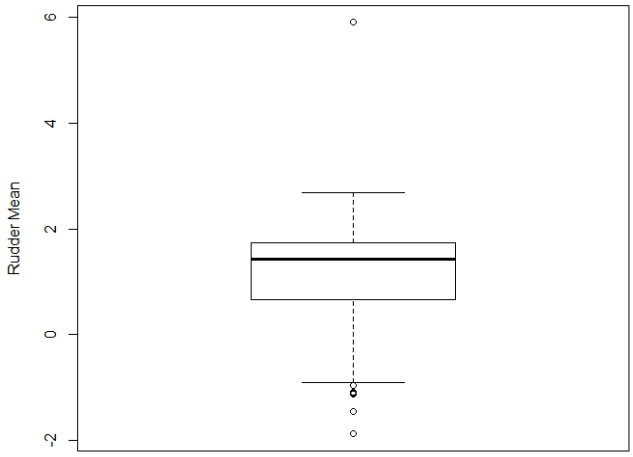

Figure 43: Rudder – Box Plot Less Turbulent Means

b. Density Plot

The density plot of the rudder means variable shows a non-normal distribution with little negative skew. The double peaks suggest the values are clustered around separate means. Bandwidth indicates a good fit to the underlying value distribution.

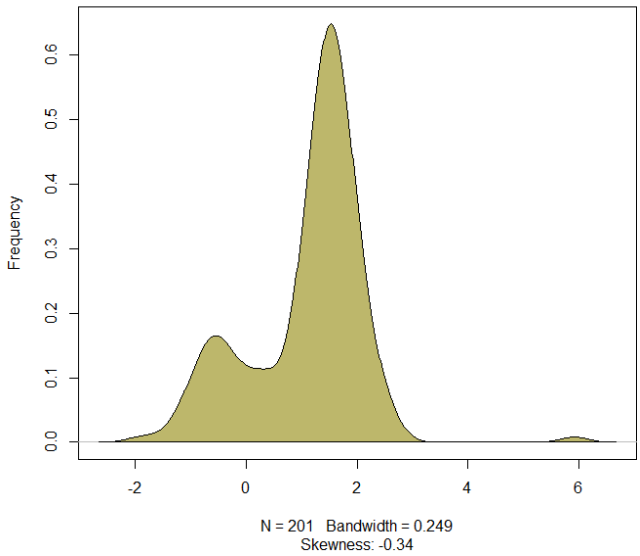

Figure 44: Rudder – Density Plot Less Turbulent Means

H.5 Elevator Sensor

a. Box Plot

The box plot of the elevator sensor means variable indicates a similar number of outlying values compared to the original means data set. The density plot should diagnose normality.

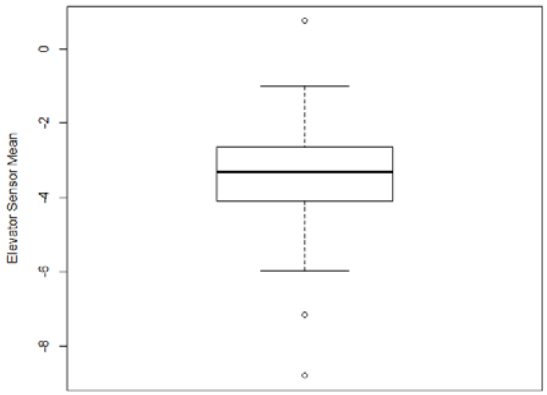

Figure 45: Elevator Sensor – Box Plot Less Turbulent Means

b. Density Plot

The density plot for the elevator sensor means variable shows a non-normal distribution with two peaks in the data values and a negative skew. Bandwidth indicates a good fit to the underlying variable value distribution.

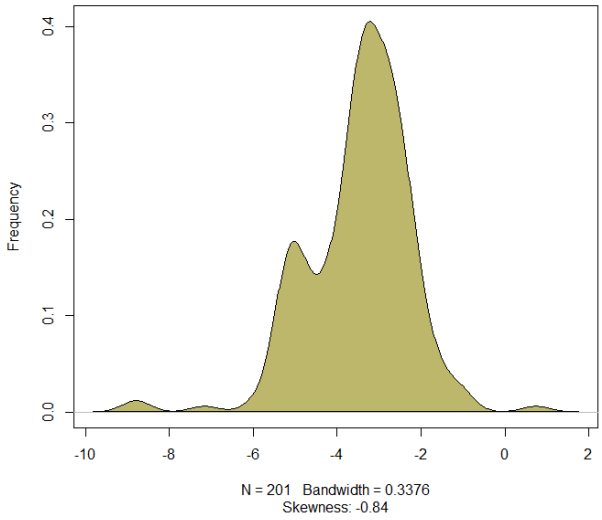

Figure 46: Elevator Sensor – Density Plot Less Turbulent Means

Appendix I – Experiment 10: Linear Regression Less Turbulent Data

Regression models using the variables of interest from the less turbulent means data were developed for three distinct cases. These cases are intended to quantify the relationship between turbulence indicators, density altitude, and engineering burn rate, between density altitude and engineering burn rate, and between turbulence indicators alone and engineering burn rate.

I.1 Turbulence Indicators and Density Altitude

A script was developed in the analysis environment to accept the means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, all four predictor variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate seem to cluster close to the regression line, though outlying values increase the spread.

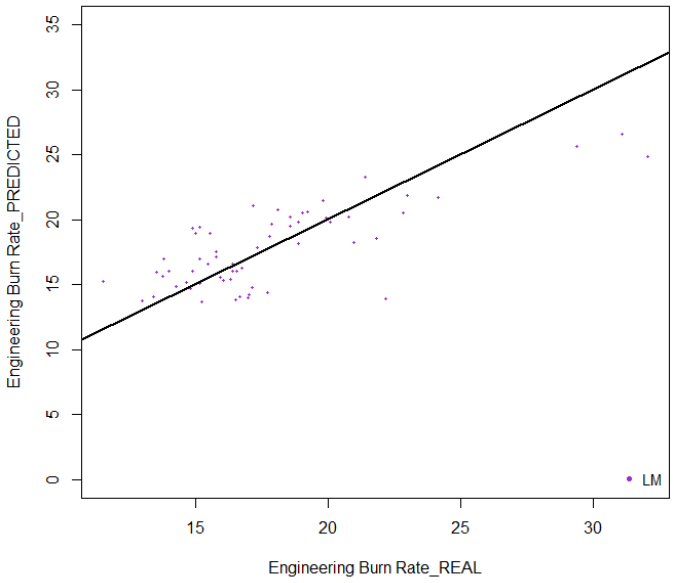

Figure 47: OLS Model – All Predictors Less Turbulent Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate seem to cluster close to the regression line, though outlying values increase the spread.

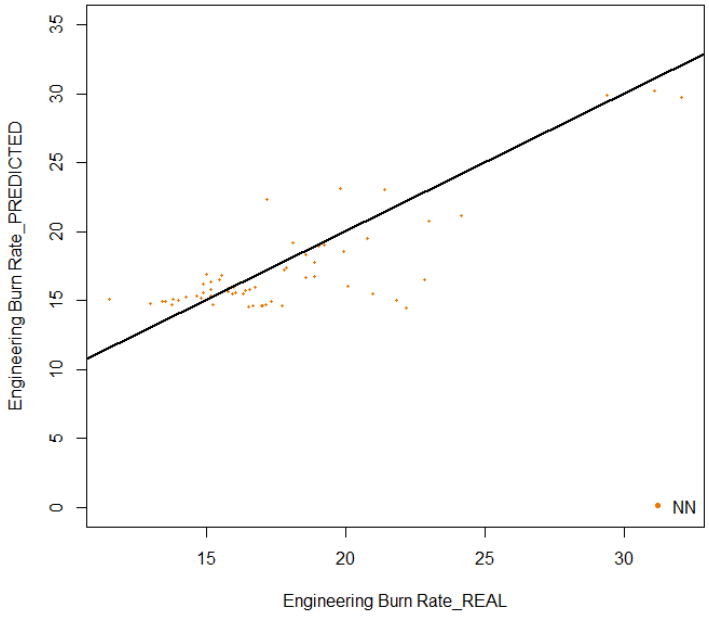

Figure 48: NN Model – All Predictors Less Turbulent Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.05372462. The p-value, which is more than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE for the OLS linear regression model is 6.53. The MSE for the neural network linear regression model is 5.75.

I.2 Density Altitude

A script was developed in the analysis environment to accept the less turbulent means data set, randomly partition the data according to a seventy/thirty split, and

calculate regression models using OLS and neural network methods. In this case, only the density altitude predictor variable was used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line to a great degree, and possess a significant amount of spread, though the range is narrow.

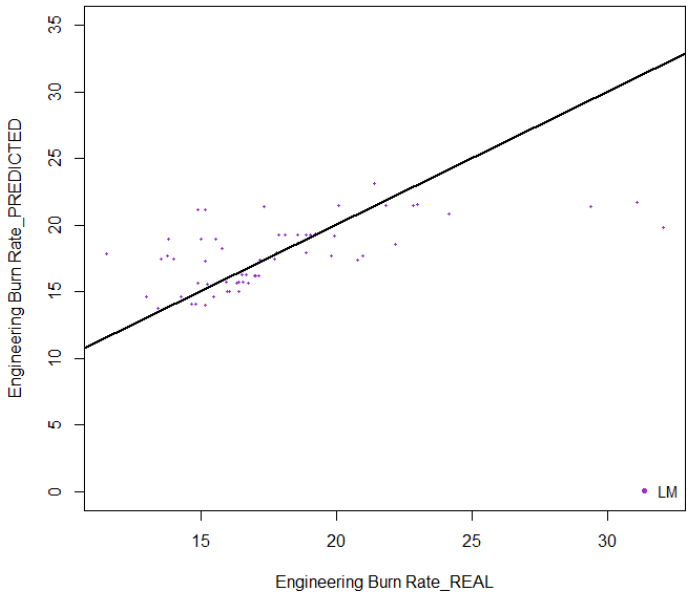

Figure 49: OLS Model – Density Altitude Less Turbulent Means

b. Neural Network Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line to a great degree, and possess a significant amount of spread, though the range is narrow.

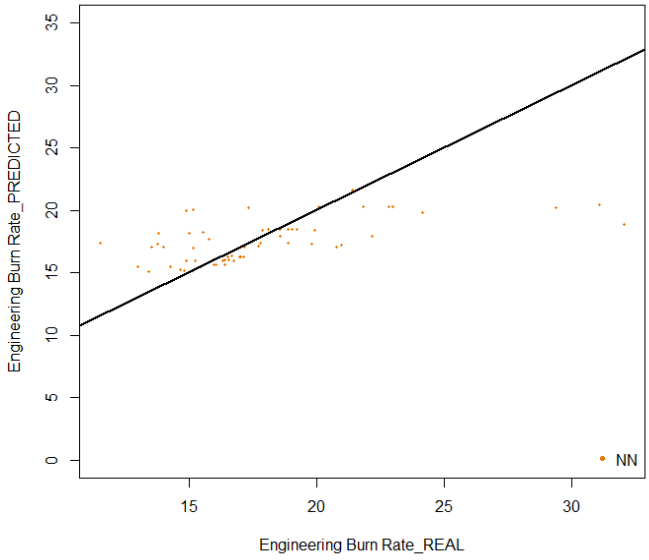

Figure 50: NN Model – Density Altitude Less Turbulent Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.0418. The p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

I.3 Turbulence Indicators (Throttle, Rudder, Elevator)

A script was developed in the analysis environment to accept the less turbulent means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the turbulence indicator variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate seem to cluster near the regression line, though outlying predictions possess a degree of spread.

