Analysis of Virtual Sensors for Predicting Aircraft Fuel Consumption

Tim Woodbury[*]

Texas A&M University, College Station, TX, 77843

Ashok N. Srivastava[†]

NASA Ames Research Center, Moffett Field, CA, 94035

Previous research described the use of machine learning algorithms to predict aircraft fuel consumption. This technique, known as Virtual Sensors, models fuel consumption as a function of aircraft Flight Operations Quality Assurance (FOQA) data. FOQA data consist of a large number of measurements that are already recorded by many commercial airlines. The predictive model is used for anomaly detection in the fuel consumption history by noting when measured fuel consumption exceeds an expected value. This exceedance may indicate overconsumption of fuel, the source of which may be identified and corrected by the aircraft operator. This would reduce both fuel emissions and operational costs. This paper gives a brief overview of the modeling approach and describes efforts to validate and analyze the initial results of this project. We examine the typical error in modeling, and compare modeling accuracy against both naïvely complex and naïvely simplistic regression approaches. We also estimate a ranking of the importance of each FOQA variable used as input, and demonstrate that FOQA variables can reliably be used to identify different modes of fuel consumption, which may be useful in future work. Analysis indicates that fuel consumption is accurately predicted while remaining theoretically sensitive to sub-nominal pilot inputs and maintenance-related issues.

Nomenclature

y	measured value of fuel consumption
$\tilde{y}$	predicted value of fuel consumption
σ_x	standard deviation in value x
city pair	grouping of a departure city with a destination city
RMSE	root mean square error
NRMSE	normalized root mean square error
GLM	generalized linear model
N.N.	neural network
G.P.	Gaussian process
Stable GP	stable Gaussian process

[*] M.S. Candidate, Aerospace Engineering Department, 301 Reed-McDonald. AIAA Student Member, twoodbury@tamu.edu

[†] Project Manager for the System-Wide Safety and Assurance Technologies Project, Intelligent Systems Division. AIAA Senior Member, ashok.n.srivastava@nasa.gov

I. Introduction

Modern commercial aircraft continuously record hundreds of measurements of external conditions, pilot settings, and the health of various onboard systems. Various data mining techniques are used to identify patterns in these very large data sets. The basic purpose of analyzing large sets in this fashion is to learn information that can be used to improve the overall health of the vehicle and reduce the operational cost for the airline. Aircraft already have warning systems to detect faults that endanger the immediate well-being of the aircraft. However, a minor issue may reduce the fuel efficiency of the aircraft. Major aircraft operators, like Southwest, consume over a billion gallons of fuel annually,[1] so even a small reduction in fuel consumption could represent a significant savings.

The dominant technique to monitor fuel consumption used by aircraft operators is to compare the total fuel consumed during a flight against an expected value based on historical data and certain flight context information. The flight is considered anomalous if the total fuel consumption during the flight exceeds the expected value plus a certain threshold.

The modeling approach considered in this paper uses regression models to detect anomalies in aircraft fuel consumption. Fundamentally, we assume that aircraft fuel consumption can be described as a nonlinear function of recorded FOQA data. A limited number of parameters, representing extrenal conditions, current flight status, and limited pilot inputs, are utilized to approximate this function. Historical flight data are used to generate a model of typical fuel consumption, and flights that deviate from the expected performance are noted.

Specifically, the instantaneous rate of fuel consumption is modeled as a function of data routinely collected by commercial aircraft. The proposed technique can identify high fuel consumption during a particular segment of flight. Furthermore, FOQA data are used as predictors, and can also be used to help determine the cause of fuel overconsumption. This approach to anomaly detection appears to be the first of its type in aviation.

The theoretical background and initial results for this study are detailed in Ref. 2. The primary purpose of this document is to determine if those results are logically consistent. The modeling process must be accurate enough to offer useful predictions. A relatively small number of FOQA variables are used as modeling inputs; this leaves regression models theoretically sensitive to a measurements that can be reflective of pilot commands or aircraft maintenance. We compare modeling accuracy using limited data as inputs versus modeling using all data as inputs, as a further check on model accuracy. Results from regression modeling are also compared against a much simpler, time-based, prediction method. For future work, it is useful to quantify the importance of each input variable; we utilize and compare two different approaches to rank the importance of the inputs. Finally, we examine the scatter in the recorded flight data by considering the typical progression of fuel consumption in time. The current paper: 1) summarizes the methodology and results previously developed; 2) describes the approaches taken to validate that methodology; 3) presents and discusses new results.

II. Prediction of Fuel Consumption

In this section, we summarize the fuel consumption models developed in Ref. 2. Representative results of that study are presented.

A. Virtual Sensors

The current research uses the method of Virtual Sensors to detect fuel overconsumption. Virtual sensors is a technique initially developed to estimate unmeasured sensor readings. The fundamental idea behind Virtual Sensors is to use time histories of state and control measurements to create an estimator for a target quantity.[3] Virtual sensors employs nonlinear regression modeling. In the initial implementation, readings from an Earth-observing satellite were used to estimate the value of unmeasured spectral data in the substantial historical data from an older satellite. The new satellite recorded spectral data at similar frequencies to the outmoded satellite, as well as new frequencies of interest. Predictive models were developed for the new readings as a function of the measurements common to both satellites. The models also offered an estimate of the uncertainty in the modeling process. This model for the unmeasured spectral values was called a "Virtual Sensor" because it allowed for the prediction of a value in the absence of a direct measurement.

Subsequently, the Virtual Sensors technique was used to correctly detect a fault from Space Shuttle Main Engine data.[4] In this case, damage in the engine fuel line led to a potentially catastrophic situation. It was demonstrated that Virtual Sensors could detect this anomaly from historical data.

The application of Virtual Sensors to detect fuel overconsumption is different than in the aforementioned cases, as measurements of fuel consumption for commercial aircraft are readily available. The current research aims to build a model of fuel consumption, as a function of correlated measurements, to compare against the recorded fuel consumption.[3] The models developed are restricted to a single make of aircraft, but are not specific to individual airplanes. Near airports, instructions from air traffic control can have a significant influence on the flight pattern and fuel consumption of the airplane, and these instructions are unrelated to maintenance or operator faults in the aircraft. Consequently, for modeling purposes, only the first 25 minutes of flight after the landing gear retract are considered. This avoids the influence of air traffic control during takeoff and landing, and also has the effect of making the flight segments considered the same length.

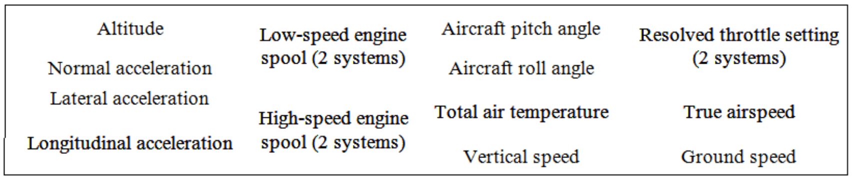
B. Modeling approach

Flight data, acquired from an aircraft operator, are divided into training and testing data sets. All aircraft are of the same model, and the data identify individual aircraft by a tail number (or equivalent specification). The Virtual Sensors prediction can identify specific aircraft that consistently have relatively high fuel consumption.

