Journal of Air Transportation Vol. 12, No. 1 -2007

DATA MINING METHODS APPLIED TO FLIGHT OPERATIONS QUALITY ASSURANCE DATA: A COMPARISON TO STANDARD STATISTICAL METHODS

Alan J. Stolzer
Saint Louis University
St. Louis, MO

Carl Halford
Saint Louis University
St. Louis, MO

ABSTRACT

In a previous study, multiple regression techniques were applied to Flight Operations Quality Assurance-derived data to develop parsimonious model(s) for fuel consumption on the Boeing 757 airplane. The present study examined several data mining algorithms, including neural networks, on the fuel consumption problem and compared them to the multiple regression results obtained earlier. Using regression methods, parsimonious models were obtained that explained approximately 85% of the variation in fuel flow. In general data mining methods were more effective in predicting fuel consumption. Classification and Regression Tree methods reported correlation coefficients of .91 to .92, and General Linear Models and Multilayer Perceptron neural networks reported correlation coefficients of about .99. These data mining models show great promise for use in further examining large FOQA databases for operational and safety improvements.

Alan Stolzer holds a Ph.D. in Technology Management—Quality Systems from Indiana State University. He holds FAA Airline Transport Pilot, Certified Flight Instructor, and Aircraft Mechanic certificates. He holds Quality Engineering, Manager, and Auditor certifications from ASQ, and a Project Manager certification from PMI. He is the Director of the Quality Engineering Systems in Transportation (QUEST) Research Center.

Carl Halford is the Research Manager at SLU's Center for Quality Engineering Systems in Transportation. Mr. Halford holds an Airline Transport Pilot certificate, with an assortment of type ratings, and has over 9000 hours of flight time in a variety of aircraft. Halford is a certified Manager of Quality/Organizational Excellence and Quality Auditor, and has been a Certified ISO 9001:2000 Lead Auditor.

INTRODUCTION

One might wonder what mining the genome, re-engineering the immigration system, and ensuring our homeland security have in common. The answer is data mining (DM).

Unlocking the secrets of the human gene is expected to yield great benefits for scientists and pharmaceutical companies battling diseases. But cataloging the estimated 100,000 human genes is no small task. Consider the fact that every human cell has 23 pairs of chromosomes containing about 3.5 billion pairs of nucleotides. The genes that carry code to make protein amount to less than 3% of all genes; the remaining 97% is genetic noise. These protein-producing genes are those that result in cancer and genetic problems when they go awry, and it is these genes that need to be understood by scientists. Unfortunately, the signals in the genes have a language all their own, and they are hidden and noisy. Among the tools used to analyze these signals is a form of DM called artificial neural networks. Neural networks help scientists locate the genes of interest through pattern recognition and understand their function—knowledge which may lead to breakthroughs in combating these health crises (Regalado, 1999).

DM is also playing a role in our efforts to control the immigration problem and ensure our homeland security. All 19 hijackers involved in the attacks on the U.S. on September 11, 2001, entered the country legally. There was no information available to the authorities that would have suggested that allowing them to enter the country was inconsistent with our national security interests. Strickland and Willard (2002) assert that effective, preventive homeland security requires a fundamental re-engineering of the immigration system based on the concept of having better information achieved through effective DM methods and processes to assure quality information. These authors propose a vastly improved system of 'knowledge development tools' to mine new data sources and identify visa applicants that warrant attention.

DM has been gaining popularity in numerous other industries in recent years, including the transportation industry. Studies of DM methods to improve traffic safety programs (Solomon, Nguyen, Liebowitz, & Agresti, 2006), applying DM techniques to forecast the number of airline passengers in Saudi Arabia (BaFail, 2004), and many others, are evidenced in the literature. Many of these studies seek to make greater use of existing databases to learn more about the problem or issue at hand than more traditional methods have afforded, or to discover what results DM methods might yield on previously performed studies. The present study seeks to do the latter using Stolzer's (2003) work to create a statistical model for predicting fuel consumption on the Boeing 757 aircraft fleet within an air carrier's operating environment.