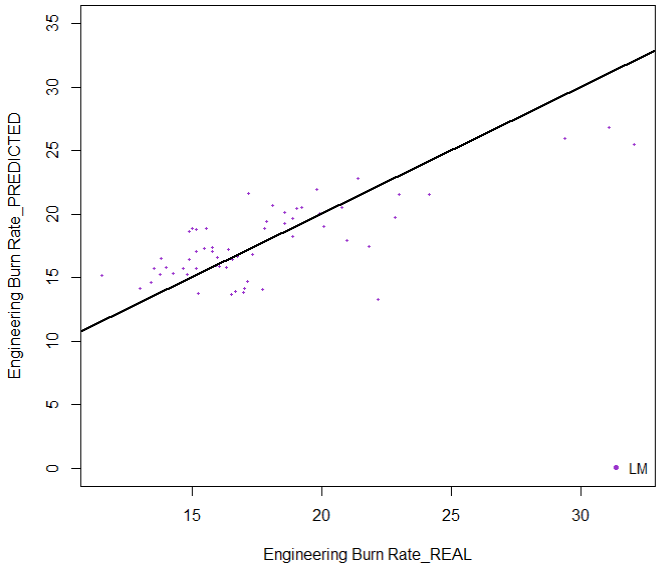

Figure 51: OLS Model – Turbulence Less Turbulent Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate seem to cluster near the regression line, though outlying predictions possess a degree of spread.

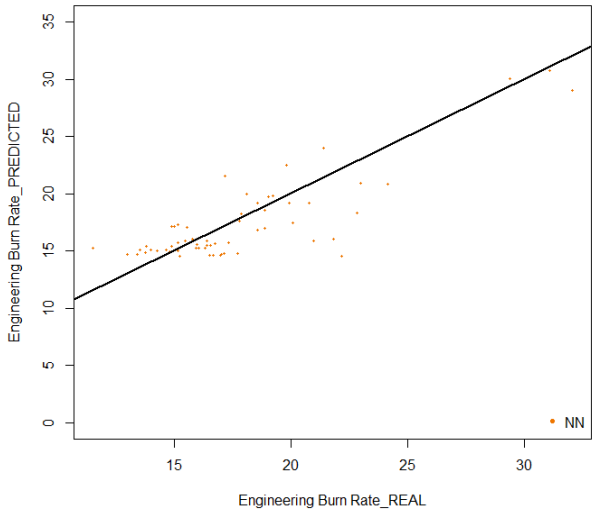

Figure 52: NN Model – Turbulence Less Turbulent Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.1217. The p-value, which is more than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE for the OLS linear regression model is 6.71. The MSE for the neural network linear regression model is 5.04.

Appendix J – Experiment 11: Variable Analysis More Turbulent Data

For each of the variables of interest from the more turbulent means data, a box plot showing outlying values, and a density plot showing the shape of the variable distribution along with skewness and bandwidth were constructed.

J.1 Engineering Burn Rate

a. Box Plot

The box plot for the engineering burn rate mean values indicates a higher number of outlying values compared to the original means data. Since outlying values are present, the distribution is suspected to be non-normal.

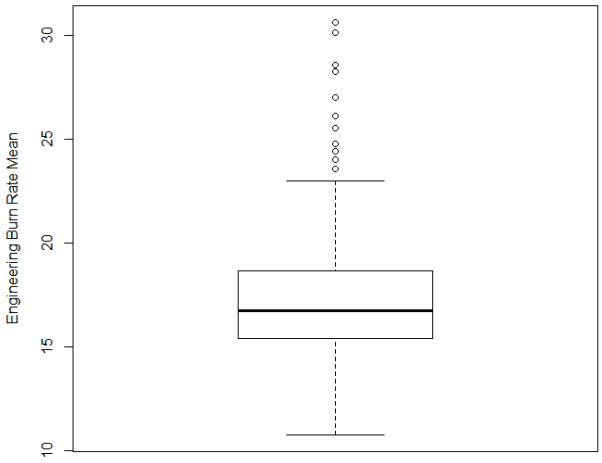

Figure 53: Engineering Burn Rate – Box Plot More Turbulent Means

b. Density Plot

The density plot for the engineering burn rate means variable shows a non-normal distribution that is skewed positively. Bandwidth indicates an acceptable fit for the underlying value distribution, even though the distribution is non-normal.

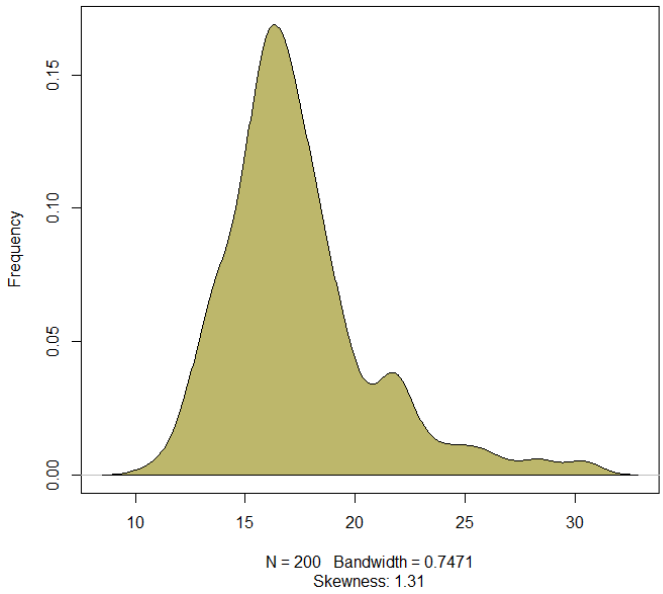

Figure 54: Engineering Burn Rate – Density Plot More Turbulent Means

J.2 Density Altitude

a. Box Plot

The box plot for the density altitude mean values shows a number of outlying values. The density plot should assess normality of data.

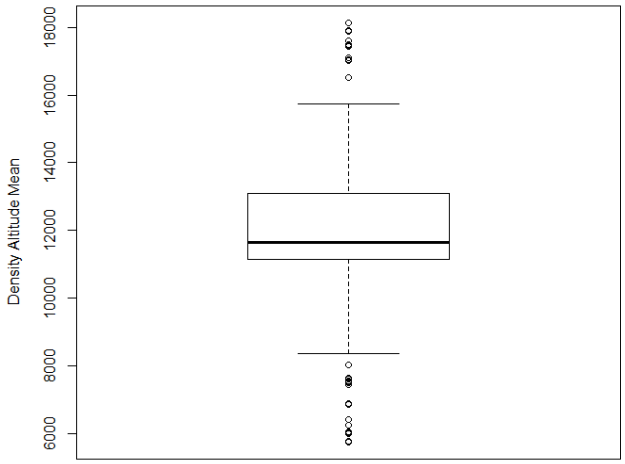

Figure 55: Density Altitude – Box Plot More Turbulent Means

b. Density Plot

The density plot for the density altitude means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates a poor fit for the underlying non-normal value distribution.

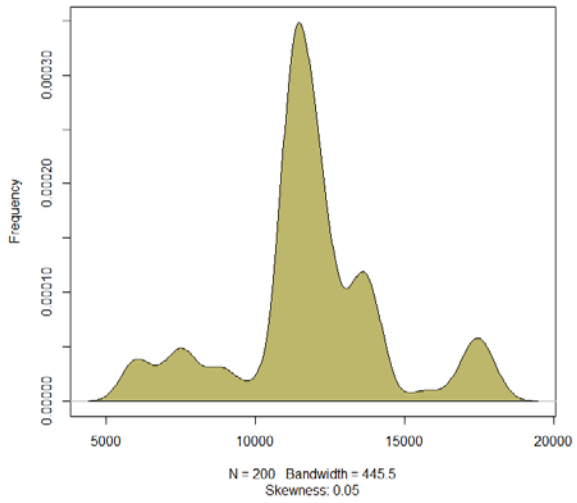

Figure 56: Density Altitude – Density Plot More Turbulent Means

J.3 Throttle

a. Box Plot

The box plot for the throttle mean values indicates few outlying values. Since outlying values are present, the distribution is suspected to be non-normal.

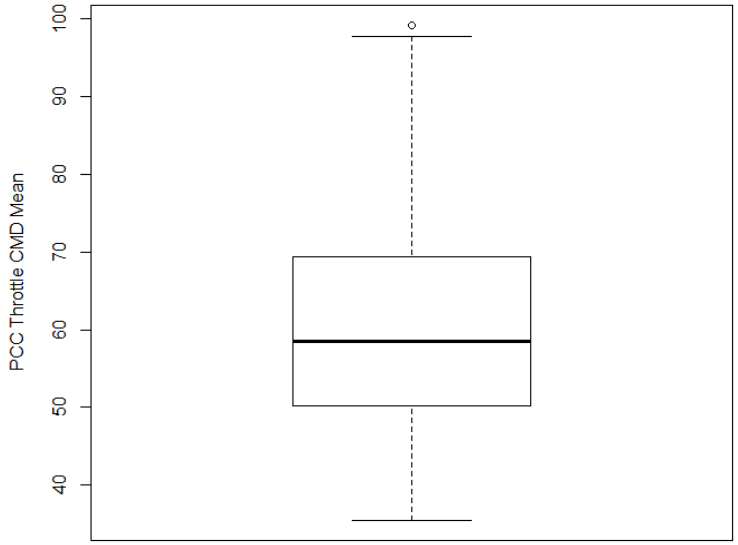

Figure 57: Throttle – Box Plot More Turbulent Means

b. Density Plot

The density plot for the throttle means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates an acceptable fit for the underlying non-normal value distribution.

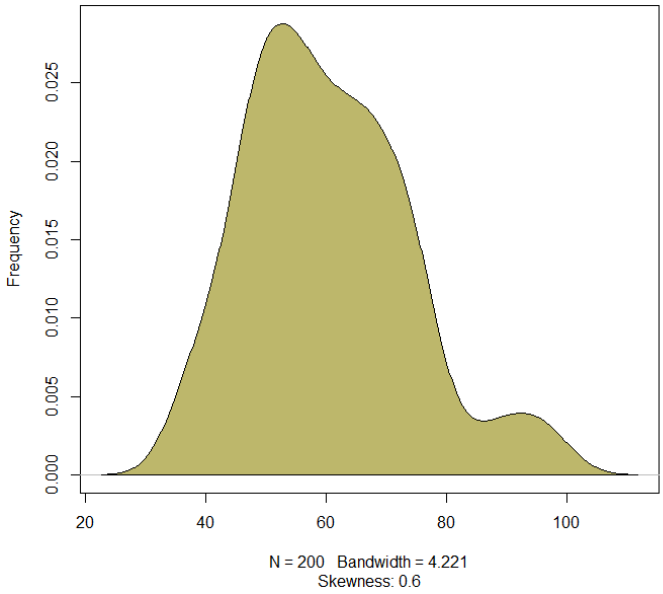

Figure 58: Throttle – Density Plot More Turbulent Means

J.4 Rudder

a. Box Plot

The box plot for the rudder means variable shows no outlying values. The density plot should diagnose non-normal distribution.

