SECOND GENERATION VIRTUAL SENSOR FOR HELICOPTER GROSS WEIGHT PREDICTION

Miguel A. Morales
CSC
mmorales@csc.com

David J. Haas
NSWC Carderock
david.haas@navy.mil

Brian L. Fuller
NSWC Carderock
brian.fuller@navy.mil

ABSTRACT

A virtual sensor for the estimation of aircraft gross weight in a large naval helicopter has been developed and implemented as part of the post flight analysis module for this aircraft. The sensor is a second generation neural network model capable of predicting gross weight at the time of takeoff, as well as producing a continuous estimate of the gross weight throughout the duration of a flight based on values of fuel quantity recorded by the Health and Usage Monitoring System (HUMS) onboard the aircraft. As a result, changes in gross weight due to fuel burn and refueling are taken into account throughout the flight. In addition to fuel, aircraft empty weight, cargo and crew constitute the remaining components that define total aircraft gross weight. Because the virtual sensor developed in this effort performs estimations of the gross weight as the aircraft takes off, changes in cargo and crew that occur on the ground can be readily accounted for. As a second generation model, the domain of operation for this virtual sensor is fully defined in multi-dimensional space and its performance assessed through the application of statistical techniques for wide data.

INTRODUCTION

Aircraft gross weight is a particularly useful quantity for aircraft operations, from the determination of aircraft performance characteristics such as the ability to hover out-of-ground effect at altitude with a given payload, to the determination of aircraft structural life. However, most aircraft gross weight estimations are still performed and logged manually. As a result, they are prone to errors, as well as potential loss or misplacement. In contrast, aircraft monitoring and analysis has steadily migrated towards increased automation transitioning or discarding manual tasks. As a consequence, the need for autonomous gross weight estimation has now become a necessity. There are several proposed methods for autonomous gross weight estimation including the use of an Extended Kalman filter working in conjunction with an internal aircraft model [1], the use of corrected moment theory [2] and the use of artificial intelligence techniques. Given that dynamic models for an aircraft are difficult to obtain and validate, the selected approach is to develop an artificial neural network model capable of performing the estimations. Previously, the feasibility of utilizing this approach was demonstrated on the SH-60 aircraft [3] that later led to the deployment of a gross weight virtual sensor [4].

In this study, neural network predictive models developed through supervised learning will be used for the determination of aircraft gross weight. When utilizing this approach, it is important to keep track of the model's state of development. In particular, accuracy and reliability are key properties that must improve as the model matures. Typically, a neural network model demonstrates its capability for performing estimations based on specific training, test and validation sets, while its domain of operation is determined based on the ranges of the parameters used for its development. In this study, this type of model development is construed as a first generation model and falls short from meeting the requirements for implementation in the field. Specifically, as the multi-dimensional composition of the data changes with new situations, the accuracy of the model's predictions may also change. In addition, its reliability can also be compromised because its domain is only given in terms of ranges. A second generation model overcomes these limitations by defining a domain of operation in multi-dimensional space, as well as undergoing a much more rigorous validation that accurately characterizes model performance. In addition, a learning curve for the model can also be produced in order to demonstrate that learning has reached stability and thus has all the information necessary to perform the predictions. The intent of this effort is to develop a second generation neural network model for gross weight prediction that can be implemented as a virtual sensor on the H-53 post-flight data processing module. In this way, flight conditions where predictions take place can be well defined avoiding potential situations for errors. A detailed discussion on model development and validation is included in the next section.

DATA CHARACTERIZATION AND MODEL VALIDATION

Validating models developed through supervised learning is often difficult and in some cases may produce misleading results if the domain of operation for the model is not properly defined. The problem originates from the lack of data that is predominately the norm for applications defined in moderate to high dimensional space. As the number of dimensions increases, the possible space of operation for a model becomes exceedingly large and therefore is unlikely to be thoroughly represented by the available sample set. Yet, a common approach for defining a model's domain is obtained by simply extracting the minimum and maximum values representing the range in the available set of observations for every dimension. This is not the appropriate way to define the domain when the available data represents only a small fraction of the space. Even in cases with low dimensionality, this situation can take place and lead to the same problem. This is illustrated in Figure 1 where four clusters of observations can be seen at specific locations in a three dimensional space.

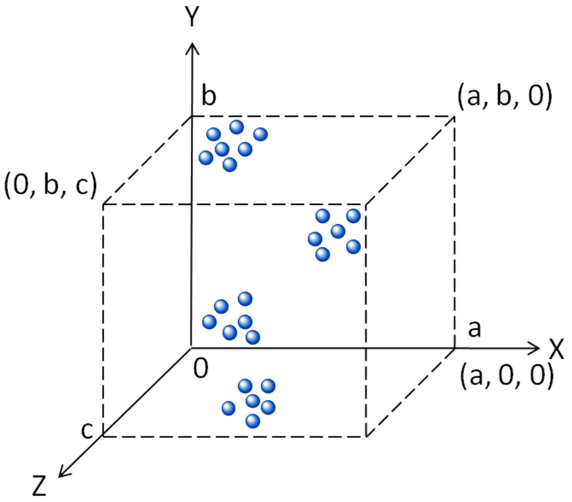

Figure 1. Sparsely populated 3-dimensional space.