Typically, the training data represent one set of city pairs, and test data represent a different set of city pairs, with no overlap between sets. This is done to consider a "worst-case" implementation of the program. We consider any unique pairing of departure and arrival locations to be its own city pair; e.g., flying from city A to city B would be labelled distinctly from a flight from city B to city A. The training data are collected into exclusive subsets of fixed sizes. Bootstrap sampling is performed on each subset some number of times, and regression models are trained to each bootstrap sample. Bootstrap sampling creates a subset by randomly selecting flights from the set, with replacement after each selection; consequently, a single flight may appear multiple times in a given subset. Regression models use the desired aircraft state histories as inputs to model fuel consumption. A sample of inputs used in one implementation is given in Table 1. The modeling process creates an ensemble of regression models for fuel consumption. The estimate is the mean of the prediction from all the trained models in the ensemble. The prediction confidence bounds are typically taken as three times the standard deviation of the ensemble prediction. The testing data are then evaluated with the resulting model, and flights for which the measured fuel consumption exceeds the confidence bounds of the estimate are noted. This technique of prediction via emsembles of regression models trained on bootstrap samples is known as bagging.[2]

For the results discussed later, training data are broken into groups of 100 flights. Each subset is bootstrap sampled three times, and one regression model is fit to each bootstrap sample. Thus, a total of $\frac{3N}{100}$ models are developed, where N is the total number of flights rounded up to hundreds. Models are built using one of four regression algorithms; all ensembles are built from the same algorithm. Essentially, four estimators are developed (one for each algorithm), each using an ensemble of $\frac{3N}{100}$ regression models. The four algorithms used are the generalized linear model (GLM), the neural network (N.N.), the Gaussian process (G.P.), and the stable Gaussian process (Stable GP). The codes for these algorithms are widely available and all four have been used previously by Ref. 5. A detailed understanding of these algorithms is not the objective of this paper, so only a brief overview of each will be given.

Table 1: Sample input variables for one implementation of the program.



Altitude	Low-speed engine spool (2 systems)	Aircraft pitch angle	Resolved throttle setting (2 systems)
Normal acceleration		Aircraft roll angle	
Lateral acceleration	High-speed engine spool (2 systems)	Total air temperature	True airspeed
Longitudinal acceleration		Vertical speed	Ground speed

C. Regression algorithms

GLMs are a generalization of linear regression that allows single-variable output data to be fit to a broad range of possible distributions, including the normal, Poisson, and exponential distributions. In general, a GLM relates the expectation of the output value, $E(y_i) = \mu_i$, to a linear combination of the input variables via a link function, $g(\mu_i)$. The link function may take many forms, but a general expression is given by Eq. 1:[6,7]

$$g(\mu_i) = g(E(y_i)) = \bar{x}_i^T \bar{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1}$$

Neural networks are a popular technique that model the desired outputs as nonlinear functions of features derived from the inputs. The basic N.N. consists of an input layer, a number of hidden layers, and an output layer. Each layer consists of nodes, which accept input from some or all of the nodes in the preceding layer, and output to at least one node in the following layer. Each node with its inputs and its output comprises a "neuron." The process of fitting a N.N. to data is form of nonlinear regression.[3]

Gaussian Process regression is a method of performing Bayesian inference on continuous values with a Gaussian Process prior. The underlying distribution is estimated and used to predict the output in the testing data.[8] Stable GP is a modification of the typical G.P. implementation. Stable GP is intended to reduce numerical errors and is better scalable to large data sets compared to the traditional G.P. implemented.[9]

D. Model testing and evaluation

The testing data are drawn from the same source as the training data. At each time step in the test data, the ensemble of models is used to predict fuel consumption from the input values. The error between the ensemble prediction and the recorded value, as well as the uncertainty in the prediction, is noted. A data point is considered high if it exceeds the upper bound, and low if it is less than the lower bound. The percentages of high and low data points are then determined for every flight in test data, regardless of the aircraft identity. The percentage of erroneous points is also determined for each unique aircraft tail number, and for all flights for each algorithm.

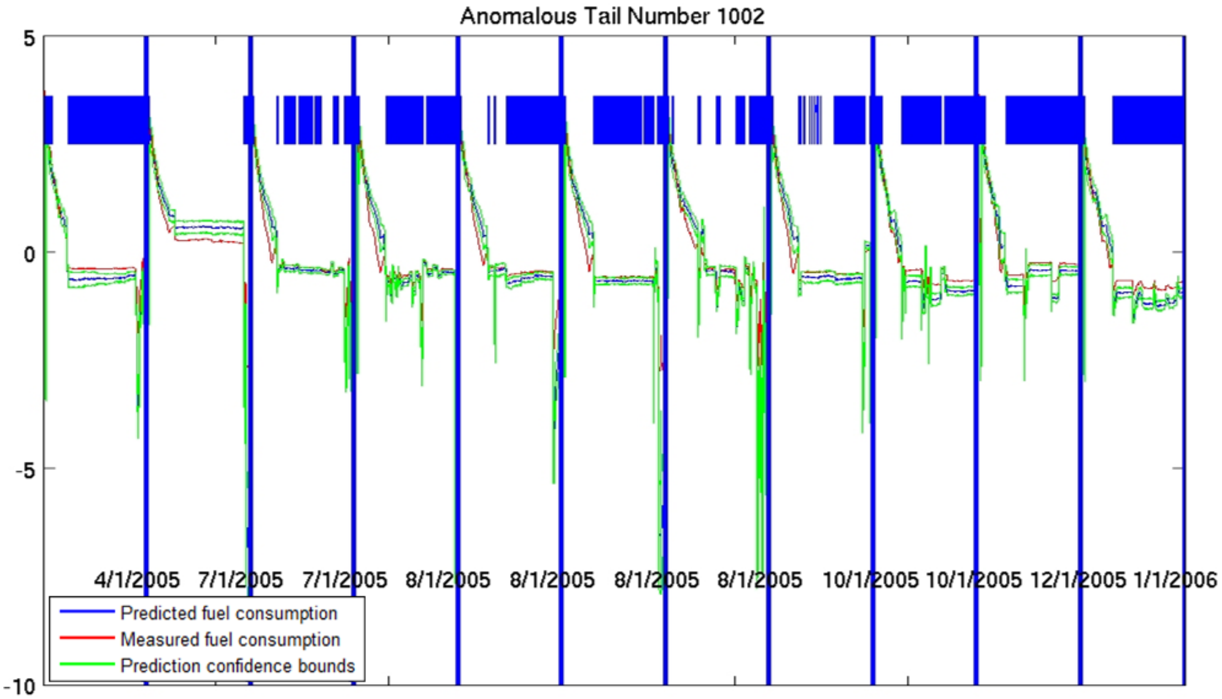


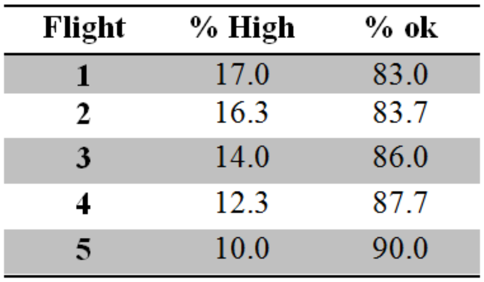
Figure 1: Measured and predicted results for several flights for an aircraft with a large percentage of "high" data points. The dates and identifying tail number shown are fictitious, to protect possibly proprietary information about the data provider. Each flight is separated from the next by blue vertical lines.