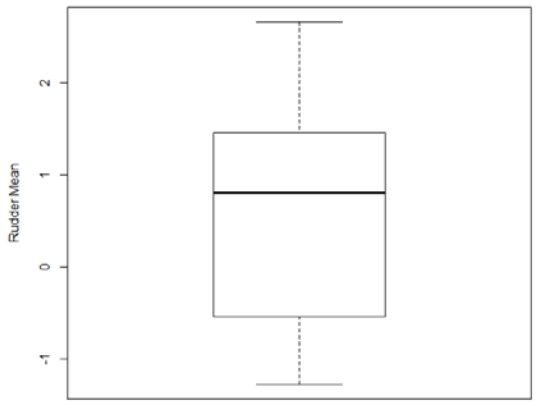

Figure 59: Rudder – Box Plot More Turbulent Means

b. Density Plot

The density plot of the rudder means variable shows a non-normal distribution with little positive skew. The double peaks suggest the values are clustered around separate means. Bandwidth indicates a good fit to the underlying value distribution.

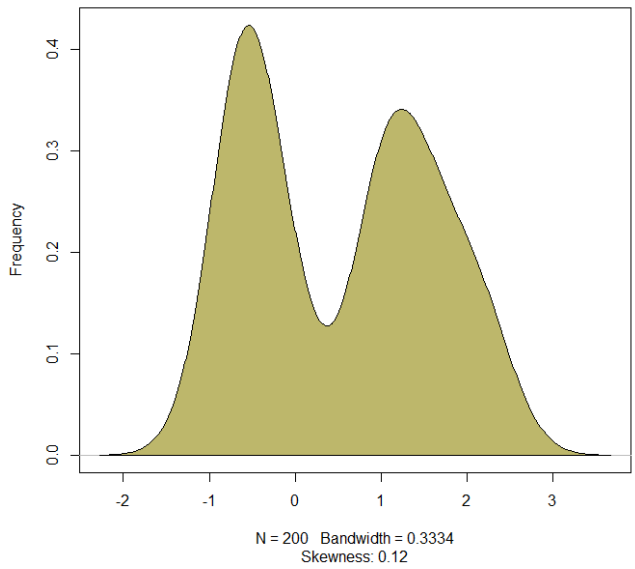

Figure 60: Rudder – Density Plot More Turbulent Means

J.5 Elevator Sensor

a. Box Plot

The box plot of the elevator sensor means variable indicates few outlying values. Normality should be diagnosed by the density plot.

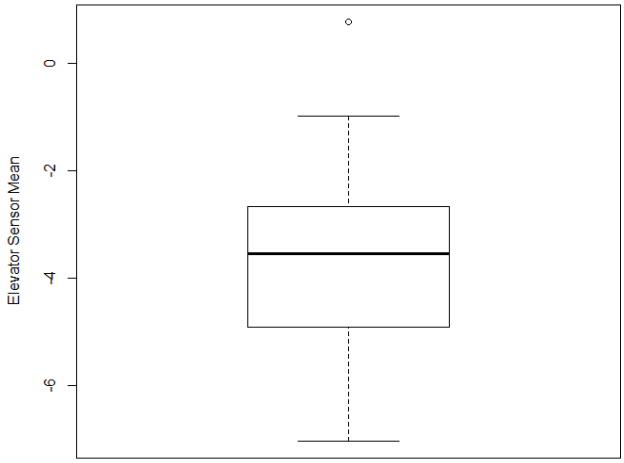

Figure 61: Elevator Sensor – Box Plot More Turbulent Means

b. Density Plot

The density plot for the elevator sensor means variable shows a non-normal distribution with two peaks in the data values and a slight positive skew. Bandwidth indicates a good fit to the underlying variable value distribution.

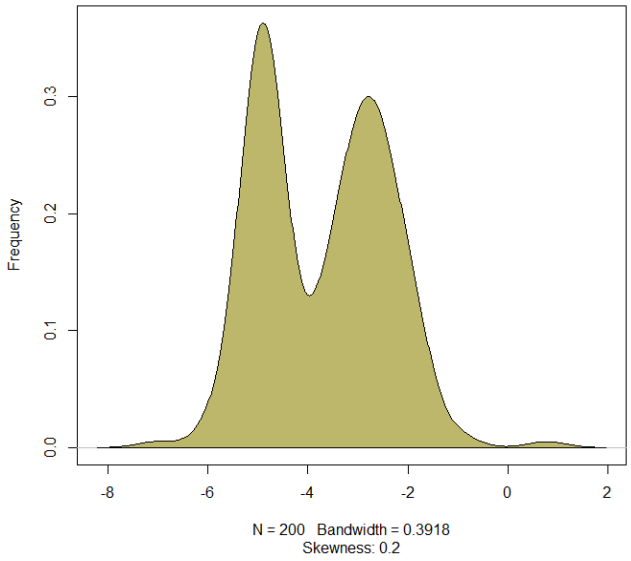

Figure 62: Elevator Sensor – Density Plot More Turbulent Means

Appendix K – Experiment 12: Linear Regression More Turbulent Data

Regression models using the variables of interest from the more turbulent means data were developed for three distinct cases. These cases are intended to quantify the relationship between turbulence indicators, density altitude, and engineering burn rate, between density altitude and engineering burn rate, and between turbulence indicators alone and engineering burn rate.

K.2 Turbulence Indicators and Density Altitude

A script was developed in the analysis environment to accept the more turbulent means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, all four predictor variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate seem to cluster close to the regression line.

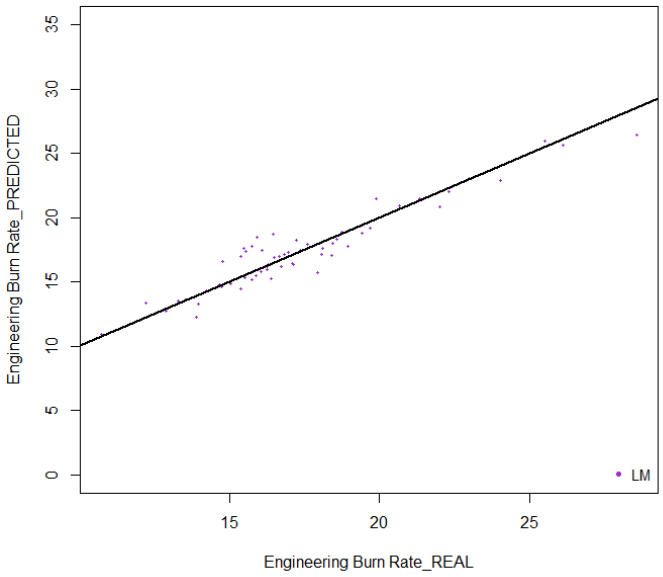

Figure 63: OLS Model – All Predictors More Turbulent Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate seem to cluster close to the regression line.

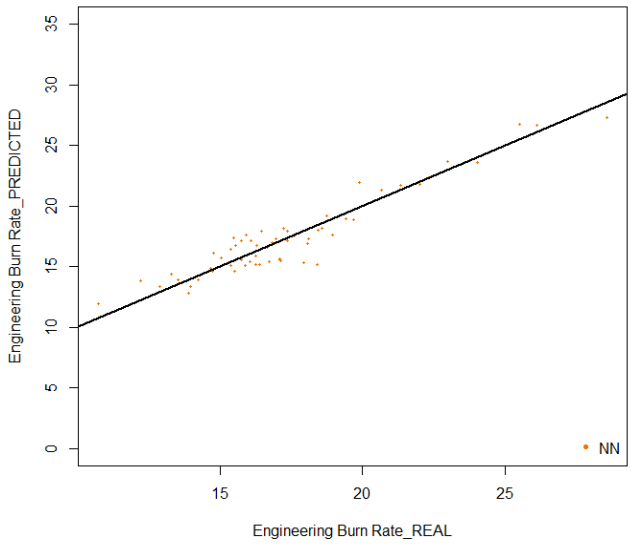

Figure 64: NN Model – All Predictors More Turbulent Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.1853. The p-value, which is more than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE for the OLS linear regression model is 1.05. The MSE for the neural network linear regression model is 1.77.

K.2 Density Altitude

A script was developed in the analysis environment to accept the more turbulent means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the density altitude predictor variable was used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line to a great degree, and possess a significant amount of spread, though the range is narrow.

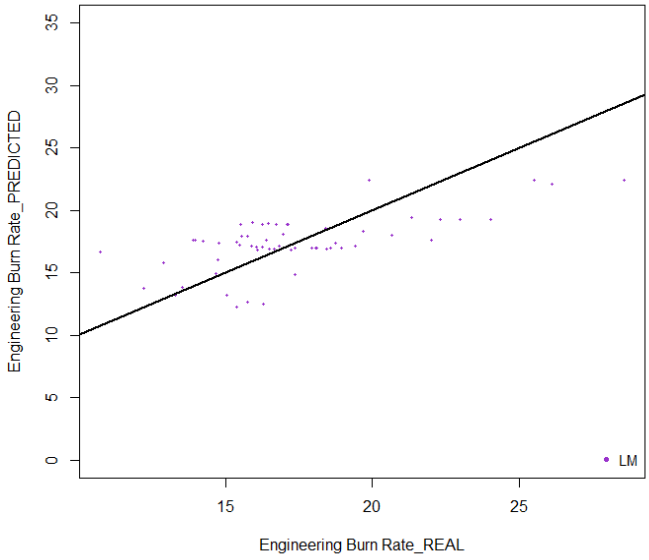

Figure 65: OLS Model – Density Altitude More Turbulent Means

b. Neural Network Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line to a great degree, and possess a significant amount of spread, though the range is narrow.

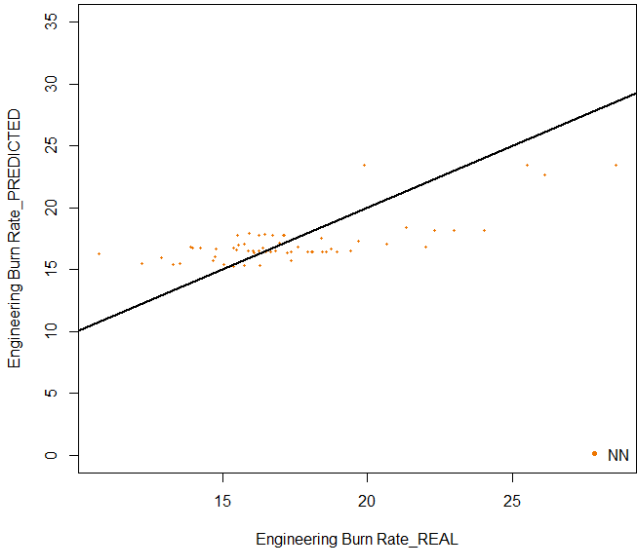

Figure 66: NN Model – Density Altitude More Turbulent Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.0099. The p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

K.3 Turbulence Indicators (Throttle, Rudder, Elevator)

A script was developed in the analysis environment to accept the more turbulent means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the turbulence indicator variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate seem to cluster near the regression line.