If the data range were to be used to define the domain for the case depicted in Figure 1, its boundaries will be given by [0, a] in the X dimension, [0, b] in the Y dimension, and [0, c] in the Z dimension. Clearly from the figure, large spaces in the vicinity of locations [a, 0, 0], [a, b, 0] and [0, b, c] are unpopulated and thus unknown. As a result, declaring a domain for a model using the minimum and maximum values in the given set of observations has the potential to include regions where the accuracy of predictions is unknown and thus has a high likelihood of producing incorrect results. A preferable way to develop the model and define its operating domain can be achieved through the use of Design of Experiments (DOE) [5]. This approach ensures a well defined operating space that can be verified in accordance to well defined mathematical concepts. As a result, the predictions of a model developed in this form can be assigned a specific level of confidence in accordance to a selected experimental design and a selected set of verification samples. The most comprehensive type of design, a full factorial, utilizes records located at all edges of the space to develop the model, while records located mid-way between the edges are used for model validation. Figure 2 illustrates the approach in the same three dimensional space as that used in Figure 1. Based on the number of samples utilized in each of the validation locations, a prediction confidence for the model can then be calculated. For example, if five samples at the locations used for model development (i.e., experiment run locations) in Figure 2 are utilized and then verified by an additional five samples at the model's validation location, a 99.5 confidence in the predictions will result everywhere in the space, assuming that the validation samples successfully confirmed the model. As can be seen, only a moderate number of samples are required to successfully develop a three dimensional model with a high level of confidence in its predictions. However, as the number of dimensions increase, the number of required samples increases as well. This increase can, nevertheless, be significantly mitigated by using alternative designs that require only a small fraction of the samples required by the full factorial design (these are thoroughly addressed in Reference 5). Another important consideration when utilizing the DOE approach is that it assumes relationships between endpoints in the model to be linear. This is a considerable drawback given that many problems do not fall into this category. DOE, nevertheless, addresses this issue by increasing the number of levels used in model development when nonlinear relationships are expected. This solution works well for low order nonlinear effects, a characteristic applicable to most problems. However, this comes at the expense of a larger number of samples required to develop the model. The case shown in Figures 2 illustrates a two level design. A three level design can be obtained by utilizing the samples at the mid-points in Figure 2 for model development and then validating at the mid-point location. Figure 3 illustrates this scenario for a full factorial design assuming all three dimensions are expected to have nonlinear relationships.

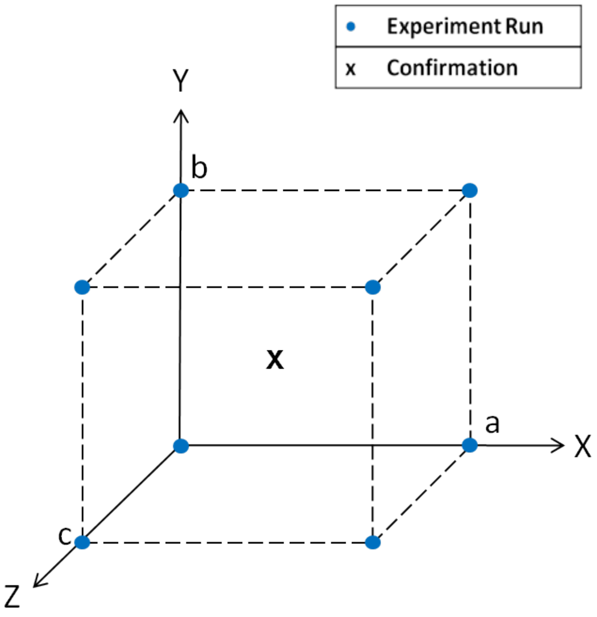

Figure 2. Design of Experiments typical approach representation in 3-dimensional space.

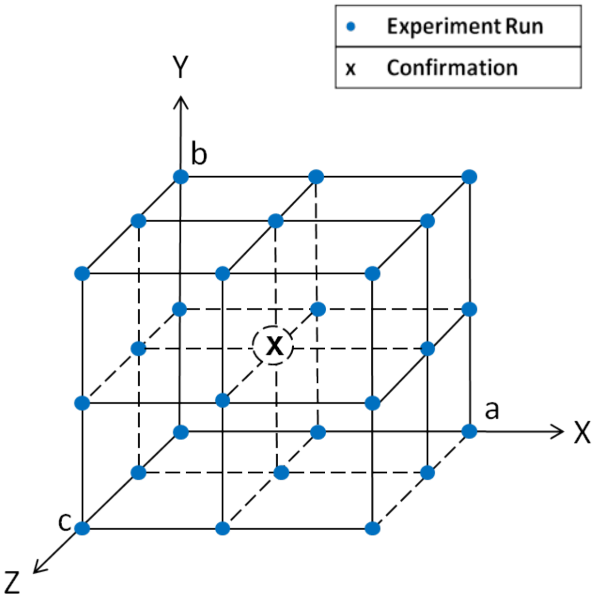

Figure 3. Design of Experiments full factorial design at three levels in 3-dimensional space.

While DOE provides a disciplined approach for the determination of a model's domain of operation, it cannot be expected to be applicable in all situations. There will be cases where the data necessary to define the various edges and mid-way locations in multi-dimensional space will not be available. Specifically, cases where the problem may be poorly understood, as well as instances where the data itself is too expensive to obtain will either force the examination of numerous variables or restrict the collection of samples. These problems are common across many business practices and in particular have been the subject of extensive study in the medical field where patient data for rare illnesses are especially hard to obtain [6]. Datasets of this type where the number of features is large compared to the number of observations are characterized as “wide.” Traditional statistical problems do not focus on wide data, but rather on “tall” data, that is, data that contain many observations per feature. As a result, most statistical tools are designed for the analysis of tall data. Nevertheless, many of these tools with some modifications can also be applied to wide data. Among the principal concerns for developing a model with wide data is to avoid over-fitting and to maximize the use of the available data. One practical approach to meet these objectives is through the use of cross-validation [7]. In this approach, the available data are first subdivided into several subsets. Next, all but one of the subsets is utilized to develop a model that is then used to predict the remaining subset. This step is repeated as many times as there are subsets so as to predict each individual subset utilizing the remainder subsets. Performing this operation ensures that prediction results are obtained for every available observation without any of them ever being utilized for model development. Typically, five subsets are utilized for cross-validation. However, there are techniques available that can determine an optimum size for the subsets if so desired.