PURPOSE OF THE STUDY

This study uses the comprehensive suite of DM tools contained in StatSoft's *STATISTICA* (2003) software to create models for predicting fuel consumption, and compares the results to those of a previous study. The earlier study developed parsimonious models for fuel consumption using multiple regression analysis to analyze Flight Operations Quality Assurance (FOQA)-derived data, with the objective of being able to identify outliers (specific flights) with respect to fuel consumption. Specifically, the goal of the present study was to ascertain whether DM methods produce fuel consumption models with superior predictive capability than traditional statistical methods such as multiple regression techniques. To accomplish this goal, we evaluated and benchmarked the results of the different DM methods offered within *STATISTICA*; and determined the optimum DM method.

BACKGROUND

What is data mining?

Data mining is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables (StatSoft, 2003). It is used for such broad areas as accurately evaluating insurance risk, predicting customer demand for goods and services, predicting the prices of stocks and commodities, monitoring expensive and critical equipment, conducting yield analysis and quality control, and predicting credit risk.

Traditional statistical techniques are not as useful on very large databases because all mean comparisons are significant and standard measures of variability are extremely small. Due in part to this limitation, DM techniques increased in popularity in the mid to late 1990s. DM tools are based on standard statistical techniques and artificial intelligence analysis techniques, and are applied to large databases for the purpose of teasing out otherwise undiscovered data attributes, trends and patterns. There are numerous methods of DM; the following is only the most cursory overview of several of the more popular methods.

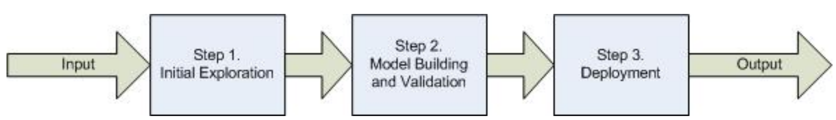
1. *Regression modeling* normally begins with a hypothesis which is tested by this common statistical technique. Linear regression (commonly used for prediction) and logistic regression (used for estimating probabilities of events) are two examples of regression modeling.

2. *Visualization* is an important concept in DM. Through the study of multidimensional graphs the analysis is able to detect trends, patterns, or relationships.
3. *Cluster analysis* is an exploratory data analysis tool that consists of several different algorithms and methods for grouping objects of similar kind into respective categories. The goal of cluster analysis is to sort different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal if they do not. Cluster analysis can be used to discover structures in data without explaining why they exist.
4. *Decision trees* are very popular classification models. They are called decision trees because the resulting model is presented in the form of a tree structure. The visual presentation makes the decision tree model very easy to understand. Decision tree methods include Classification and Regression Trees (C&RT) and Chi-squared Automatic Interaction Detection (CHAID).
5. *Neural networks* are analytic techniques that are intended to simulate cognitive functions. These techniques learn with each iteration through the data, and are capable of predicting new observations (on specific variables) from other observations (on the same or other variables).

Steps in DM

There are three basic stages to most DM projects, as depicted in Figure 1: initial exploration; model building and validation; and deployment. Initial exploration refers to the preparation of the data, which may include cleaning of the data, data transformations, selecting subsets of records, and performing feature selection operations. Model building and validation involves evaluating various models for predictive performance and choosing the most appropriate one for the project. Deployment refers to the application of the chosen model or models to generate predictions or estimates of the outcome.

Figure 1. Steps in Data Mining



Crucial concepts in DM

Of course, not all projects are the same and few involve the full range of DM tools and methods, but some familiarity with the crucial concepts in DM is important. These concepts are summarized below (StatSoft, 2003; Wang, 2003).