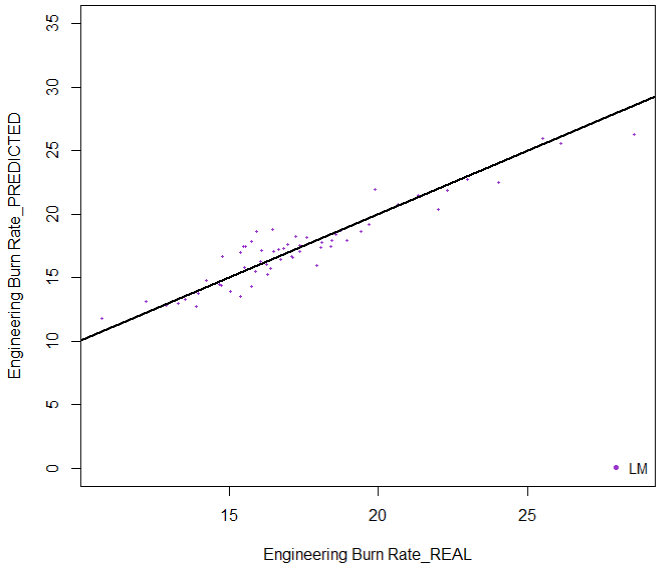

Figure 67: OLS Model – Turbulence More Turbulent Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate seem to cluster near the regression line.

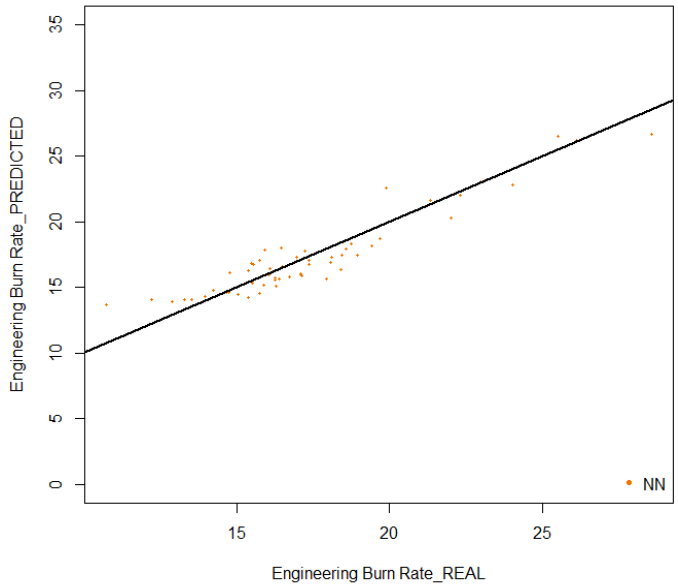

Figure 68: NN Model – Turbulence More Turbulent Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.0257. The p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should be rejected. The MSE for both regression models will not be calculated, as the figures would not be accurate.

Appendix L – Experiment 13: Variable Analysis Higher Density Altitude Data

For each of the variables of interest from the higher density altitude means data, a box plot showing outlying values, and a density plot showing the shape of the variable distribution along with skewness and bandwidth were constructed.

L.1 Engineering Burn Rate

a. Box Plot

The box plot for the engineering burn rate mean values shows some outlying values. Since outlying values are present, the distribution is suspected to be non-normal.

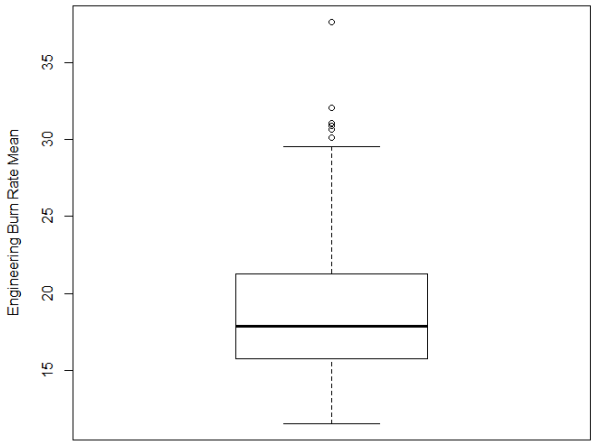

Figure 69: Engineering Burn Rate – Box Plot Higher Density Altitude Means

b. Density Plot

The density plot for the engineering burn rate means variable shows a non-normal distribution that is skewed positively. Bandwidth indicates an acceptable fit for the underlying value distribution, even though the distribution is non-normal.

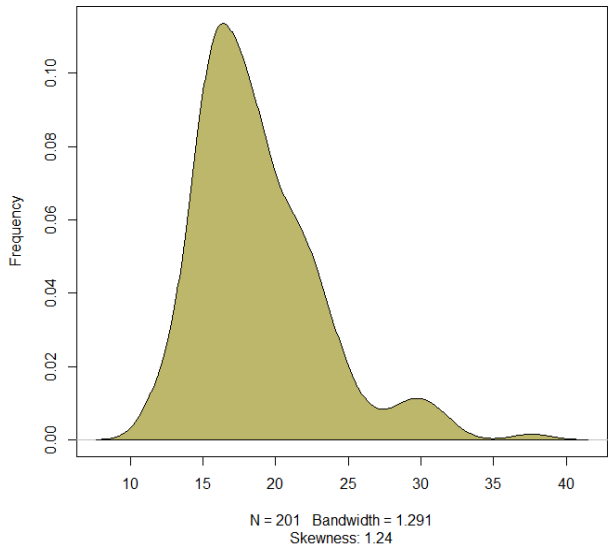

Figure 70: Engineering Burn Rate – Density Plot Higher Density Altitude Means

L.2 Density Altitude

a. Box Plot

The box plot for the density altitude mean values shows no outlying values. The density plot should assess normality of data.

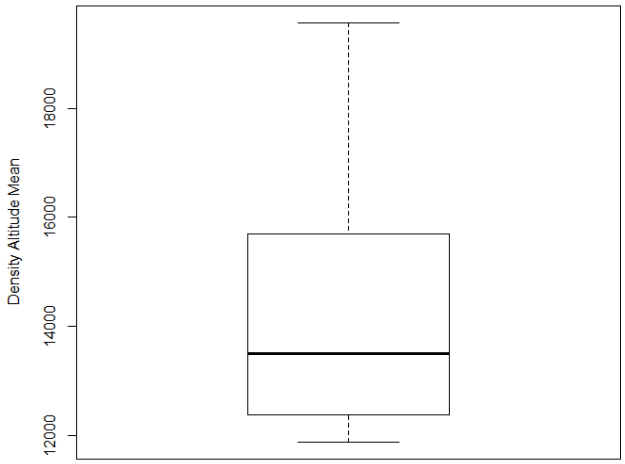

Figure 71: Density Altitude – Box Plot Higher Density Altitude Means

b. Density Plot

The density plot for the density altitude means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates a poor fit for the underlying non-normal value distribution.

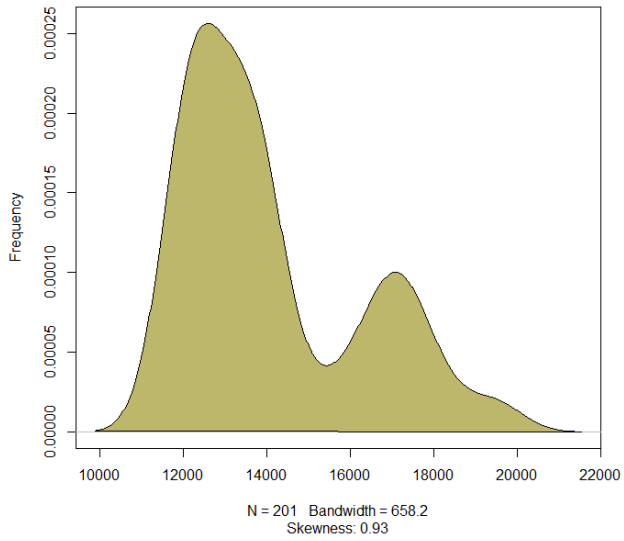

Figure 72: Density Altitude – Density Plot Higher Density Altitude Means

L.3 Throttle

a. Box Plot

The box plot for the throttle mean values indicates no outlying values. The density plot should assess normality.

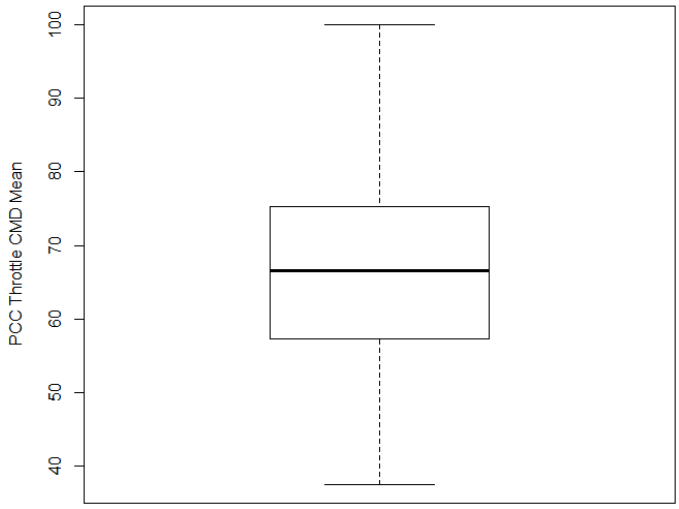

Figure 73: Throttle – Box Plot Higher Density Altitude Means

b. Density Plot

The density plot for the throttle means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates an acceptable fit for the underlying non-normal value distribution.

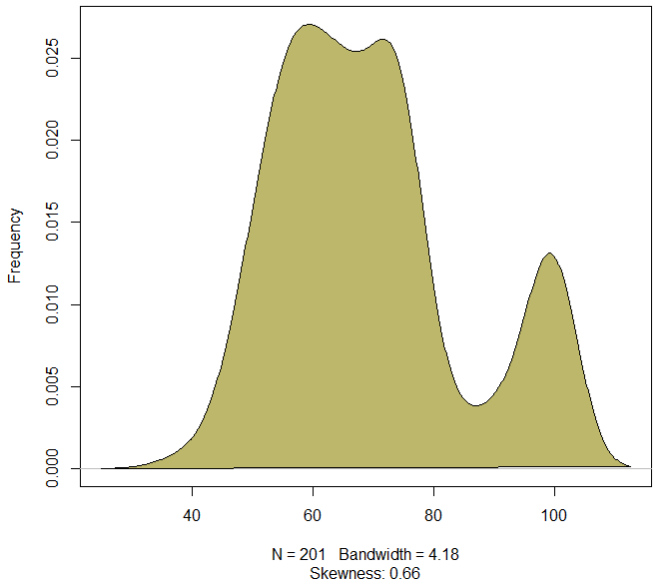

Figure 74: Throttle – Density Plot Higher Density Altitude Means

L.4 Rudder

a. Box Plot

The box plot for the rudder means variable shows no outlying values. The density plot should diagnose non-normal distribution.