We typically plot the measured and predicted fuel consumption with confidence bounds for airplanes with the greatest percentage of high data points. Figure 1 shows results for one aircraft with frequent exceedances. Eleven flights of this aircraft are shown consecutively; flights are separated by blue vertical lines. Each flight begins when the landing gear retract and ends 25 minutes later. The measured fuel consumption is represented by a red curve; the prediction and confidence bounds are indicated, respectively, by blue and green curves. The x-axis and y-axis represent normalized time and fuel consumption, respectively. The dates of each flight shown are fictitious, to protect possibly proprietary airline data. The blue marks above the curves indicate points of exceedances. The sample aircraft shown has an unusually high percentage of exceedances.

Both Table 2 and Table 3 summarize testing data results from the GLM. Table 2 shows the percentage of high and nominal data points in the individual flights with the highest percentage of exceedances. These data are drawn from a different data set than are the results shown in Fig. 1. The data in Tables 2 and 3 are those described in Sec. IV.A and used in most of the analysis in this paper. Note that even flights with the most exceedances in this data set fall within the confidence bounds more than 80% of the time. Table 3 shows the individual aircraft with the highest percentage of exceedances across all flights. The number of flights is shown as an indicator of the consistency of the aircraft's performance. Note that the aircraft with the lowest percentage of nominal data points has only one flight, so the large percentage of exceedances may be a result of bad data or some other factor. The other aircraft in Table 3 have data from multiple flights. Although the percentage of exceedances is typically less than 2%, exceedances that occur infrequently might be early indicators of more severe faults.

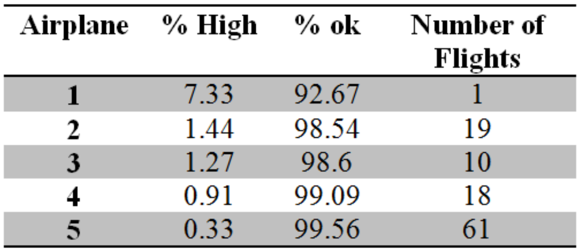
The cases with the greatest percentage of exceedances are shown, but only about 11% of all flights had any points fall outside of the confidence bounds. The majority of flights registered as nominal. Furthermore, most aircraft experienced agreement with the GLM prediction 99% of the time or more. Only a few aircraft seem to display any kind of consistent error. The commercial aircraft operator has very strong incentives to minimize fuel waste; therefore, we anticipate that the majority of aircraft will have nominal fuel consumption, with exceedances occuring very rarely. The predictive model appears to model most aircraft accurately while detecting a small minority of anomalies, a result that meets our expectation.

Table 2: Results for individual flights with highest percentage of exceedences. Results are shown for one set of input data and sorted by decreasing % High, then by increasing % ok. Results are from the GLM.



Flight	% High	% ok
1	17.0	83.0
2	16.3	83.7
3	14.0	86.0
4	12.3	87.7
5	10.0	90.0

Table 3: Results for specific airplanes with highest percentage of exceedances in all their flights. Results are sorted by decreasing % High, then by increasing % ok. Results are from the GLM.



Airplane	% High	% ok	Number of Flights
1	7.33	92.67	1
2	1.44	98.54	19
3	1.27	98.6	10
4	0.91	99.09	18
5	0.33	99.56	61

III. Validation and Analysis

Recent efforts on this topic have focused on validating and analyzing results of this approach. Ultimately, the methodology must be validated by identifying a maintenance or operator fault on an airplane detected by the model as having high fuel consumption. As of the time of this analysis, we are in the process of performing such a validation. In the interim, we seek to show that the results offer acceptable accuracy, and that the data appear to be suitable for the regression approach employed. The relevant results are summarized in Section IV; this section covers the background of the results.

A. Normalized root mean squared error

Previously, the only comparative metric between flights and algorithms was the percentage of high and low flights. Root mean square error (RMSE) and normalized root mean square error (NRMSE) were introduced as metrics of the modeling accuracy. Recall that the prediction confidence bounds are proportional to the standard deviation of the ensemble model. If the input data are poor predictors of fuel consumption, the ensemble will have very wide confidence bounds. This could make the models insensitive to anomalies. To address this concern, NRMSE is used to quantify model accuracy. For a data set of n measured outputs y and predicted outputs $\tilde{y}$, RMSE and NRMSE are defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} y_i - \tilde{y}_i}{n}} \tag{2}$$

$$NRMSE = \frac{RMSE}{\sigma_{\tilde{y}}} \tag{3}$$

NRMSE was first used to compare the performance of different algorithms, different tail numbers, and different flights of specific tail numbers. It was later used to examine the performance of airplanes flying between the same city pairs. Since models were trained and tested on different city pairs, an analysis of the performance of each city pair in the test data was conducted, to determine if the predictive models were biased towards any particular pair.

As a check on the modeling process, two alternate predictors were employed. The first variation used all the continuously recorded parameters as inputs for model-building (obvious analogues for fuel consumption were not included, as this would defeat the purpose of the Virtual Sensors technique). This analysis was designed to demonstrate that the limited number of parameters used in the standard program were sufficient to make a good approximation to the fuel consumption, and that influential input variables had not been omitted.

The second variation did not use regression models. As before, the training data were broken into groups of 100, and each group was bootstrap sampled into three subsets of 100. The mean of the y values of each subset at each time step was saved as the "prediction" of the subset at that time. At each time step, the average of predictions of the subsets was used as the predicted fuel consumptionm, yielding a very simple approximation for the fuel consumption as a function of time. This exercise was performed to demonstrate that the regression models offered a significant improvement over the simplest prediction method.

B. Ranking of input variables

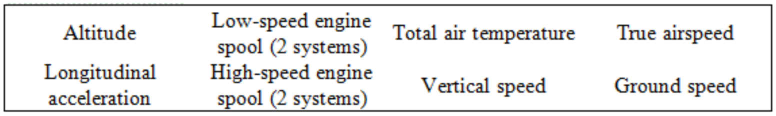
We would like to quantify the importance of each of the inputs used; knowing which variables are most related to accurate models may be useful for input selection in future projects. Two approaches are used to rank the significance of each input variable. The first approach exploited the nature of the GLM. As discussed in Sec. II.C, the GLM relates a function of the expectation of the output to a linear equation of the input variables. For the purposes of discussion, a simplified model of the GLM is offered, in which the expectation of the output is given by Eq. 4 and the error is given by Eq. 5. In Eq. 5, $Q(\bar{\beta})$ is a nonlinear function that depends on the magnitude of $\bar{\beta}$, and λ is a scalar constant.[2]

$$y_t \approx \bar{x}_t^T \bar{\beta} + \beta_0 \tag{4}$$

$$e \approx \sum_{t=1}^{T} ||y_t - \bar{x}_t^T \bar{\beta} + \beta_0||_2 + \lambda Q(\bar{\beta}) \tag{5}$$

Of interest here is the term $\lambda Q(\beta)$. The error term is a combination of the modeling error in y_t plus a function of the magnitude of $\bar{\beta}$. As λ increases from 0, the error becomes dominated by the second term, and the error is minimized by decreasing the magnitude of the terms in $\bar{\beta}$ while approximating y_t as well as possible. As such, the term in λ has the effect of driving to zero those coefficients in $\bar{\beta}$ that have the least influence on the regression model's accuracy. Essentially, the magnitude of each term in the coefficient vector $\bar{\beta}$ ranks the value of each input variable in predicting the output variable. To examine this "ranking", we examined the output of generalized linear models for values $0 < \lambda < 1$, and the resulting value of each input variable.