1. *Data preparation, cleaning, and transformation*. Many times this is the most time-consuming aspect of the project, and one that is often given little attention. Data that is collected via an automatic process, which probably includes most input data in DM projects, frequently contains data that contain out of range values, impossible data combinations, and other irregularities. Various methods are employed to clean the data to make it usable, or to eliminate the data from the analysis.
2. *Feature selection*. A feature selection technique enables the analyst to include the best variables for the project when the data set includes more variables than can be reasonably used.
3. *Feature extraction*. Feature extraction techniques attempt to aggregate the predictors in some way in order to extract the common information contained in them that is most useful for model building. Typical methods include Factor Analysis and Principal Components Analysis, Multidimensional Scaling, Partial Least Squares methods, and others.
4. *Predictive DM*. This type of DM project is intended to develop statistical or neural network models that can be used to predict objects of interest.
5. *Sampling, training, and testing (hold-out) samples*. In most DM projects, only a randomly chosen subset of the data is used. This enables the analyst to evaluate multiple methods using different samples, and then test these methods to gain insight into the predictive capability of the results.
6. *Over-sampling particular strata to over-represent rare events (stratified sampling).* Sometimes it is necessary to employ stratified sampling to systematically over-sample rare events of interest. This precludes predictions of a no response for all cases if simple random sampling were used when, in fact, these (rare) events are present.
7. *Machine learning*. Machine learning refers to the application of generic model-fitting or classification algorithms for predictive DM, and reminds us that the emphasis in DM is *accuracy* of prediction rather than having a clear and interpretable understanding of the prediction.
8. *Deployment*. Deployment is the application of a trained model so that predictions can be obtained for new data.

STATISTICA

STATISTICA, a suite of analytic software products produced by StatSoft (2003), was used for this study. *STATISTICA* provides a comprehensive array of data analysis, data management, data visualization, and DM procedures. Its techniques include a wide selection of predictive modeling, clustering, classification, and exploratory techniques in a single software platform. *STATISTICA* includes an extensive array of analytic, graphical, and data management functions, as well as DM and machine learning algorithms, including: support vector machines, EM (Expectation Maximization) and k-Means clustering, CART, generalized additive models, independent component analysis, stochastic gradient boosted trees, ensembles of neural networks, automatic feature selection, MARSplines (Multivariate Adaptive Regression Splines), CHAID trees, nearest neighbor methods, association rules, random forests, and others (StatSoft, 2003).

Articles/studies on DM for airline safety

Today DM techniques are used for many different purposes in many industries, including the aviation industry. For example, an exploratory study on FOQA database at a major air carrier took place in 2005 (Global Aviation Information Network, 2005). The cooperative study involved the air carrier, the Federal Aviation Administration (FAA), the Global Aviation Information Network, and a DM software provider, and was intended to provide guidance on tools that may be useful in enhancing the current analysis of airline digital flight data. This study focused on principal components analysis, correlation of different events, conditional (Trellis) graphics, tree-based models, and neural networks. In part, the DM study found that certain methods showed promise in improving efficiency by automating some of the query and output process. Principal components analysis and clustering methods were deemed helpful for data reduction and characterization of correlation structures. Tree-based models provided a modeling structure for understanding the relationship between flight events and flight parameters, and for assessing the importance of variables. Neural network models were deemed less useful due to their inability to distinguish between landing approaches that resulted in a successful landing from those that resulted in a go around. The study also noted an additional disadvantage that neural networks are more difficult to interpret than tree-based models.

Another similar study funded by the FAA involved the analysis of FOQA data on the airline's Boeing 777 and 747 fleets. The objective of this study was to determine whether DM techniques can help improve airline or system safety by identifying risks, and assess the effectiveness of operational changes. Three learning algorithms, that is, decision trees, clustering and association rules, were applied to the data. In general, the DM tools identified many interesting patterns and associations beneath the surface that had not been identified by the air carrier's flight data monitoring program (Global Aviation Information Network, 2004).

Helicopter health and usage management systems also generate large amounts of data that are used mainly for diagnostic purposes to detect helicopter faults. An initiative by the Ministry of Defense in the United Kingdom has been to apply tools that improved analysis capability, increase levels of automation, and provide enhanced use of resources. The study evaluated several supervised and unsupervised methods, and also explored fusing the results of unsupervised techniques with the judgments of other mathematical and artificial intelligence tools, such as logic, fuzzy logic, and Bayesian networks (Knight, Cook, & Azzam, 2005).