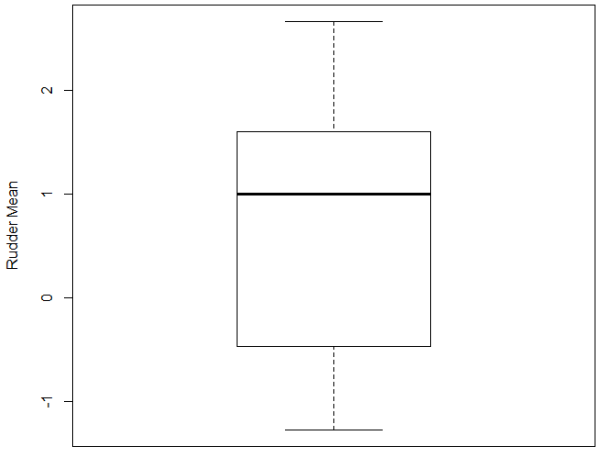

Figure 75: Rudder – Box Plot Higher Density Altitude Means

b. Density Plot

The density plot of the rudder means variable shows a non-normal distribution with little negative skew. The double peaks suggest the values are clustered around separate means. Bandwidth indicates a good fit to the underlying value distribution.

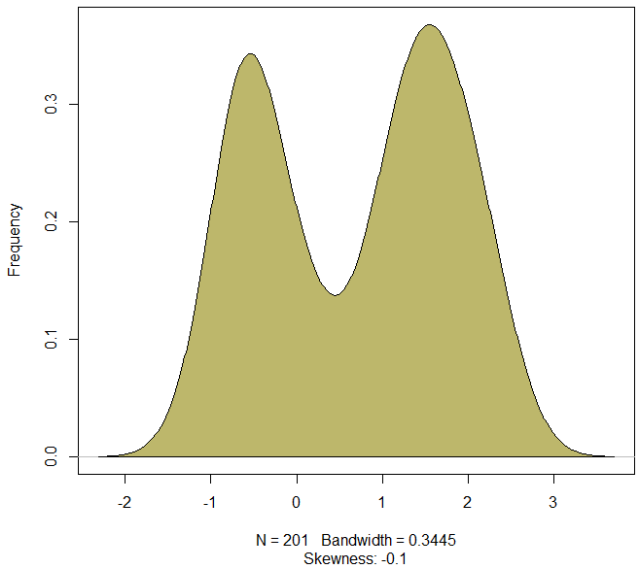

Figure 76: Rudder – Density Plot Higher Density Altitude Means

L.5 Elevator Sensor

a. Box Plot

The box plot of the elevator sensor means variable indicates few outlying values. The density plot should diagnose normality.

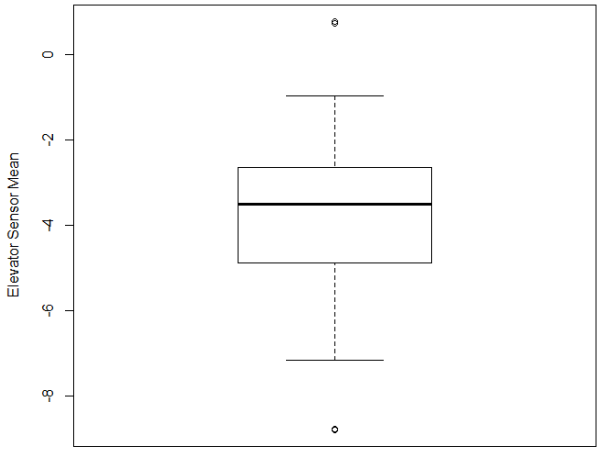

Figure 77: Elevator Sensor – Box Plot Higher Density Altitude Means

b. Density Plot

The density plot for the elevator sensor means variable shows a non-normal distribution with two peaks in the data values and a slight negative skew. Bandwidth indicates a good fit to the underlying variable value distribution.

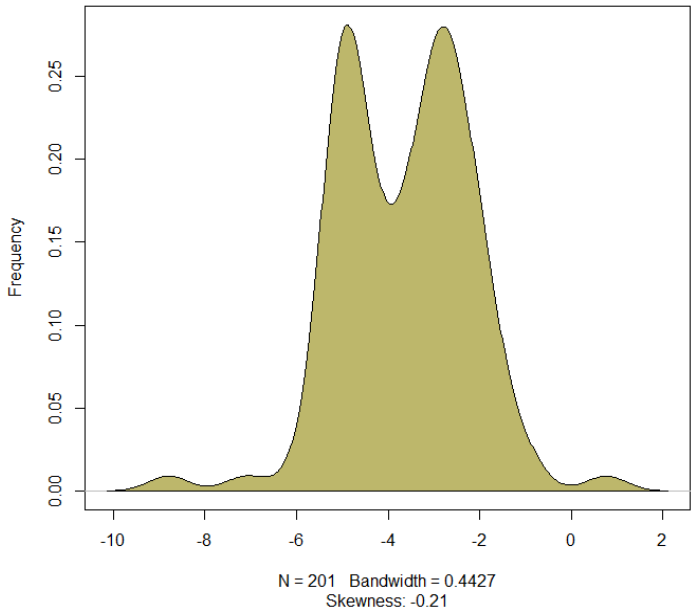

Figure 78: Elevator Sensor – Density Plot Higher Density Altitude Means

Appendix M – Experiment 14: Linear Regression Higher Density Altitude Data

Regression models using the variables of interest from the higher density altitude means data were developed for three distinct cases. These cases are intended to quantify the relationship between turbulence indicators, density altitude, and engineering burn rate, between density altitude and engineering burn rate, and between turbulence indicators alone and engineering burn rate.

M.1 Turbulence Indicators and Density Altitude

A script was developed in the analysis environment to accept the lower density altitude means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, all four predictor variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not seem to cluster close to the regression line.

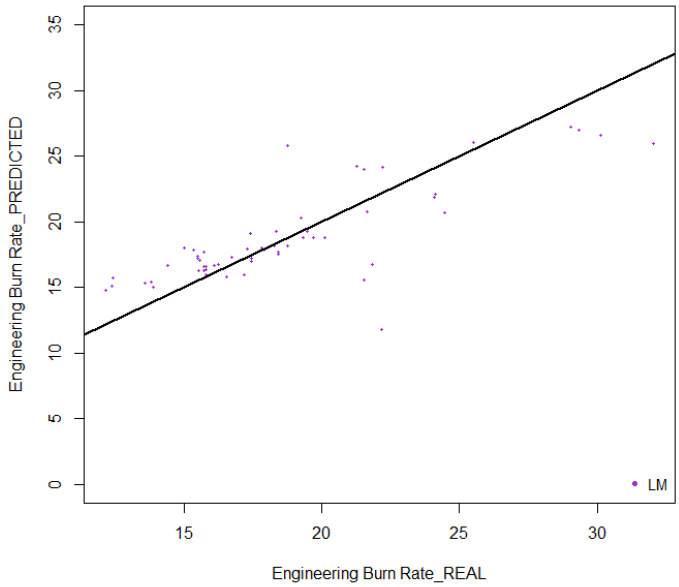

Figure 79: OLS Model – All Predictors Higher Density Altitude Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate do not seem to cluster close to the regression line.

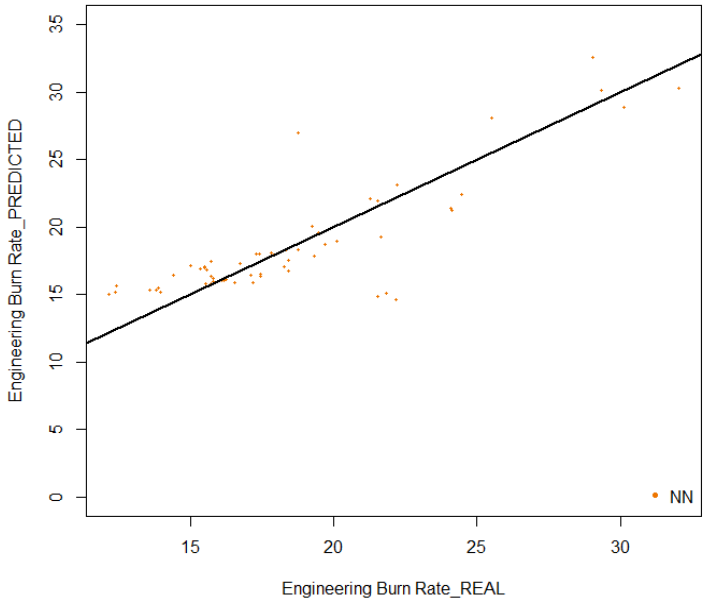

Figure 80: NN Model – All Predictors Higher Density Altitude Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.7333. The p-value, which is more than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE for the OLS linear regression model is 6.65. The MSE for the neural network linear regression model is 5.66.

M.2 Density Altitude

A script was developed in the analysis environment to accept the higher density altitude means data set, randomly partition the data according to a seventy/thirty split,

and calculate regression models using OLS and neural network methods. In this case, only the density altitude predictor variable was used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line, and possess a significant amount of spread.

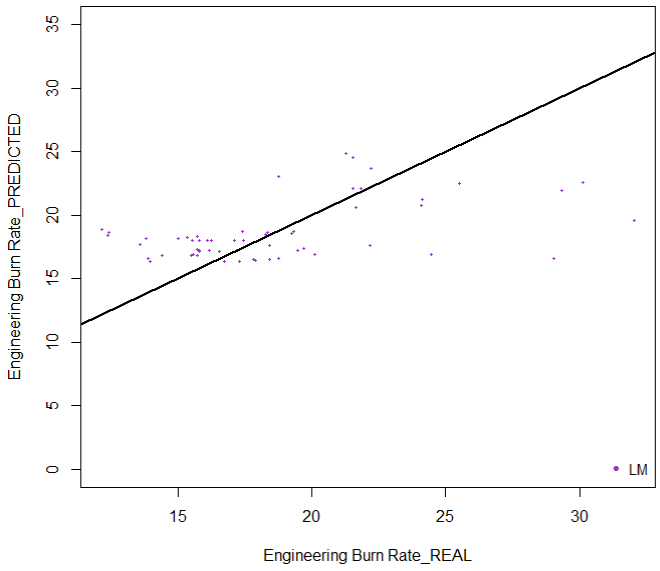

Figure 81: OLS Model – Density Altitude Higher Density Altitude Means

b. Neural Network Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line, and possess a significant amount of spread.