Table 4: Summary of input variables for brute-force iteration over the inputs.



Altitude	Low-speed engine spool (2 systems)	Total air temperature	True airspeed
Longitudinal acceleration	High-speed engine spool (2 systems)	Vertical speed	Ground speed

For the second approach to ranking, generalized linear models were built using ALL unique combinations of input variables. The errors in the resulting models were examined. The problem was intractable with all sixteen original inputs. To reduce the computational time required, the number of input variables was reduced to ten (see Table 4 for a listing). The variables selected were the ten with the largest coefficients in $\bar{\beta}$ from the original model-building process.

C. Development of fuel consumption in time

The progression of the data over the duration of the flight was analyzed. Recall that the flight intervals considered began when the landing gear retracted, such that the start of each flight roughly corresponded to a common state soon after takeoff. It was necessary that most flights would show a common general profile of fuel consumption versus time; if flights showed excessive variation in time, then it would be difficult to build a reliable, general model for fuel consumption. To examine the variation in fuel consumption as a function of time, the time axis was discretized into constant steps Δt. The mean and standard deviation of all points of all flights for each city pair in each time region were computed and plotted.

Results of the previously described analysis led us to perform a simple classification analysis on the data. It was noticed that for a particular sequence of time steps, two distinct "modes" of fuel consumption were seen in the histograms. Linear discriminant analysis (LDA) was performed to determine if the recorded input variables could be used to predict which fuel mode would be used. LDA is a basic classifier in which a hyperplane (line in two dimensions, plane in three dimensions) in the input vector space is used to separate a specified number of classes.[10]

IV. Results

Results are covered in five sections. The first section briefly describes the data set used in the following analysis. The second section describes analysis was performed on previous results using NRMSE as an error metric. The third section summarizes findings from the two simplistic models employed (using all the input variables for modeling, and using no model). The fourth section details the results of the GLM rankings of the input variables. The fifth section describes findings of the analysis of fuel consumption in time.

A. Data source

Data used in the analysis shown in Tables 2 and 3, and in all subsequent analysis, were provided by easyJet Airline Company Ltd, a U.K. operator of Airbus narrow-body aircraft. This data set consisted of recorded measurements from 7,905 flights between the ten most frequently visited city pairs between April and December 2010. Data from four city pairs, comprising 3,257 flights, were used for training; the remaining 4,648 flights were used for testing. A handful (less than 1%) of recorded flights in both sets was rejected as being too short (i.e., less than 25 minutes). The data consisted of measurements of hundreds of discrete and continuous variables recorded every second of the flight. To make the program more tractable, every fifth recording was used for modeling (as though data were recorded at 0.2 Hz). The sixteen input variables used with this data set are those shown in Table 1.

B. Analysis of previous results

The analysis summarized here was performed on results from the easyJet data set. Figure 2 shows histograms of the NRMSE of each flight for each algorithm utilized. We consider NRMSE less than about 20% to be acceptably low. The relatively low average NRMSE displayed by each algorithm suggests that the regression models do a good job of accurately predicting the fuel consumption in most cases. The Stable GP and GLM have noticeably lower NRMSE than the N.N. and G.P. The latter algorithms both have an average NRMSE

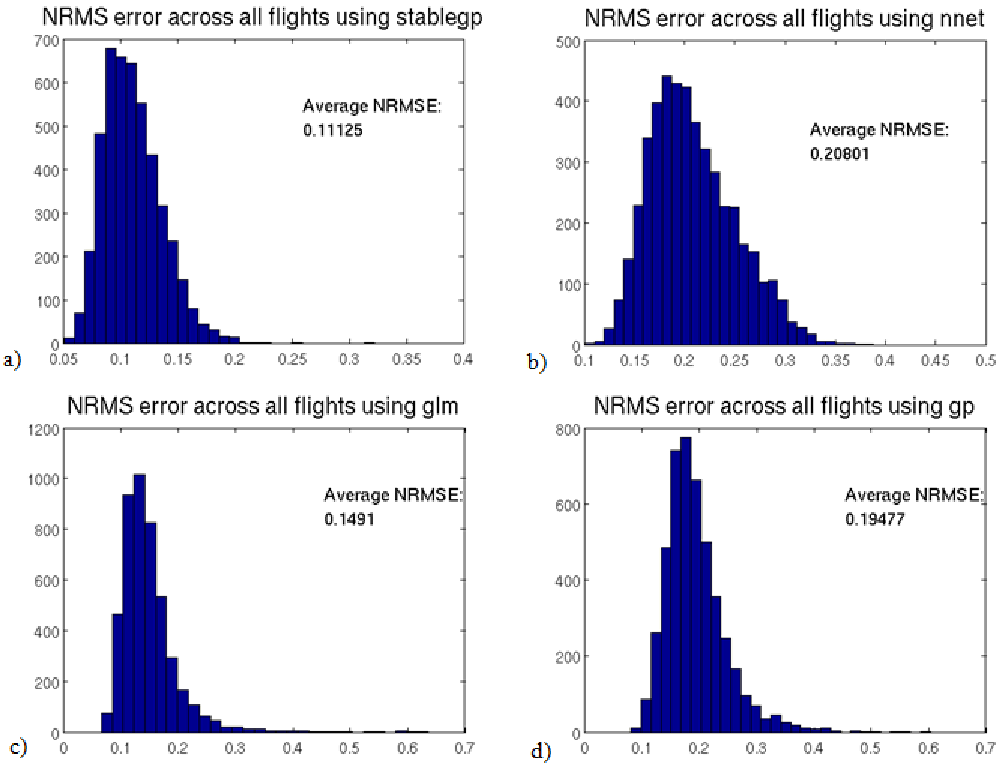


Figure 2: Histograms of NRMSE for each flight in the testing data, for: a) Stable GP, b) N.N., c) GLM, d) G.P. NRMSE of less than 20% typically indicates relatively good agreement between measured and predicted values.