PREVIOUS STUDY

Our previous study was designed to develop a parsimonious model(s) for fuel consumption using multiple regression analysis to analyze FOQA-derived data, with the objective of being able to identify outliers (specific flights) with respect to fuel consumption (Stolzer, 2003). The data used for the study were provided by a major air carrier, and consisted of 1,863 routine passenger-carrying flights on Boeing 757 aircraft.

Depending on the aircraft involved, data is captured on a few dozen to thousands of parameters (e.g., altitude, airspeed, throttle position, aileron deflection) each second; more than 180 parameters were contained in the subject dataset. Since the object of interest was limited to predicting fuel flow, the vast majority of these parameters were eliminated based on relevance. Following a reasoned elimination of other variables due to multicollinearity, curvilinearity, skewness and other adverse conditions, the remaining variables (i.e., 10) were entered into a standard, non-stepwise regression with fuel flow (ff) as the dependent variable. Since there is fuel flow on two engines on a Boeing 757 aircraft and parameters are recorded for each, two equations were produced; one for engine 1 (ENG1ff) and one for engine 2 (ENG2ff).

Fuel flow was best predicted by calibrated airspeed (CAS), gross weight (GWeight), and engine N2 (ENG*x*n2; i.e., high compressor speed, see Table 1 for a definition of each of the FOQA parameters used in the study). The resulting equations were as follows:

ENG1ff: – 9170.077 + 10.943 CAS + 0.008657 GWeight + 93.701 ENG1n2, with an R^2 (coefficient of determination) of .853

ENG2ff: – 9347.178 + 10.835 CAS + 0.008726 GWeight + 95.616 ENG2n2, with an R^2 of .872

The models formulated were checked for adequacy through the examination of residuals, and testing for a linear fit of the predictors to the dependent variable. Based on an analysis of residuals and tests for linear fit, there did not appear to be any correlation between random errors, the variables appeared to be linearly related, and there appeared to be reasonably consistent variances in the data for both models.

To validate the models, data on 179 additional flights were obtained. These data were fitted using the derived models and the performance of both models suggested that they were likely to be successful as predictors. In fact, the R^2 on engines 1 and 2 with the new data were 86.3% and 87.2%, respectively, which was approximately equivalent to the fit of the original data.

Table 1. Flight Operations Quality Assurance (FOQA) Parameters

FOQA Parameter Name	Definition
Mach	Mach
CAS	Calibrated airspeed
TAT	Total air temperature
ALT	Altitude
GWeight	Gross weight
ENG1epr, ENG2epr	Engine 1 and 2exhaust pressure ratio
ENG1ff, ENG2ff	Engine 1 and 2 fuel flow
ENG1n1, ENG2n1	Engine 1 and 2 low compressor speed
ENG1n2, ENG2n2	Engine 1 and 2 high compressor speed
ENG1egt, ENG2egt	Engine 1 and 2 exhaust gas temperature
AOA	Angle of attack
ATTroll	Angle of bank
ATTpitch	Pitch attitude
SFCstab	Stabilizer position
CTLspdbrk	Speedbrake control position
SFCalrn	Left aileron position
SFCalrnrt	Right aileron position
SFCrudder	Rudder position
SFCelev	Left elevator position
SFCelevrt	Right elevator position
SFCflap	Flap position

METHODOLOGY

In the previous study, much effort was made to transform the data that was problematic or to perform a reasoned elimination of some of the variables. In fact, a nontrivial number of variables had to be eliminated in order to avoid violations of assumptions and, thus, have confidence in the results. Admittedly, this had the effect of reducing the performance of the regression models, but the trade-off between model performance and confidence in the result is a conundrum routinely faced by analysts. By contrast, DM methods are generally robust to non-linear data, complex relationships, and non-normal distributions; thus, no pre-processing or transformations were performed as part of the DM project.