Cross-validation effectively addresses the way to validate a model developed with wide data, but it cannot determine the model’s domain of operation. Proper characterization of the multi-dimensional space is necessary to accomplish this task, since the definition of the domain will need to be made in these terms to ensure accurate predictions. Another important concern that needs to be addressed is whether or not the available dataset completely describes the problem. A way to establish this is by extracting a random subset of the data similar in size to those used for cross-validation. Then, by utilizing several samples at a time from the remaining dataset, models are developed to predict the originally extracted subset. As more samples are used to develop the particular model, the error in the predictions is expected to decrease until it becomes stable and further addition of samples do not improve predictions. Observing this behavior is evidence that sufficient data is available to describe the problem. If the predictions, however, do not exhibit stability, then it can be concluded that more data is needed in order to properly represent the problem. Furthermore, if the stability condition is reached, a model constructed at the time stability is observed should in principle be able to predict the remaining samples at a similar level of accuracy as the originally extracted subset. It should be noted that there may still be cases not yet collected that are not represented in the dataset. However, the significance of this scenario can be dismissed by effectively defining the model’s domain of operation.

In this effort, a wide dataset is used for the development of the gross-weight virtual sensor given that a significant number of variables are evaluated before a determination is made as to their contribution, while at the same time, the number of available samples is limited. As a result, a five-fold cross-validation is used to verify the performance of every model. Once the best performing model is obtained, an operating domain is defined and enforced by constructing a multi-dimensional space identifier that acts as a pre-processor to the model. This pre-processor allows predictions only for known regions of the multi-dimensional space. In addition, the final models are also tested to determine whether they have reached stability in learning and thus the problem is properly represented. The models are further verified by an independent set of flights that became available after model completion.

MODEL DEVELOPMENT AND RESULTS

In order to successfully develop a neural network model for the prediction of aircraft gross weight, the characteristics of how this parameter can vary must first be established. Principal among these is the distinctive quality of gross weight for not having any frequency content. Gross weight is a steadily decreasing function determined only by fuel consumption in the absence of any weight-changing events. Weight-changing events include cargo and crew loading and unloading, and amount to a stepwise change in the gross weight. Only one event, refueling, produces a steady increase in the gross weight parameter. The challenge with developing an estimation model for gross weight is that the most influential parameters available for gross weight prediction are dynamic and thus have high frequency content. Unless the contribution of the various dynamic components for these signals can effectively cancel each other’s influence as to result in a steadily decreasing function, attempting to predict the gross weight parameter from dynamic signals will not generate a well conditioned model. Instead, model development should utilize the signals from flight regimes where signals are steady or can otherwise be accurately characterized by statistical measures over a period of time.

In this effort, a total of 40 flights obtained from nine H-53 aircraft conducting regular operations are utilized for model development in three distinct flight regimes. Given that the resulting virtual sensor is intended to conduct regular gross weight estimations for the H-53 fleet, both performance and frequency of prediction constitute the criteria for model success. In this section, model development is described individually for each of the three regimes investigated and the best of the resulting models is then selected. In addition, integration of this model as a virtual sensor is also discussed.

Steady Hover Regime

The first of the regimes evaluated is steady hover. This regime is the same as that utilized in Reference 3 and is selected because signals are reasonably steady in this regime. However, in this study the frequency of encountering this regime is also considered. To this end, the ten second window utilized in Reference 3 is reduced to five seconds in an effort to capture more instances of steady hover. Specifically, the steady hover condition requires that the engine torque, as well as the radar altitude remain within 3% and 3ft respectively for the duration of the time window where they are evaluated. In addition, the GPS velocity must be below 5 Knots.

When evaluating a regime for gross weight predictions, it is first necessary to determine if there are parameters possessing a strong relationship with gross weight that can serve as the inputs to the predictions. Close inspection of the steady hover regime reveals a good relationship between engine torque and gross weight. Figure 4 illustrates this relationship by displaying those instances where a steady hover is identified within a five second window for the 40 flights in the development dataset. These instances are subdivided into four distinct gross weight ranges and assigned a unique color from lowest gross weight to highest gross weight in the sample. These values correspond to 43,400 lbs and 54,300 lbs respectively. The figure displays the relationship between engine torque and gross weight for varying radar altitude since operating in-ground-effect or out-of-ground-effect will affect this relationship. However, in this example, all altitudes for steady hover correspond to in-ground-effect and the relationship is only noticeable through close observation of the bottom values for engine torque that are seen to rise as the radar altitude increases. It should also be noted that if the conditions for steady hover are maintained beyond the required five seconds, the window is expanded. As a result, each point in the graph represents a single event. There are 169 events in the flight sample that are identified in 22 distinct flights.

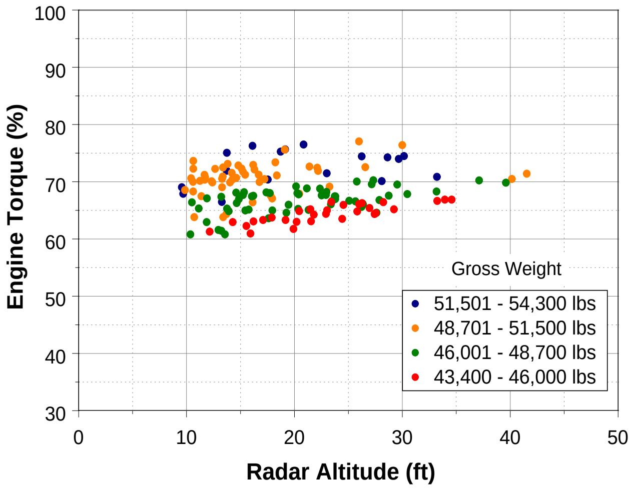

Figure 4. Relationship between engine torque and gross weight vs. radar altitude for the steady hover regime.

It can be observed from the figure above that for a given value of the radar altitude, higher values of torque correspond to higher values of gross weight. However, the frequency of prediction is significantly restricted as this maneuver could only be detected in approximately half of the flights. As a result, this regime is discarded for virtual sensor development based on its low frequency of prediction. Yet, for purposes of comparison, a model utilizing the same inputs as those found in Reference 3 was constructed and obtained a Root Mean Squared (RMS) error of 1,230 lbs.