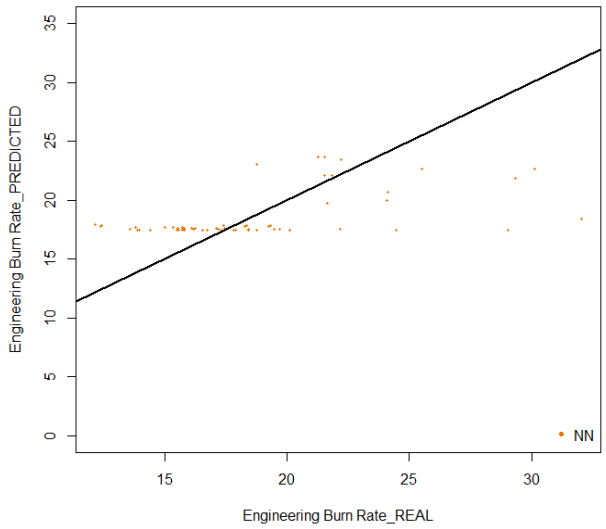

Figure 82: NN Model – Density Altitude Higher Density Altitude Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.0226. The p-value, which is less than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) can be rejected. Since the presence of heteroscedasticity is confirmed, the MSE of both the OLS and Neural Network models is not calculated. The calculated MSE would not be accurate for the majority of the data.

M.3 Turbulence Indicators (Throttle, Rudder, Elevator)

A script was developed in the analysis environment to accept the higher density altitude means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the turbulence indicator variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate seem to cluster near the regression line, though outlying values increase the spread.

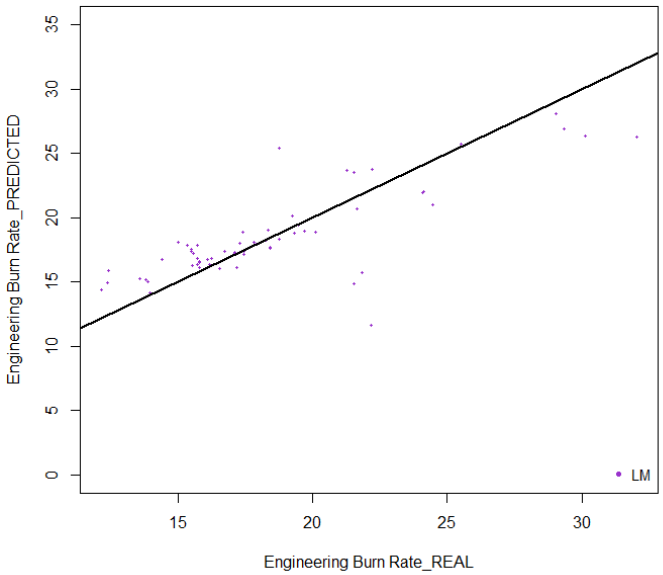

Figure 83: OLS Model – Turbulence Higher Density Altitude Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate seem to cluster near the regression line, though outlying values increase the spread.

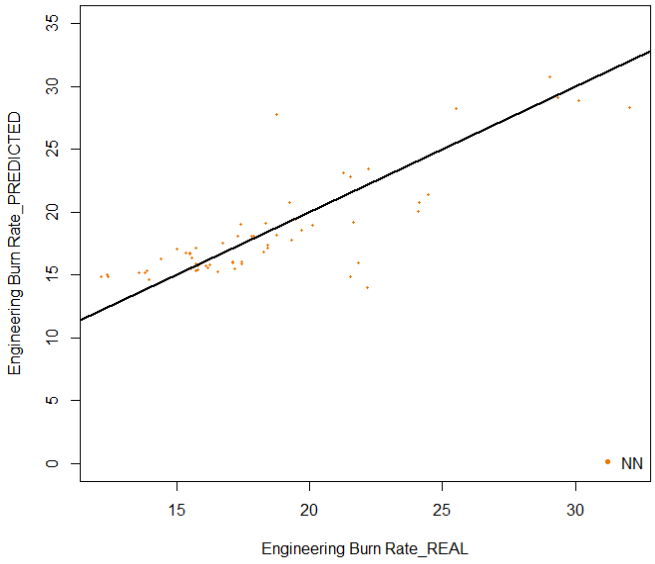

Figure 84: NN Model – Turbulence Higher Density Altitude Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.8330. The p-value, which is more the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE of the OLS linear regression model is 6.73. The MSE of the neural network model is 6.24.

Appendix N – Experiment 15: Variable Analysis Lower Density Altitude Data

For each of the variables of interest from the lower density altitude means data, a box plot showing outlying values, and a density plot showing the shape of the variable distribution along with skewness and bandwidth were constructed.

N.1 Engineering Burn Rate

a. Box Plot

The box plot for the engineering burn rate mean values indicates the presence of outlying values. Since outlying values are present, the distribution is suspected to be non-normal.

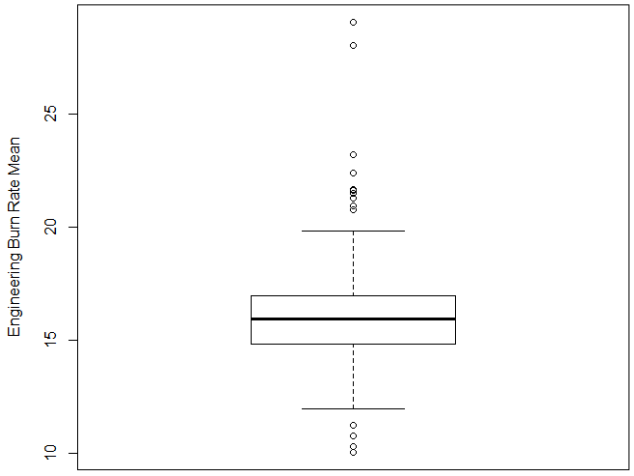

Figure 85: Engineering Burn Rate – Box Plot Lower Density Altitude Means

b. Density Plot

The density plot for the engineering burn rate means variable shows a non-normal distribution that is skewed positively. Bandwidth indicates an acceptable fit for the underlying value distribution, even though the distribution is non-normal.

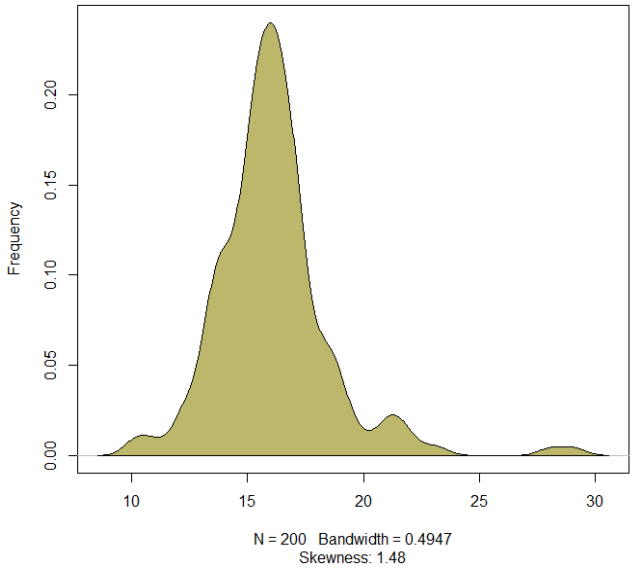

Figure 86: Engineering Burn Rate – Density Plot Lower Density Altitude Means

N.2 Density Altitude

a. Box Plot

The box plot for the density altitude mean values shows no outlying values. The density plot should assess normality of data.

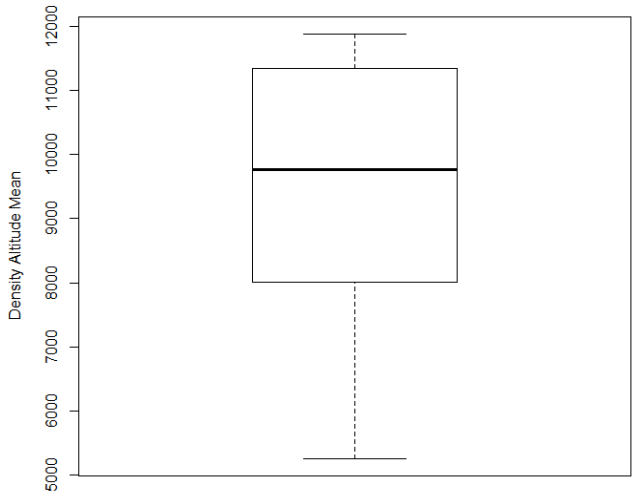

Figure 87: Density Altitude – Box Plot Lower Density Altitude Means

b. Density Plot

The density plot for the density altitude means variable shows a non-normal distribution that has a slight negative skew. Bandwidth indicates a poor fit for the underlying non-normal value distribution.

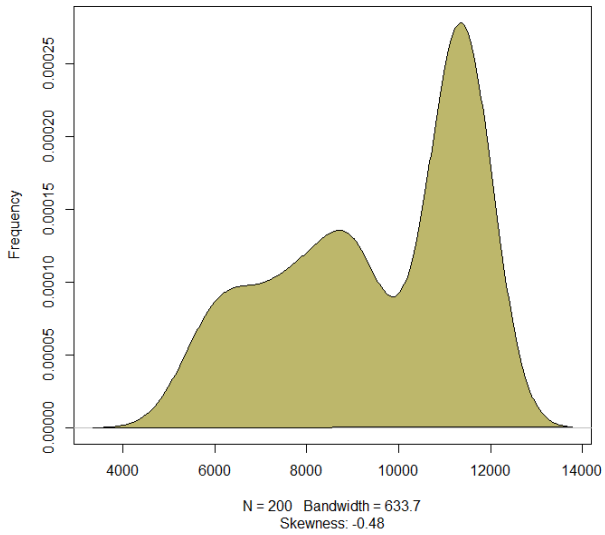

Figure 88: Density Altitude – Density Plot Lower Density Altitude Means

N.3 Throttle

a. Box Plot

The box plot for the throttle mean values indicates few outlying values. Since outlying values are present, the distribution is suspected to be non-normal.

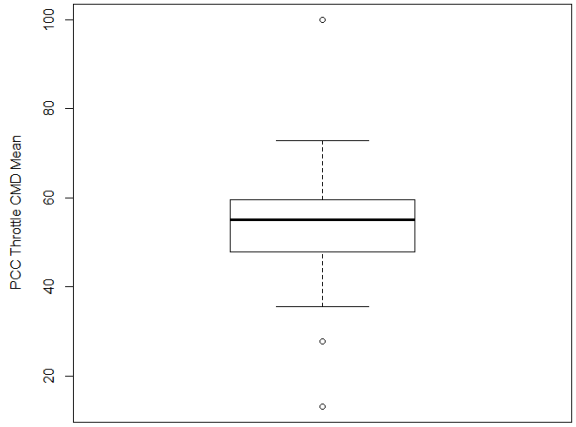

Figure 89: Throttle – Box Plot Lower Density Altitude Means

b. Density Plot

The density plot for the throttle means variable shows a non-normal distribution that has a slight positive skew. Bandwidth indicates an acceptable fit for the underlying non-normal value distribution.