It should be noted that the regression analyses performed in the previous study were ultimately performed using clean data that met all reasonable assumptions for regression studies, and so a high predictive capability of the models was anticipated even though only a small subset of predictors were used. Given these conditions, it was not anticipated that DM methods would perform significantly better than multiple linear regression since the regression models' explained variance was .853 (ENG1ff) and .872 (ENG2ff).

To facilitate the desired comparison, a standard recursive partitioning (i.e., tree) method called Classification and Regression Tree Models (C&RT) was performed due to its popularity and ease of interpretation. The C&RT method builds classification and regression trees for predicting variables. *STATISTICA* contains numerous algorithms for predicting continuous or categorical variables from a set of continuous predictors and/or categorical factor effects. Each child node in the tree diagram represents a bivariate split on one of the predictors. Terminal nodes indicate actual predicted values for sets of cases. The dendrograms created in this process are quite easy to review and interpret to understand the sets of if/then statements created by the model.

This was followed by an Advanced Comprehensive Regression Models (ACRM) project. This model has several pre-arranged nodes for fitting linear, nonlinear, regression-tree, CHAID and Exhaustive CHAID, and different neural network architectures to a continuous dependent variable, and for automatically generating deployment information.

Finally, *STATISTICA*'s Intelligent Problem Solver (IPS) procedure was used. The IPS is a sophisticated tool for the creation and testing of neural networks for data analysis and prediction problems. It designs a number of networks to solve the problem, copies these into the current network set, and then selects those networks into the results dialog, allowing testing to be performed in a variety of ways. These latter two projects are *STATISTICA* methods that allow a comparison of numerous DM algorithms simultaneously on a dataset.

In addition to standard analysis techniques, goodness of fit tests were run to compare the performance of various methods.

RESULTS

Initial exploration

The analyst is familiar with the dataset since it was used in the previous study; however, it was examined again for out of range values, impossible data combinations, and other irregularities. It was determined that the dataset was more than adequate for the present study.

Model building and validation (and deployment)

C&RTs were performed. The C&RT method was run using V-fold cross-validation (a technique where repeated (v) random samples are drawn from the data for the analysis). The variables contained in the tree diagram for the Engine 1 model included CAS, GWeight, ENG1n1, ENG1egt, and ALT. A goodness of fit test performed on this model yields the results as depicted in Table 2.

Table 2. Summary of Goodness of Fit—Engine 1 Fuel Flow

Factor	Predicted
Mean Square Error	13449.18
Mean Absolute Error	89.06
Mean Relative Squared Error	0.00
Mean Relative Absolute Error	0.03
Correlation Coefficient	0.92

The C&RT analysis was also performed on the ENG2ff model. The tree diagram for ENG2ff included CAS, GWeight, ENG2n1, and ENG2n2. A goodness of fit test performed on this model yields the results as depicted in Table 3.

Table 3. Summary of Goodness of Fit—Engine 2 Fuel Flow

Factor	Predicted
Mean Square Error	13674.90
Mean Absolute Error	89.25
Mean Relative Squared Error	0.00
Mean Relative Absolute Error	0.03
Correlation Coefficient	0.91

The next method used was *STATISTICA*'s ACRM project. This model fits several DM methods to a continuous dependent variable, and automatically generates deployment information. Figure 3 depicts the *STATISTICA* workspace as it is configured to run this project.

Figure 2. STATISTICA Workspace for Advanced Comprehensive Regression Model Project