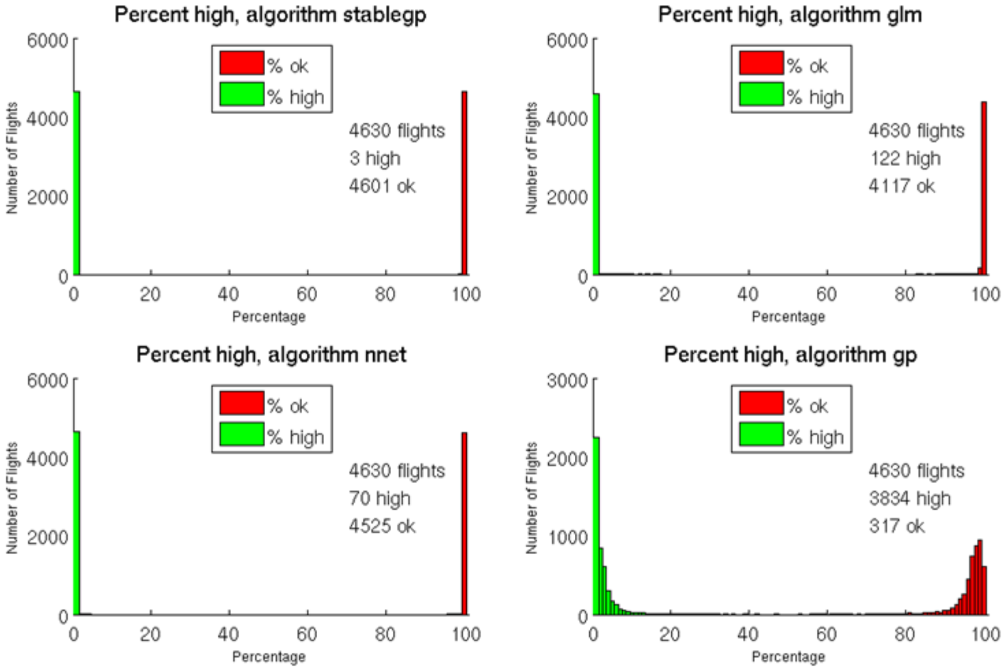


Figure 3: Histograms of the number of high and nominal flights for each algorithm used. Plotted are histograms of the percentages of high and nominal data points in each flight of a given test set. The number of high and nominal flights is displayed on each plot.

of approximately 20%. The N.N. has a greater spread in the NRMSE, and the G.P. shows a few flights with high NRMSE. Overall, results indicate the model fitting is working as it should.

Figure 3 contains histograms of the percentage of high and nominal points in all flights. Each histogram also displays the number of flights with any exceedances ("high" flights), and the number of flights for which all points fall within the confidence bounds ("ok"). In general, these figures show that close to 100 % of data points fall within the confidence bounds, and only a very small percentage of points exceed the model prediction. The Stable GP, GLM, and N.N. have very few off-nominal flights. The G.P. shows only a few hundred flights with no exceedances. Given the relative accuracy of the other algorithms, this appears to be indicative of poor model fitting. Since the NRMSE for the G.P. is comparable to the N.N., this suggests that the confidence bounds are narrower for the G.P. than for the N.N. This would lead to a large percentage of flights classified as off-nominal, and may be indicative of overfitting in the trained models.

C. Variations on the standard approach

The same four algorithms were used to develop regression models using all the continuously recorded variables as inputs. This list comprised 123 variables. Obvious analogues to fuel consumption, such as aircraft gross weight, were removed from the list of inputs. The purpose of this study was to compare the standard modeling process using sixteen inputs with the most complex model that could be created using the same process. If the models with all inputs showed a substantial improvement in error metrics, that could indicate that the sixteen variables selected were inadequate as predictors, and should be replaced or supplemented with different inputs.

Figure 4 shows summary histograms of the results. The average NRMSE decreased to approximately 5% for the Stable GP, N.N., and GLM. The average NRMSE for the G.P. actually increased by about 2%. In three of the four cases, using more variables as inputs reduced the error metric (as would generally be expected); however, we believe it is better to continue using a reduced number of inputs to reduce the complexity of the model.

The other alternate model tested naïvely predicted fuel consumption as a function of time alone, as

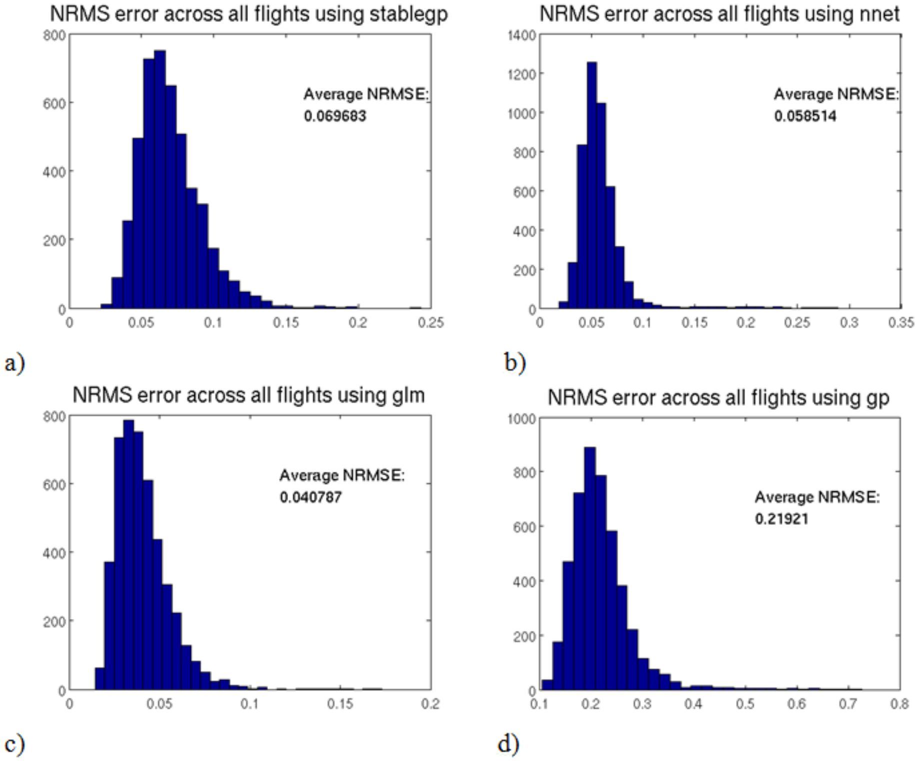


Figure 4: Histograms of NRMSE, using all applicable input variables, for each algorithm used. The NRMSE decreases for Stable GP, N.N., and GLM algorithms, but not for the G.P.

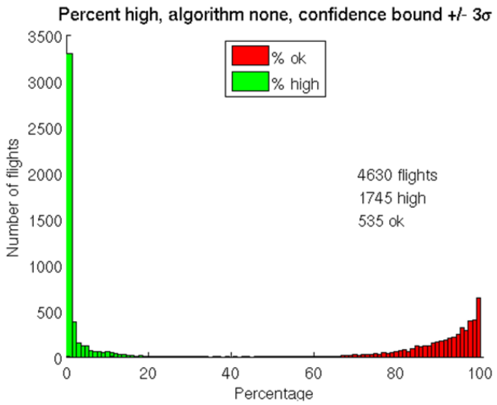
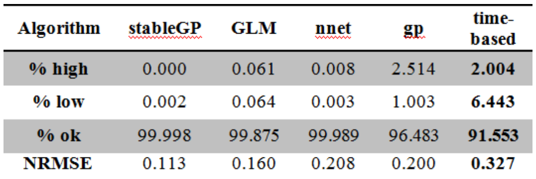


Figure 5: Summary of results using all input data for the time-based prediction.

described earlier. This exercise compared the standard modeling process with the simplest model possible, to demonstrate that the regression algorithms offered a significant improvement. When data from all flights were used for prediction, the NRMSE was about 33%. This is notably higher than the error in the standard regression models. Figure 5 shows a plot, similar to Fig. 3, using the means from all test data as predictors. All but a few hundred flights display exceedances. The time-based model is compared to each regression algorithm in Table 5. The percentage of nominal data points is about 92%, which is not much lower than than the other models; however, from the much higher NRMSE, it is clear that regression modeling is substantially much accurate than the naïvemodel.