High Speed Steady Level Flight Regime

The second flight regime considered for gross weight is high speed steady level flight. In this case, the relationship between gross weight and engine torque for different values of airspeed is investigated. Figure 5 illustrates this relationship for all instances meeting the conditions for steady level flight in a 10 second window.

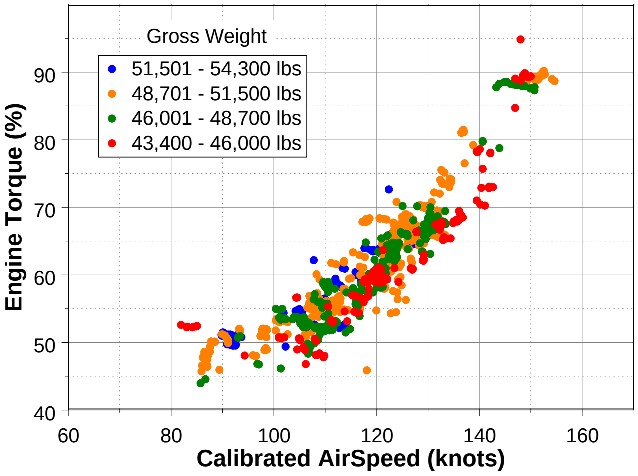

Figure 5. Relationship between the engine torque and gross weight as a function of calibrated airspeed for high speed steady level flight.

As in Figure 4, gross weight is subdivided into four distinct ranges for clarity. However, in this case, multiple instances of the same event are displayed if the conditions for the 10 second window are met as the window moves forward in time. As can be observed in the figure, no clear increase of the gross weight for higher values of engine torque is detected for any given value of the airspeed. Despite this result, prediction of the gross weight is still attempted assuming additional parameters significantly influence the relationship. Model development is performed by constructing windows of different durations where statistics of selected signals are calculated to serve as inputs to the model. Window durations of 10, 8 and 7 seconds are constructed. In each window, average values of engine torque, calibrated airspeed (CAS), longitudinal stick position, speed of sound, pitch attitude and density altitude are calculated and utilized as inputs to the models after careful evaluation of model response. The 10 second window model showed the best training results and attained a RMS error of 604 lbs, followed by the 8 second window model with 982 lbs RMS error and the 7 second window model with 1,268 lbs RMS error. Although longer window times improve accuracy, as the length of the window increases, the model becomes more restrictive and the frequency of prediction decreases. The 7 second window model is able to perform predictions in 39 of the 40 flights, the 8 second window model is able to perform predictions in 34 flights and the 10 second window model is only able to perform predictions in 22 flights (just over half of the total). However, in addition to the drop in the frequency of prediction, all models in this regime exhibited poor performance in the cross-validation tests by producing RMS errors in the 2,500 lbs range, a clear indication that the models memorized the training data set. Attempts to generalize the models by decreasing the number of nodes did not improve predictions. As a result, this regime was not selected for gross weight prediction.

Ascents from Takeoff Regime

The third flight regime evaluated is defined from the time of takeoff until the aircraft reaches a specific altitude. In this case, the intent is to measure the energy required to hover for a given time period in order to average the effects of fluctuations in torque, altitude and other parameters that are prevalent during steady hover conditions. Most importantly, this quantity is directly related to the aircraft gross weight, density altitude, radar altitude (i.e., in-ground-effect) and aircraft trim and can be expressed as,

$$E_{\mathrm{Q}} = \sum_{i=2}^{T} \mathrm{Q}_i \,(\mathrm{t}_i - \mathrm{t}_{i-1}) \qquad (1)$$

where $\mathrm{E_Q}$ represents the energy term for a time period T (i.e., engine torque integral), t is time and Qi is the engine torque at a given time sample i. By defining the regime in this way, the aircraft is required to go through a fixed altitude range that maintains a consistent nominal in-ground-effect, and thus accounts for altitude variation. Density altitude effects are accounted for directly by utilizing density altitude as an input to the model, and trim is accounted for by utilizing the pitch attitude as another input. It is further assumed that this energy term is the main component of the work performed by the engines and that changes in potential energy, and kinetic energy are not significant (calibrated airspeed is restricted to less than 41 knots throughout the duration of the ascent for predictions to take place). Utilizing this approach generates a steady quantity that can more effectively be used to predict the steady gross weight function. Ascents from takeoff to three distinct altitudes are considered. Altitudes of 15ft, 30ft and 50ft are chosen because ascents from takeoff are commonly performed to different altitudes. As the ascent altitude increases, the number of ascents reaching these higher altitudes decreases. Therefore, given that one of the measures of model success includes frequency of prediction, the altitudes are kept relatively low to ensure ascents will either reach or go beyond the prescribed altitude. As expected, most events are found at 15ft, followed by 30ft and 50ft. It should also be noted that if the prescribed altitude is exceeded, the interval considered ends when the aircraft reaches that altitude.

The integral of engine torque is examined to determine if it has a strong relationship with gross weight. The results are shown in figures 6 through 8. As in figures 4 and 5, the gross weight is subdivided into four groups to highlight the gross weight relationship. However, in this case, every point represents a unique ascent from takeoff so there are no repetitions. Ascents are limited to nominal takeoffs (i.e., those that do not exhibit extreme values of the controls) and are accounted for by defining an operating domain.

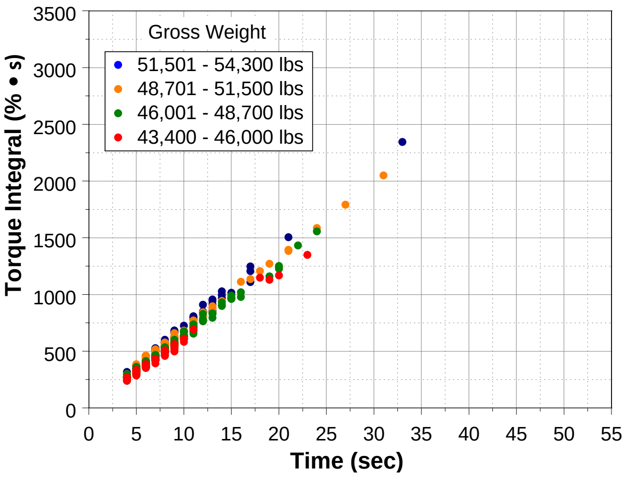

Figure 6. Relationship between the engine torque integral and gross weight as a function of duration for ascents from takeoff to 15ft.