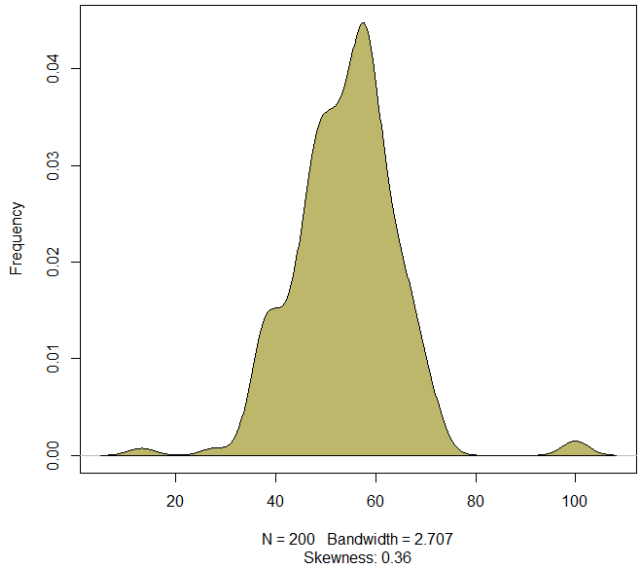

Figure 90: Throttle – Density Plot Lower Density Altitude Means

N.4 Rudder

a. Box Plot

The box plot for the rudder means variable shows few outlying values. The density plot should diagnose non-normal distribution.

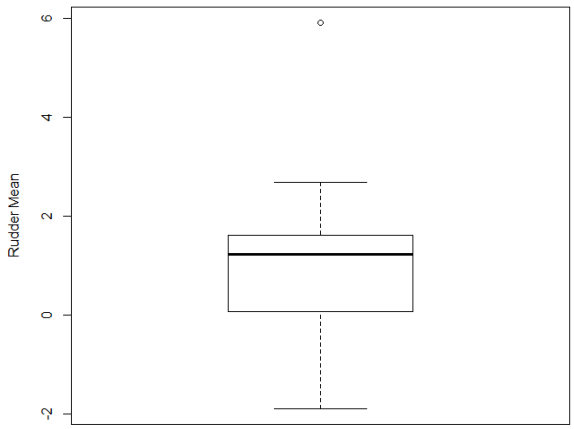

Figure 91: Rudder – Box Plot Lower Density Altitude Means

b. Density Plot

The density plot of the rudder means variable shows a non-normal distribution with little negative skew. The double peaks suggest the values are clustered around separate means. Bandwidth indicates a good fit to the underlying value distribution.

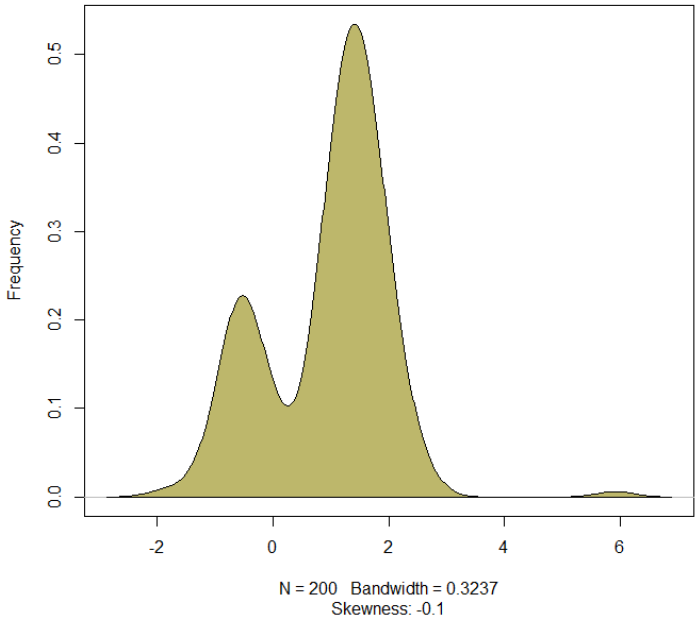

Figure 92: Rudder – Density Plot Lower Density Altitude Means

N.5 Elevator Sensor

a. Box Plot

The box plot of the elevator sensor means variable indicates no outlying values. The density plot should diagnose normality.

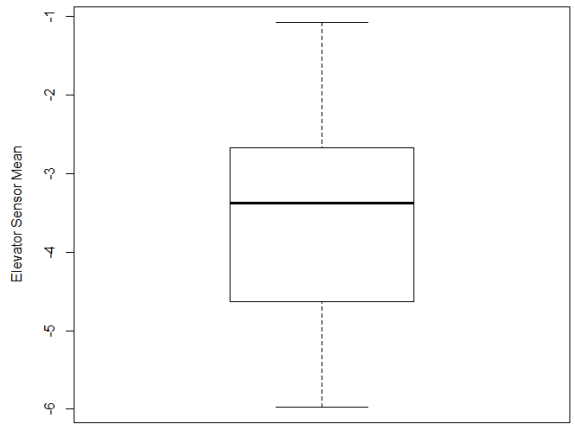

Figure 93: Elevator Sensor – Box Plot Lower Density Altitude Means

b. Density Plot

The density plot for the elevator sensor means variable shows a non-normal distribution with two peaks in the data values and a slight negative skew. Bandwidth indicates a good fit to the underlying variable value distribution.

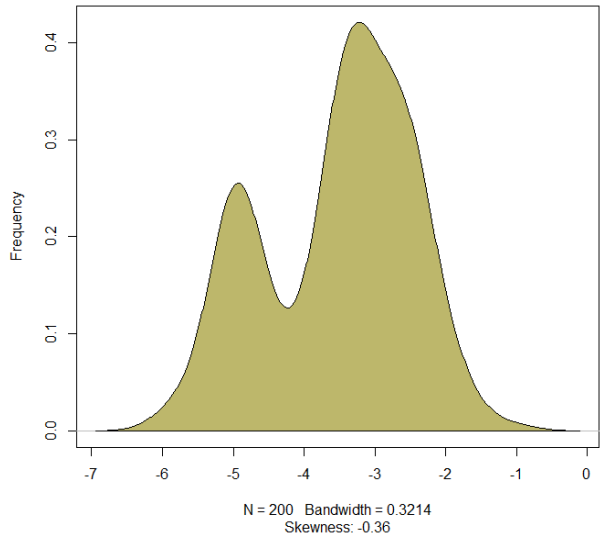

Figure 94: Elevator Sensor – Density Plot Lower Density Altitude Means

Appendix O – Experiment 16: Linear Regression Lower Density Altitude Data

Regression models using the variables of interest from the lower density altitude means data were developed for three distinct cases. These cases are intended to quantify the relationship between turbulence indicators, density altitude, and engineering burn rate, between density altitude and engineering burn rate, and between turbulence indicators alone and engineering burn rate.

O.1 Turbulence Indicators and Density Altitude

A script was developed in the analysis environment to accept the lower density altitude means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, all four predictor variables were used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate seem to cluster close to the regression line, but outlying values increase the spread.

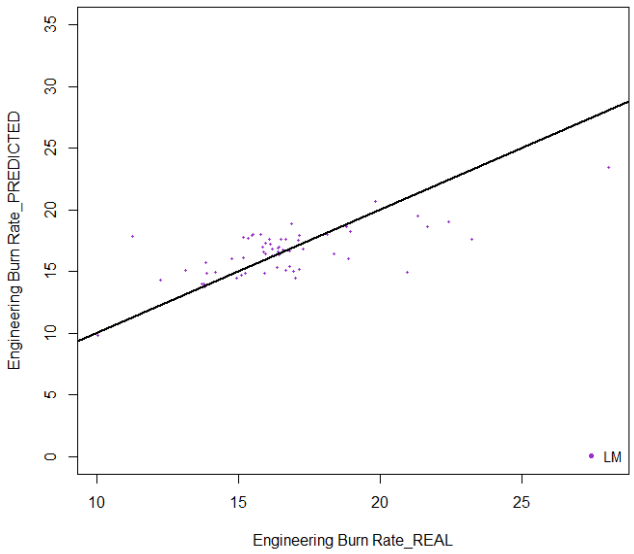

Figure 95: OLS Model – All Predictors Lower Density Altitude Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate seem to cluster close to the regression line, but outlying values increase the spread.

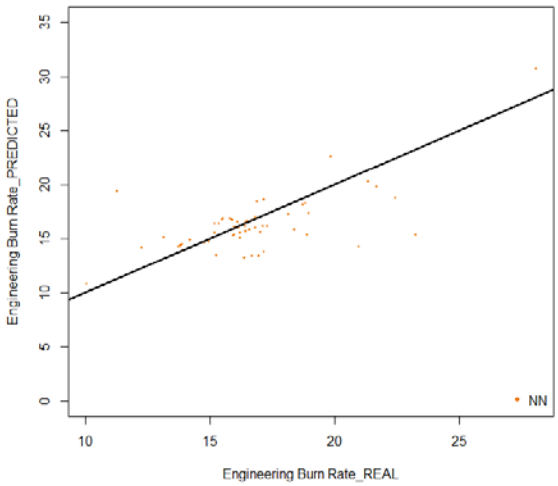

Figure 96: NN Model – All Predictors Lower Density Altitude Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.6145. The p-value, which is more than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE for the OLS linear regression model is 4.18. The MSE for the neural network linear regression model is 5.08.

O.2 Density Altitude

A script was developed in the analysis environment to accept lower density altitude means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the density altitude predictor variable was used.

a. OLS Regression

OLS linear regression shows the predictions for engineering burn rate do not cluster near the regression line to a great degree, and possess a significant amount of spread, though the range is narrow.

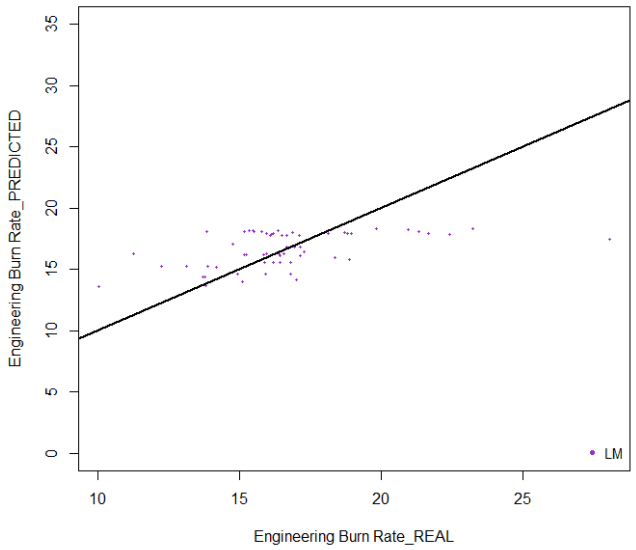

Figure 97: OLS Model – Density Altitude Lower Density Altitude Means

b. Neural Network Regression

Neural network linear regression shows the predictions for engineering burn rate do not cluster near the regression line, and possess a significant amount of spread, though the range is narrow.