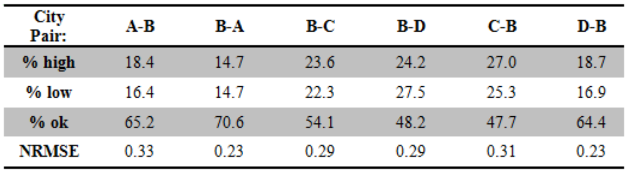
As a further check, the time-based estimation was "trained" and evaluated on individual city pairs. Table 6 shows the percentage of nominal and erroneous data points for each city pair. NRMSE improves for some city pairs, with a minimum of about 23% for two cases. However, the percentage of nominal data points descreases for all city pairs, to between 48-71%. Obviously, fuel consumption cannot be predicted as a simple function of time. Regression algorithms reduce both the NRMSE and percentage of anomalous data points significantly compared to the time-based model.

Table 5: Comparison of prediction accuracy for all algorithms and for simplistic model, ignoring city pair data. Percentages shown are the percentage of data points across all flights at which the measured fuel consumption was less than the estimate, within the confidence bounds, or exceeded the estimate, respectively.



Algorithm	**stableGP**	**GLM**	**nnet**	**gp**	**time-based**
% high	0.000	0.061	0.008	2.514	**2.004**
% low	0.002	0.064	0.003	1.003	**6.443**
% ok	99.998	99.875	99.989	96.483	**91.553**
NRMSE	0.113	0.160	0.208	0.200	**0.327**

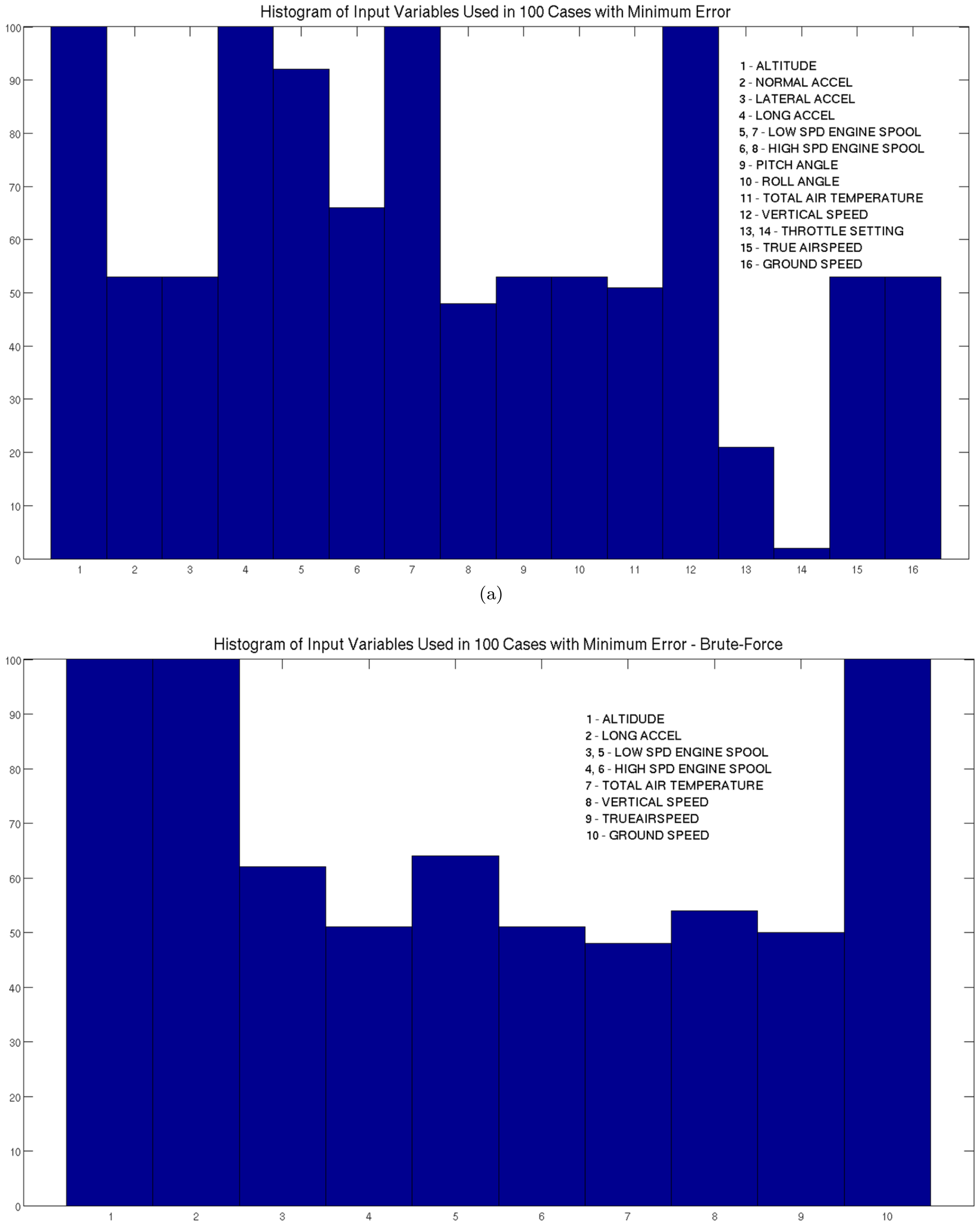
Table 6: Comparison of results using specific city pairs for the "algorithm-free" simple prediction. Shown are the percentage of data points for each city pair at which the measured fuel consumption was less than the estimate, within the confidence bounds, or exceeded the estimate, respectively.



City Pair:	**A-B**	**B-A**	**B-C**	**B-D**	**C-B**	**D-B**
% high	18.4	14.7	23.6	24.2	27.0	18.7
% low	16.4	14.7	22.3	27.5	25.3	16.9
% ok	65.2	70.6	54.1	48.2	47.7	64.4
NRMSE	0.33	0.23	0.29	0.29	0.31	0.23

D. Ranking of input variables

To rank the importance of each input variable, Eq. 5 was solved for different values of λ while iterating from $\lambda = 0$ to $\lambda = 1$. The error, defined as the $\mathcal{L}_2$ norm of the difference between the measured and predicted vectors of fuel consumption, was determined for each model built. In this case, values near zero in the vector $\bar{\beta}$ were judged to correspond to relatively unimportant input variables. For comparison, we performed a brute-force iteration, using unique combinations of the input variables in Table 4, to build GLMs, and the

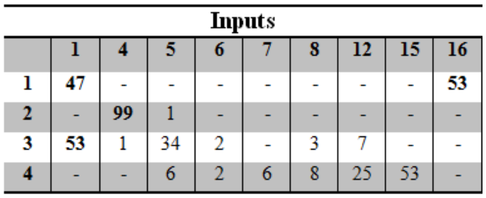


(a)

(b)

Figure 6: Histogram tallying the presence of each input variable in the 100 cases with the lowest error for the input ranking trials. Error is defined as the $\mathcal{L}_2$ norm of the vector $y - \tilde{y}$. a) corresponds to the iteration over different values of λ. It shows the frequency with which the coefficient of each input variable exceeded the tolerance for the 100 models with the lowest error. b) corresponds to the brute-force iteration over input variables to build GLMs. It shows the frequency with which each input was used in building the 100 GLMs that had the lowest error.