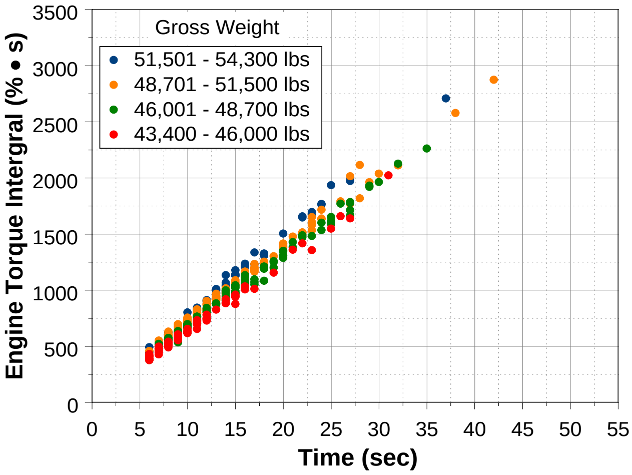

Figure 7. Relationship between the engine torque integral and gross weight as a function of duration for ascents from takeoff to 30ft.

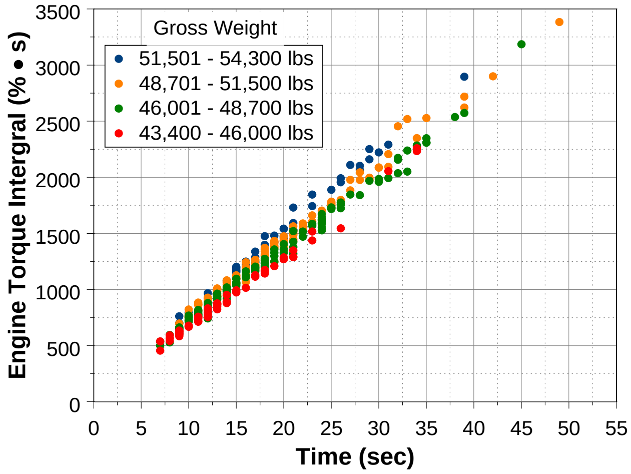

Figure 8. Relationship between the engine torque integral and gross weight as a function of duration for ascents from takeoffs to 50ft.

As can be observed in figures 6 through 8, the relationship between gross weight and torque integral as a function of ascent duration is well behaved and shows higher gross weight as the values of torque integral increases for all ascent altitudes. As a result, it can be concluded that there is a good relationship between gross weight and the integral of torque in this flight regime. However, as previously indicated, the torque integral and ascent duration are not the only parameters that have a physical relationship with gross weight. Density altitude, as well as aircraft trim are also important. Density altitude can easily be incorporated into the model as an input, but all the information necessary to determine aircraft trim, such as c.g. position are not available. Although this is a limitation, partial information for aircraft trim can still be derived from pitch attitude. It should be noted that density altitude exhibits little or no variation between zero and fifty feet, thus an average value is sufficient to define it. However, the pitch attitude is constantly changing throughout the ascent. It is for this reason that pitch attitude is incorporated into the model by obtaining the torque integral components at every instant through the ascent as defined by,

$$E_{Q1} = \sum_{i=2}^{T} \mathrm{Q}_i \cos \theta_i \, (\mathrm{t}_i - \mathrm{t}_{i-1}) \qquad (2)$$

$$E_{Q2} = \sum_{i=2}^{T} \mathrm{Q}_i \, \sin \theta_i \, (\mathrm{t}_i - \mathrm{t}_{i-1}) \qquad (3)$$

where E_{Q1} and E_{Q2} are the components of the integrated engine torque and θ is the pitch attitude. These two quantities in addition to average density altitude and ascent duration are referred to as the baseline input set.

The model architecture selected to perform predictions is a back-propagation neural network with two layers composed of eight nodes in the first layer and four nodes in the second layer. This architecture is kept constant throughout the experiments in order to isolate the influence of each input parameter. For each experiment, a five-fold cross-validation is applied to assess model performance. This requires the main dataset to be broken into five distinct data subsets. Four of these are used to train the model to predict the remaining subset. This is done once for each subset. As a result, each experiment requires the development of five distinct models, each predicting a unique subset. The results reported are only those where the subset is not used for model development and thus are reflective of data that has never been seen by the models. It is also important to note that while each subset has approximately the same number of events, these are assigned to a subset on a flight-by-flight basis, that is, only whole flights are added to an individual subset at a time. In this way, information for the fundamental flight unit never appears in more than one subset and thus preserves the intent of cross-validation.

In an attempt to improve model performance, several parameters in addition to those in the baseline input set are considered one at a time. However, to avoid numerous experiments that are unnecessary from a DOE perspective, only ascents to 50 ft are utilized for this purpose since they constitute the upper edge of the altitudes considered. The results are shown in Figure 9. Only the RMS error averages for all models involved in cross-validation for each experiment are displayed.

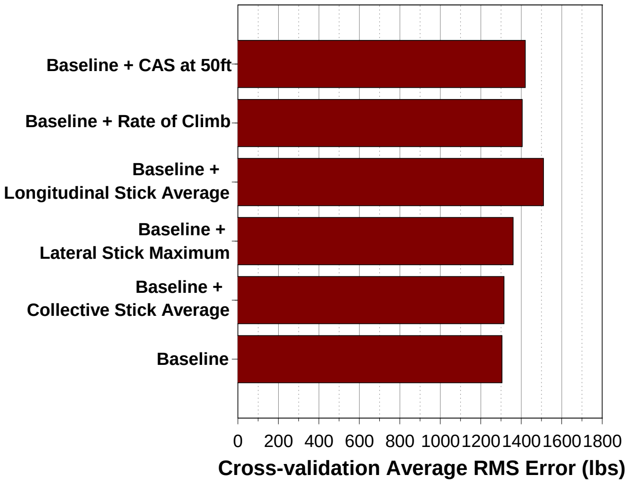

Figure 9. Gross weight prediction model performance for different model inputs for ascents from takeoff to 50ft.