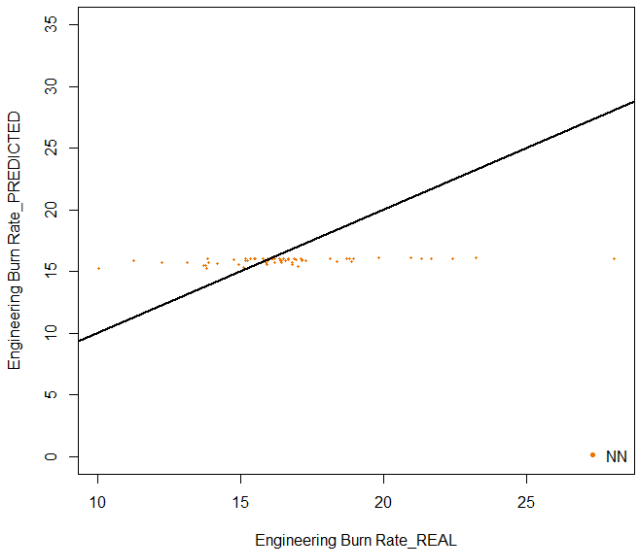

Figure 98: NN Model – Density Altitude Lower Density Altitude Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.3206. The p-value, which is more than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE of the OLS linear regression model is 6.10. The MSE of the neural network linear regression model is 8.04.

O.3 Turbulence Indicators (Throttle, Rudder, Elevator)

A script was developed in the analysis environment to accept the lower density altitude means data set, randomly partition the data according to a seventy/thirty split, and calculate regression models using OLS and neural network methods. In this case, only the turbulence indicator variables were used.

e. OLS Regression

OLS linear regression shows the predictions for engineering burn rate seem to cluster near the regression line, with outlying values increasing the spread.

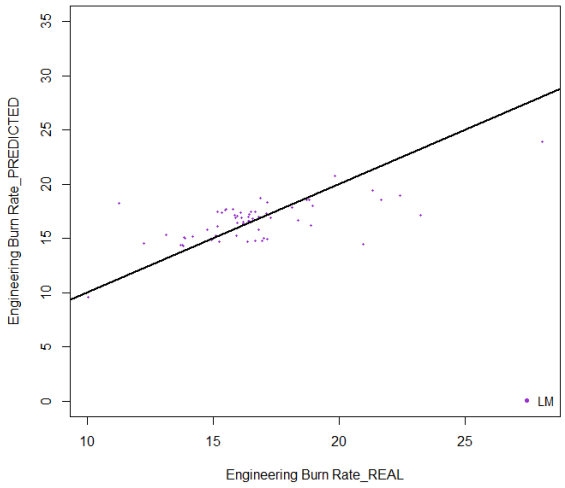

Figure 99: OLS Model – Turbulence Lower Density Altitude Means

b. Neural Network Regression

Neural network regression shows the predictions for engineering burn rate seem to cluster near the regression line, with outlying values increasing the spread.

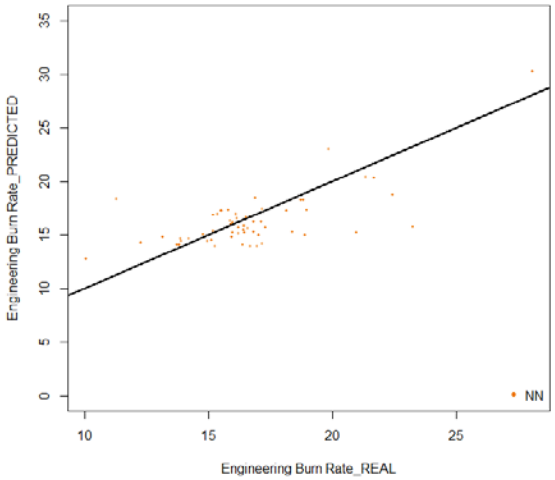

Figure 100: NN Model – Turbulence Lower Density Altitude Means

c. Breusch-Pagan Test

The Breusch-Pagan Test shows a p-value of 0.2344. The p-value, which is more than the significance level of 0.05, indicates the null hypothesis (variance of the residual errors is constant) should not be rejected.

d. MSE

The MSE of the OLS linear regression model is 4.27. The MSE of the neural network linear regression model is 4.65.

Bibliography

[1] T. H. Davenport and D. J. Patil, “Data scientist: the sexiest job of the 21st century.,” *Harv. Bus. Rev.*, vol. 90, no. 10, p. 2012, 2012.

[2] M. A. Waller and S. E. Fawcett, “Data Science , Predictive Analytics , and Big Data: A Revolution That Will Transform Supply Chain Design and Management,” *J. Busienss Logist.*, vol. 34, no. 2, pp. 77–84, 2013.

[3] C. Gandrud, *Reproducible Research with R and R Studio*, First. New York: CRC Press, 2015.

[4] H. Wickham, “Tidy Data,” *J. Stat. Softw.*, vol. 59, no. 10, pp. 1–11, 2014.

[5] F. Geerts, G. Mecca, P. Papotti, and D. Santoro, “The LLUNATIC Data-Cleaning Framework,” in *Proceedings of the VLDB Endowment*, 2013, vol. 6, no. 9, pp. 625–636.

[6] Y. Tian, P. Michiardi, and M. Vukolic, “Bleach: A Distributed Stream Data Cleaning System,” 2016.

[7] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, K. Goldberg, and U. C. Berkeley, “ActiveClean : Interactive Data Cleaning For Statistical Modeling,” in *Proeedings of the VLDB Endowment*, 2016, vol. 9, no. 12, pp. 948–959.

[8] X. Chu and I. Ilyas, “Qualitative Data Cleaning,” in *Proceedings of the VLDB Endowment*, 2016, vol. 9, no. 13, pp. 1605–1608.

[9] J. Freire, A. Bessa, F. Chirigati, H. Vo, and K. Zhao, “Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips,” *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.*, pp. 63–77, 2016.

[10] S. Ramanan *et al.*, “Determinants of theory of mind performance in Alzheimer’s disease: A data-mining study,” *Cortex*, vol. 8, pp. 8–18, 2016.

[11] X. Cong *et al.*, “Influence of Feeding Type on Gut Microbiome Development in Hospitalized Preterm Infants,” *Nurs. Res.*, vol. 66, no. 2, pp. 123–133, 2017.

[12] J.-H. Kao *et al.*, “Spatial analysis and data mining techniques for identifying risk factors of Out-of-Hospital Cardiac Arrest,” *Int. J. Inf. Manage.*, vol. 37, no. 1, pp. 1528–1538, 2017.

[13] J. Zhang, *Time Series Analysis Methods and Applications for Flight Data*, First. Berlin-Heidelburg: Springer, 2017.

[14] B. Collins, “Estimation of Aircraft Fuel Consumption,” *J. Aircr.*, vol. 19, no. 11, pp. 969–975, 1982.

[15] Y. S. Chati and H. Balakrishnan, “A Gaussian Process Regression Approach to Model Aircraft Engine Fuel Flow Rate Base of Aircraft Data (BADA),” in *Proceedings of The 8th ACM/IEEE International Conference on Cyber-Physical Systems, Pittsburgh, PA USA, April 2017 (ICCPS 2017)*.

[16] E. T. Turgut, “An Analysis of the Effect of Non-Payload Weight on Fuel Consumption for a Wide-Bodied Aircraft,” *ANADOLU Univ. J. Sci. Technol. A - Appl. Sci. Eng.*, vol. 18, no. 1, pp. 59–59, 2017.

[17] Federal Aviation Administration, *Pilot ’ s Handbook of Aeronautical Knowledge*. 2016.

[18] G. James, *An Introduction to Statistical Learning*, Sixth. New York: Springer, 2013.

[19] N. Gupta, “Artificial Neural Network,” in *International Conference on Recent Trends in Applied Sciences with Engineering Applications*, 2013, vol. 3, no. 1, pp. 24–28.

[20] F. Günther and S. Fritsch, “Neuralnet: Training of Neural Networks,” *R J.*, vol. 2, no. 1, pp. 30–38, 2010.

[21] N. Murata, S. Yoshizawa, and S. Amari, “Network information criterion --- determining the number of hidden units for an artificial neural network model,” *IEEE Trans. Neural Networks*, vol. 5, no. 6, pp. 865–872, 1994.

[22] T. S. Breusch and A. R. Pagan, “a Simple Test for Heteroscedasticity and Random Coefficient Variation,” *Econometrica*, vol. 47, no. 5, pp. 1287–1294, 1979.

[23] A. F. Hayes and L. Cai, “Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation,” *Behav. Res. Methods*, vol. 39, no. 4, pp. 709–722, 2007.

[24] National Weather Service Engineering Division, *Percision Digital Barometer Specification*. 1998.

[25] M. Crawley, *Statistics: An Introduction Using R*, Second. Chinchester, West Sussex, UK: John Wiley and Sons, 2015.

[26] B. M. Damghani, D. Welch, C. O’Malley, and S. Knights, “The Misleading Value of Measured Correlation,” *Wilmott*, vol. 2012, no. 62, pp. 64–73, 2012.

[27] J. L. Rodgers and W. A. Nicewander, “Thirteen Ways to Look at the Correlation Coefficient,” *Am. Stat.*, vol. 42, no. 1, p. 59, 1988.

[28] J. Aldrich, “Correlations Genuine and Spurious in Pearson and Yule,” *Stat. Sci.*, vol. 10, no. 4, pp. 364–376, 1995.

[29] C. Achen, “Measuring Representation: Perils of the Correlation Coefficient,” *Am. J. Pol. Sci.*, vol. 21, no. 4, pp. 805–815, 1977.

[30] T. Lumley, P. Diehr, S. Emerson, and L. Chen, “The Importance of the Normality Assumption in Large Public Health Data Sets,” *Annu. Rev. Public Heal.*, vol. 23, pp. 151–169, 2002.

[31] D. Nikolic, R. C. Muresan, W. Feng, and W. Singer, “Scaled correlation analysis: A better way to compute a cross-correlogram,” *Eur. J. Neurosci.*, vol. 35, no. 5, pp. 742–762, 2012.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
16-06-2017	Master's Thesis	December 2016 – June 2017

TITLE AND SUBTITLE

Statistically Modeling Fuel Consumption with Heteroscedastic Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dazzio, L. Elaine, GS-12, DAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/ENG)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENG-MS-17-J-075

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

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
Aircraft operate in unpredictable environmental conditions. As a result, autopilot design is difficult, as optimal responses cannot be anticipated for all conditions. Consequently, the autopilot might overcorrect for conditions, using more fuel than necessary. By analyzing performance data on a subject aircraft, the relationships between environmental condition variables and fuel consumption using linear regression models have been characterized. These relationships are accurate, even though the data is non-normal and heteroscedastic.

15. SUBJECT TERMS
Aerospace, linear regression, heteroscedasticity, aircraft fuel consumption, autopilot performance analysis

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. Scott Graham, AFIT/ENG
U	**U**	**U**	**UU**	158	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4581 (Scott.Graham@afit.edu)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18