Table 7: Frequency with which each input variable had one of the four largest coefficients in $\bar{\beta}$, in the 100 cases with minimum error for iteration over λ in GLM building. Input labels correspond to those in Fig. 6a. Only inputs with nonzero occurrences are shown.



	Inputs								
	1	4	5	6	7	8	12	15	16
1	**47**	-	-	-	-	-	-	-	**53**
2	-	**99**	1	-	-	-	-	-	-
3	**53**	1	34	2	-	3	7	-	-
4	-	-	6	2	6	8	25	53	-

errors were determined for those models in the same fashion.

To compare the results of these two approaches, Fig. 6 contains histograms of the frequency of occurrence of each input in the 100 cases of lowest error for both the iteration over λ and for the brute-force GLM-building exercise. In both approaches, altitude and longitudinal acceleration are inputs for 100% of the minimum error models; clearly these are important inputs for accurate modeling. Other inputs, such as low speed engine spool rpm, vertical speed, and ground speed appear in 100% of models for one approach, but only in about half of models for the other approach.

Further insight can be gained from the vector $\bar{\beta}$ in the GLM, since the absolute value of each coefficient effectively ranks the importance of each input variable. To get a better understanding of the results in the histograms, $\bar{\beta}$ was used to rank the input variables in each of the 100 cases for the iteration over λ. The variable with the largest coefficient was ranked as first, the one with the second-largest coefficient was ranked as second, and so on. There is not a perfectly analagous metric for the brute-force approach because not all inputs were used to build models. Table 7 shows the frequency with which variables were ranked among the top four.

Based on Table 7, we can further conclude that ground speed is very significant, as it appears in 100% of the best brute-force models, and is ranked first in about half of the iterations over λ. True airspeed and one of the low-speed engine spool sensors may also be also be very important; true airspeed is ranked fourth in about half of the cases in the iterative approach and appears in about half of the brute-force cases. The low-speed engine spool measurements appear in over 90% of the iterative cases, and were used in about half of the best brute-force cases. Based on the results of this section, we conclude that altitude, longitudinal acceleration, ground speed, true airspeed, and engine rpm are some of the most significant predictors of aircraft fuel consumption, of the inputs considered.

E. Analysis of fuel consumption in time

There was interest in examining the variation in fuel consumption between different flights to the same city pair. The fuel consumption measurements for all flights to the same city pair were sorted into 25 second intervals. The mean and standard deviation of measured fuel consumption across each interval, for all flights to that city pair, was computed. Figure 8 shows representative results for two city pairs. The standard deviation appears to be relatively small throughout the flights, indicating that most flights follow a similar underlying trend. (This does not mean that flights are very precisely predicted by that trend, as discussed in Section IV.C.)

To better represent the development of fuel consumption with respect to time, histograms of all measured output values at specific times for all flights were created. In general, values tended to be distributed about a central mean in an approximately normal distribution, whose mean gradually decreased with time. However, at certain time steps, two distinct mean values were observed. This behavior is shown in Fig. 7. Rather than gradually decreasing, the peak at the initial time, corresponding to the higher fuel consumption rate, slowly decreased as the second peak, corresponding to lower fuel consumption, grew.

LDA was performed with the training data, and susequently used to classify test data points into one of the two fuel consumption modes based on the input values. The classification was accurate in 97% of cases, indicating that the fuel consumption mode was well predicted by the inputs. Analyzing the vertical speed of the aircraft during this time period indicated that the lower fuel consumption mode corresponded to near-zero vertical speed, while the higher mode corresponded to an upward vertical velocity. This fact suggests that the lower mode corresponds to a level cruise, which ideally should be the most fuel-efficient flight phase

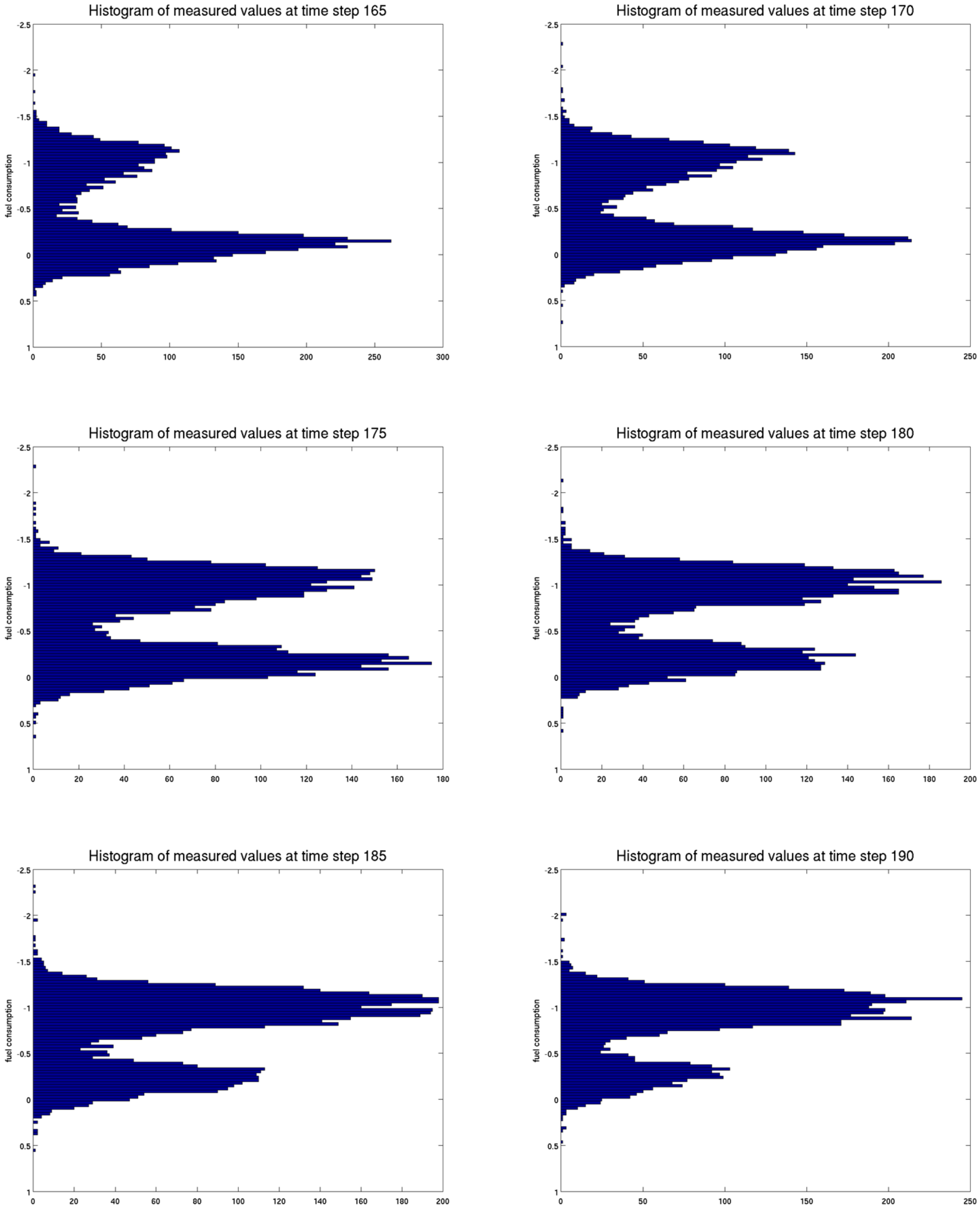
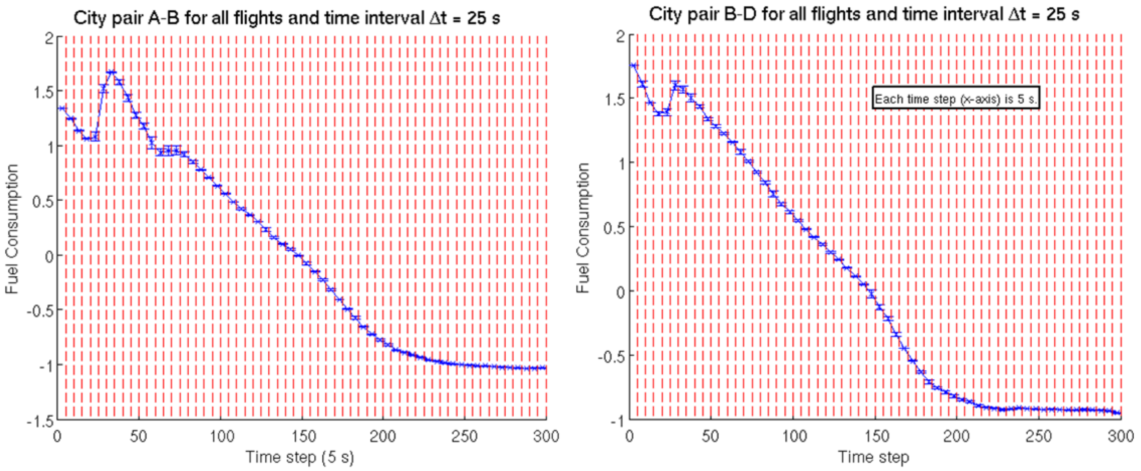