As can be seen in the figure, no improvements are obtained by including additional inputs. As a result, the baseline input set is selected for predictions. Only one additional modification to the definition of the regime is implemented. Specifically, the start of the regime is set at 2 ft in order to unambiguously determine the start of the aircraft's ascent. This modification proved to be of significant benefit to the predictions.

Having defined the input parameter set, the next step is to determine where predictions show the best performance and are the most reliable. To this end, models are developed for ascents to 15ft and 30ft in addition to the ones already available for 50ft. Results for all models for the 40 flight set in terms of RMS error, maximum over and under-predictions and number of events predicted are given in Table1. At least one event is predicted in all 40 flights for ascents to 15ft and 30ft, while predictions are performed in 39 flights for ascents to 50ft. In addition, a comparison between actual and predicted gross weight for each regime is shown in figures 10 to 12. Since maximum errors are desired to be within ±2,000 lbs, error boundary lines have been included in the figures for reference.

Table 1. Gross weight prediction performance for ascents from takeoff to various altitudes.

Regime	Events Predicted	RMS Error (lbs)	Maximum	
			Over-prediction (lbs)	Under-prediction (lbs)
Ascents to 15ft	388	1,181	3,428	5,121
Ascents to 30ft	372	1,003	2,871	2,838
Ascents to 50ft	338	1,048	2,783	3,226

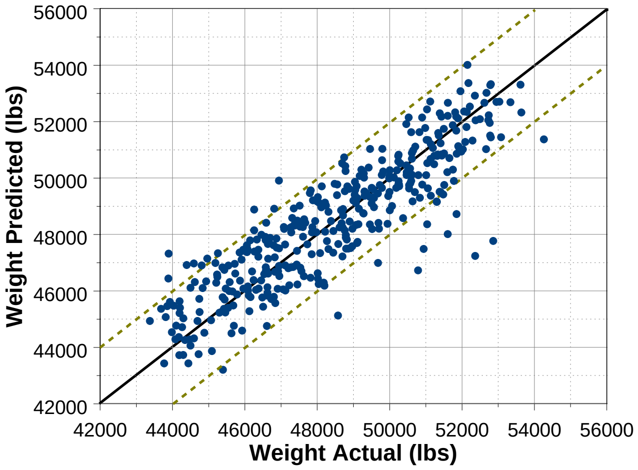

Figure 10. Baseline input set gross weight prediction performance for ascents from takeoff to 15ft.

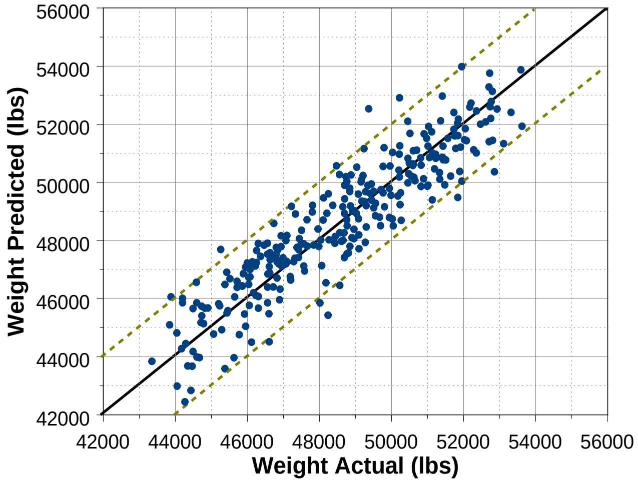

Figure 11. Baseline input set gross weight prediction performance for ascents from takeoff to 30ft.

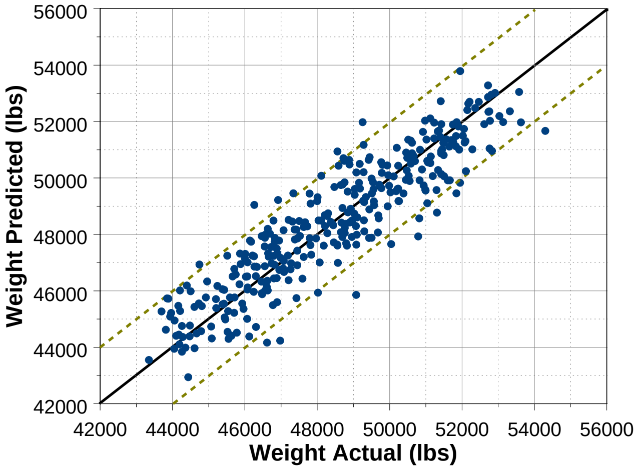

Figure 12. Baseline input set gross weight prediction performance for ascents from takeoff to 50ft.

As can be seen in the figures, while ascents to 15ft contain the largest number of events, there is a noticeable number of significant outliers in the predictions. As a result, this regime is not sufficiently reliable. In contrast, ascents to 30ft still contain a significant number of events and demonstrate better performance with only a few minor outliers. Results from this regime are also better in terms of performance and frequency of prediction than those obtained for the 50ft regime. As a result, it is the ascents to 30ft regime that is selected for final virtual sensor development. It is speculated that the lower accuracy encountered in the 15ft regime is due to the fact that small errors originating from signal sampling frequency have a more significant effect for shorter ascents, while errors in the 50ft regime are comparable to those in the 30ft regime, but capture fewer events. As a result, a compromise between the two regimes at the edges produces the best results.