Figure 7: Development of distinct modes of fuel consumption between time 825 seconds and 950 seconds. The histograms shown are for all flights.



(a) (b)

Figure 8: Variation in fuel consumption throughout all flights for city pairs a) A-B and b) B-D. Plotted are the mean in each segment with standard deviation as the error bars. The x-axis labels are in time increments of 5 seconds, such that the entire flight considered is 1500 seconds. The y-axis has been normalized.

for a commercial airplane. The higher mode probably corresponds to the initial climb after takeoff. The examination of fuel consumption in distinct flight phases was not a major concern of this study; however, it might be useful in future studies to predict different modes of fuel consumption based on measured inputs.

V. Conclusions

A great amount of background work has been performed to develop a methodology for using data mining to predict the fuel consumption rate of commercial aircraft. This work, which is the topic of Ref. 2, accurately predicted the measured metric of fuel consumption for most aircraft and identified a small number of airplanes with higher than predicted fuel consumption. This is approximately the distribution that was expected. The occurrence of higher than predicted fuel consumption may be indicative of an ongoing fault in those aircraft, or in their operation. To validate this program, such a fault must be identified in one of the aircraft labeled as having high fuel consumption. Such an identification has not yet been attempted.

The current paper has focused on analyzing the results of this methodology and demonstrating that the modeling process works as it should, assuming it is valid. This analysis has indicated that the regression algorithms used, particularly the GLM and the Stable GP, are a good fit to predicting fuel consumption. The regression algorithms generally have normalized root mean square errors below 20%. It has been shown that the modeling approach, which uses 10-20 input variables, offers a reasonably good prediction of fuel consumption, even compared to a program using all available continuously recorded variables. Furthermore, the limited number of inputs allows the model to retain the ability to predict off-nominal pilot inputs. Based on model-building with the GLM, the most important variables for modeling fuel consumption seem to be altitude, longitudinal acceleration, ground speed, true airspeed, and low speed engine spool rpm. It has been demonstrated that even the relatively short flight segments considered have two distinct modes of fuel consumption that are easily detectable and well-predicted by the input variables. Although the current study did not utilize this fact, it may be useful for future work.

Overall, analysis has not found any mistakes in previous work, and none of the findings suggest that the application of Virtual Sensors to predict fuel consumption is unreliable. Ultimately, validating this methodology requires an inspection of the aircraft in the data set, which is ongoing. Future work will apply this technique to analyze other data sets from different airlines and aircraft.

Acknowledgments

The data set used in this study was provided by easyJet Airline Company Ltd. EasyJet is a British airline that flies customers to destinations in Europe, North Africa, and West Asia. EasyJet has grown quickly since its founding in 1995 and currently carries more passengers than any other United Kingdom-based airline. EasyJet operates 196 aircraft, most of which are Airbus 319, from 19 bases across Europe[11].[12] NASA was granted access to the data used for this study under terms of an agreement for collaborative research in which access to operational data enables NASA to develop new methodologies such as that described here. The authors are grateful to easyJet for their cooperation and support.

This project was supported by NASA's Undergraduate Student Research Program. T. Woodbury thanks Bryan Matthews and Irving Statler, both of NASA Ames, for reviewing and providing feedback on this paper.

References

[1] Airlines, S., "Southwest Airlines 2010 One Report: Financial Summary," `http://www.southwestonereport.com/financial-10yr.php`, Accessed July 17, 2011.

[2] Srivastava, A. N., "Greener Aviation with Virtual Sensors: A Case Study," *Data Mining and Knowledge Discovery*, Vol. 24, No. 2, 2012, pp. 443–471.

[3] Srivastava, A. N., Oza, N. C., and Stroeve, J., "Virtual Sensors: Using Data Mining Techniques to Efficiently Estimate Remote Sensing Spectra," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 43, No. 3, 2005, pp. 443–471.

[4] Matthews, B. L. and Srivastava, A. N., "Adaptive Fault Detection on Liquid Propulsion Systems with Virtual Sensors: Algorithms and Architectures," JANNAF Joint Propulsion Meeting, 2010.

[5] NASA, "Discovery and Systems Health," `http://ti.arc.nasa.gov/tech/dash`, Accessed August 10, 2011.

[6] Myers, R. H., Montgomery, D. C., and Vining, G. G., *Generalized Linear Models*, John Wiley & Sons, Inc., 2002.

[7] McCullagh, P. and Nelder, J., *Generalized Linear Models*, Chapman&Hall, 2nd ed., 1989.

[8] Ebden, M., "Gaussian Processes for Regression: A Quick Introduction," `http://www.robots.ox.ac.uk/~mebden/reports/GPtutorial.pdf`, Accessed July 17, 2011.

[9] Foster, L., Waagen, A., and Aijaz, N., "Stable and Efficient Gaussian Process Calculations," *Journal of Machine Learning Research*, Vol. 10, 2009, pp. 857–882.

[10] Michie, D., Spiegelhalter, D., and Taylor, C., *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, 1994.

[11] "easyJet," `http://www.theairdb.com/airline/easyjet.html`, Accessed August 24, 2011.

[12] "2010 Annual Report," `http://2010annualreport.easyjet.com/`, Accessed August 24, 2011.