At this point, it becomes necessary to determine if the current dataset is sufficient for model development. As previously discussed, this can be determined by selecting a random set of events that is then used to validate models that progressively utilize more events. The RMS error for each model can then be plotted as a function of the number of samples to produce a learning curve. If the curve is seen to reach stability (i.e., its slope approaches zero), it implies there is sufficient data to represent the problem within its operating domain. The learning curve for the selected model type is shown in Figure 13. Forty nine events from the 362 event set in the ascents to 30ft regime are chosen as the validation set.

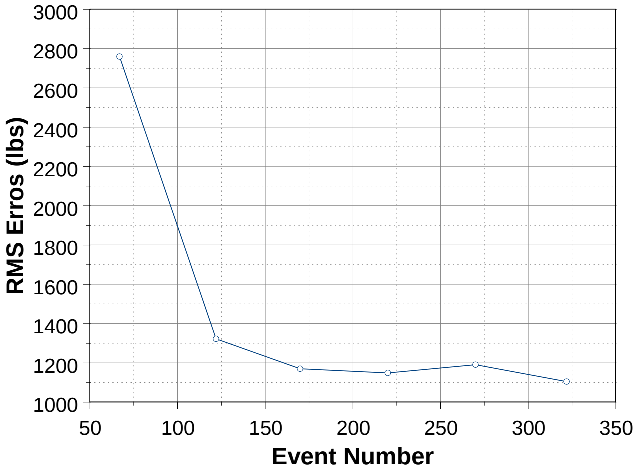

Figure 13. Learning curve for final model.

As can be seen in the figure, the gradient of the curve is low enough to declare the model is reasonably stable and thus is properly characterized. As such, the model provides a good solution to the problem.

Virtual Sensor Integration

The last step necessary to complete the virtual sensor is the preparation of a pre-processor that will enforce its operating domain. This is obtained by defining confidence regions in multi-dimensional space based on the available data. Initially, small conservative boundaries for the regions are selected, but these are carefully extended by modifying each input by a given percent of its range. With each modification, results are observed to see how much they affect the original predictions. If a small effect is observed, the boundary for the parameter being varied is increased in all regions. If, however, a large effect is observed no change in the boundary is conducted. After all parameters have been varied throughout their ranges and the boundaries adjusted for all confidence regions, the pre-processor defining the model's operating domain in multi-dimensional space is finalized.

Because development of the virtual sensor is conducted by utilizing five-fold cross-validation, every experiment produces five distinct models. No one model in this set is necessarily better than the other ones since distinct data is utilized for its development. As a result, in the final virtual sensor implementation, all five models are used and the average of their predictions constitutes the reported gross weight predicted value. A comparison of the actual vs. predicted gross weight as reported by the virtual sensor for the entire development dataset is shown in Figure 14.

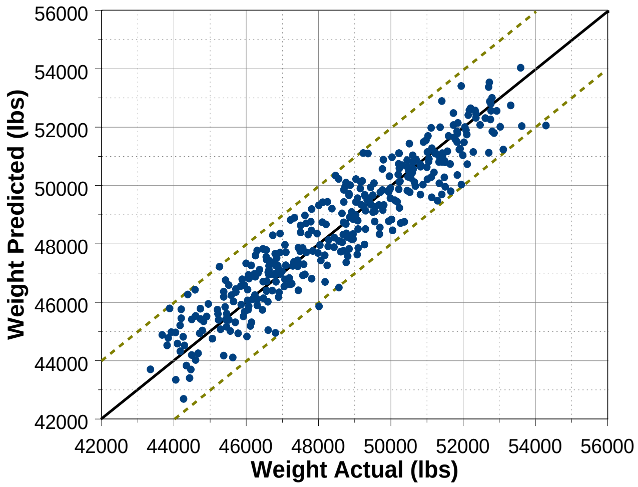

Figure 14. Virtual sensor performance on development dataset.

As can be seen in the figure, the performance of the virtual sensor is noticeably better than that reported for cross-validation shown in Figure 11. The maximum over and under-predictions are 1,900 lbs and 2,229 lbs respectively and the RMS error is only 831 lbs. This is expected since four out of the five models for any one event utilize the event during their development. Therefore, these results should be viewed as the ones to expect when events are very similar to those used during development. In contrast, the results from cross-validation are those to be expected for events that have never been seen before by the models, but are still within its operating range.

After model completion, an additional eight flights with complete gross weight logs became available and were utilized for virtual sensor validation. Results are shown in Figure 15.

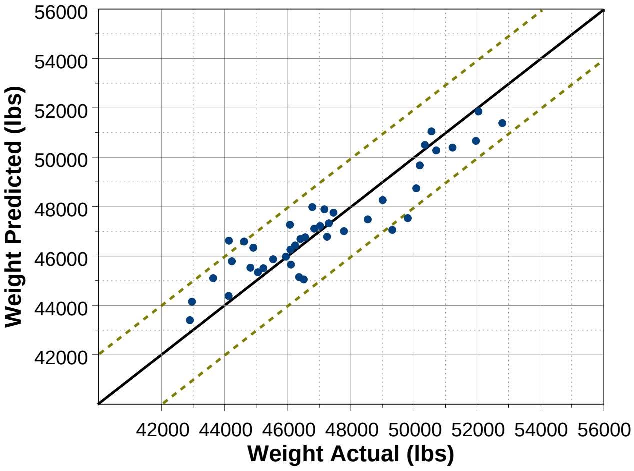

Figure 15. Virtual sensor performance on 8 flight validation set.

As can be seen in the figure, the results conform to expectations as the maximum over and under predictions are within the cross-validation values calculated for the virtual sensor and the RMS error is 1,052 lbs and is within 49 lbs of the calculated cross-validation value. In addition, the virtual sensor correctly calculates the gross weight below its development range in two distinct instances.

In its final implementation as part of the post analysis module for the H-53, the gross weight virtual sensor also utilizes the fuel recorded history from the HUMS to recreate the gross weight throughout the entire flight. A fuel reconstruction algorithm is implemented as a pre-processor to the fuel signal in an attempt to minimize the time when fuel signals are unavailable. A gross weight reconstruction time history for one of the flights in the validation set is shown in Figure 16.

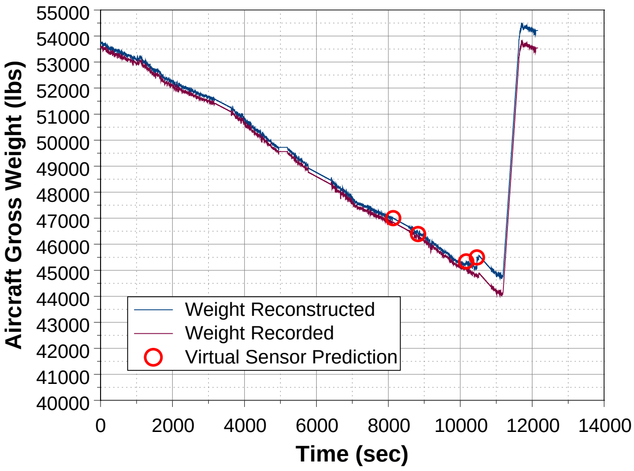

Figure 16. Gross weight reconstruction for a flight in the validation dataset.

A particularly important function that the current implementation of the gross weight virtual sensor is not able to perform is to account for sling load pick-up and drop. This is because the helicopter needs to hover at a prescribed altitude above ground to perform this operation, and therefore, does not meet the selected prediction regime requirements. Nevertheless, sling load pick-up and drop patterns can be identified by autonomous procedures. Therefore, in principle, a similar regime as that currently used to perform gross weight predictions could be defined from higher start and end altitudes to account for sling loads. This is expected to be part of a future effort.

SUMMARY AND CONCLUSIONS

A second generation virtual sensor for the prediction of aircraft gross weight in the H-53 helicopter was successfully developed. The sensor was designed to perform predictions for ascents from takeoff to 30ft and then reconstruct aircraft gross weight throughout a flight with the help of fuel signals monitored by the HUMS. The operating domain for the sensor was fully defined in multi-dimensional space in order to ensure accuracy and reliability through a pre-processor. A nominal takeoff was also defined and implemented as part of the pre-processor. The accuracy for the virtual sensor was calculated at 1,003 lbs of RMS error for a range of gross weights between 43,400 lbs and 54,300 lbs. Maximum over and under-estimations on a 40 flight set were found to be 2,871 lbs and 2,838 lbs respectively (under three times the calculated RMS error expected for a normal distribution). A frequency of estimation of 78% was attained for the takeoffs in the dataset (i.e., more than three out of every four takeoffs were predicted). Yet, all 40 flights in the sample were predicted in at least one instance because some flights contained multiple takeoffs, thus a 100% frequency of prediction on a per flight basis was achieved.

It was also found that the steady hover regime does not generate frequent enough opportunities for prediction and as a result was not selected. Similarly, attempts at predicting gross weight during high speed steady level flight resulted in unreliable predictions as determined through cross-validation. The initial model performance assessment in training demonstrated promising results, but then performed poorly when cross-validated. This is a significant concern since it indicates that there are trends in the data that can minimize the difference between the actual gross weight and the predictions. However, these trends were not quantifiable. As a result, the high speed level flight regime was not selected for gross weight prediction. In contrast to these regimes, the prediction approach applied to ascents from takeoff to a prescribed altitude was found to be the most reliable and showed consistent results at three different prescribed altitudes. At the same time, it was found that the number of opportunities for prediction is significantly greater than the steady hover or high speed level flight regimes. In addition, the frequency of prediction only decreases by a small amount as the prescribed ascent altitude increases. As a result, the prediction approach applied to ascents from takeoff was found to be the most useful for gross weight prediction. In particular, it was found that ascents to 15ft are too short to properly measure the energy expended during the ascent, while ascents to 50ft had a reduced frequency of prediction. Therefore, the ascents from takeoff to 30ft were found to constitute the best regime for predictions.

Given that the virtual sensor for gross weight prediction developed in this effort is a second generation model verified through statistical measures applicable to wide data, it is expected that its accuracy and frequency of prediction will conform to the calculated values for RMS error and flight prediction frequency when processing flights outside its development set. An initial opportunity to demonstrate this capability became available through the identification of eight flights with complete gross weight log records after the gross weight virtual sensor was completed. Gross weight prediction for these flights returned 1,052 lbs of RMS error with one or more predictions taking place for every flight (100% flight prediction frequency). As a result, by independently verifying the performance values for the virtual sensor, these results provide significant evidence that the virtual sensor will perform as intended in the flight post-processing module of the H-53 aircraft.

REFERENCES

[1] Abraham, M. and Costello, M., “In-Flight Estimation of Helicopter Gross Weight and Mass Center Location,” *Journal of Aircraft*, Vol. 46, No. 3, May-June 2009.

[2] Teal, R., Evernham, J., Larchuk, T., Miller, D., Marquith, D., White, F. and Diebler, D., “Regime Recognition for MH-47E Structural Usage Monitoring,” Proceedings of the American Helicopter Society 53rd Annual Forum, 1997, pp. 1267-1284.

[3] Morales, M. and Haas, D., “Feasibility of Aircraft Gross Weight Estimation Using Artificial Neural Networks,” Proceedings of the American Helicopter Society 57th Annual Forum, 2001, pp. 1872-1880.

[4]Ben-Zeev, O., Haas, D.J., and Morales, M.A.,“A Virtual Sensor System for Helicopter Usage Monitoring,” Proceedings of the American Helicopter Society 58th Annual Forum, Montreal, Canada, June 11-13, 2002

[5] Schmidt, S. R., Launsby, R.G., *Understanding Industrial Designed Experiments*, Air Academy Press & Associates, Colorado Springs, CO, 2005, Chapter 4.

[6] Rabinowitz et al., “Use of L1 Norm for Selection of Sparse Parameters Sets that Accurately Predict Drug Response Phenotype from Vival Genetic Sequences,” *Bioinformatics*, October 2005.

[7] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning: Data Mining, Interference, and Prediction*, Springer Publishing, New York, NY, 2001, pp. 214-216.