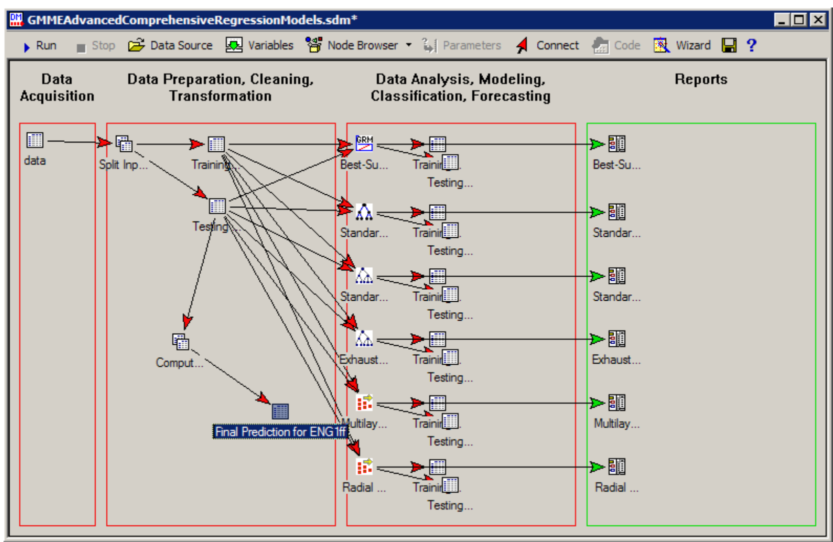


Table 4 contains the summary output from goodness of fit tests on the various methods explored by the ACRM tool on ENG1ff.

Table 4. Summary of Goodness of Fit for Engine 1 Fuel Flow: Advanced Comprehensive Regression Model

Factor	GLM Predicted	Trees Predicted	CHAID Predicted	ECHAID Predicted	MLP Predicted	RBF Predicted
Mean Square Error	670.201	9025.980	56545.54	46538.480	553.511	55059.900
Mean Absolute Error	19.253	71.926	181.990	166.860	17.7905	181.690
Mean Relative Squared Error	0.000	0.001	0.000	0.000	0.000	0.000
Mean Relative Absolute Error	0.006	0.021	0.050	0.050	0.005	0.050
Correlation Coefficient	0.996	0.941	0.530	0.640	0.997	0.550

GLM – Generalized Linear Model
CHAID - Chi-squared Automatic Interaction Detection Model
ECHAID - Exhaustive Chi-square Automatic Interaction Detection Model
MLP - Multilayer Perceptron Model
RBF - Radial Basis Function Model

Both the Generalized Linear Model (GLM) and the Multilayer Perceptron (MLP) had very high correlation coefficients exceeding 0.995 and relatively low error measures. Figure 4 depicts a plot of the predicted variable versus the observed, and Figure 5 depicts a plot of the residuals versus the observed variable for the GLM for ENG1ff.

Figure 3. General Linear Model of Engine 1 Fuel Flow: Predicted versus Observed

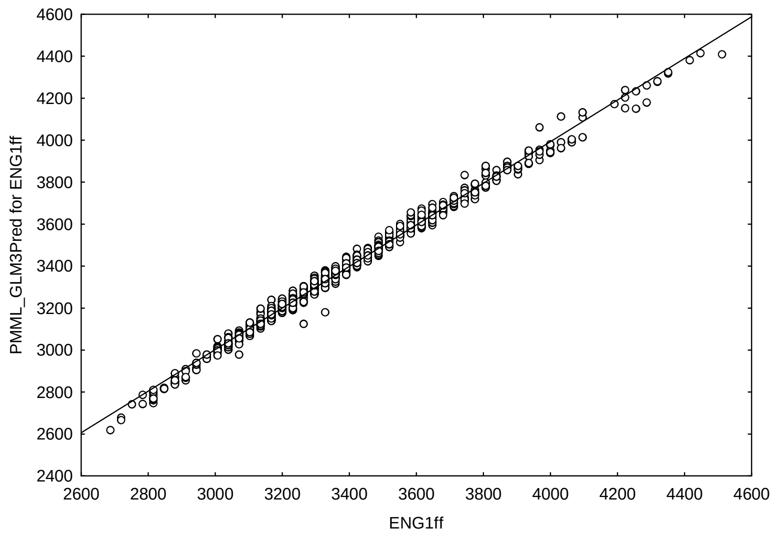


Figure 4. General Linear Model of Engine 1 Fuel Flow: Residuals versus Observed

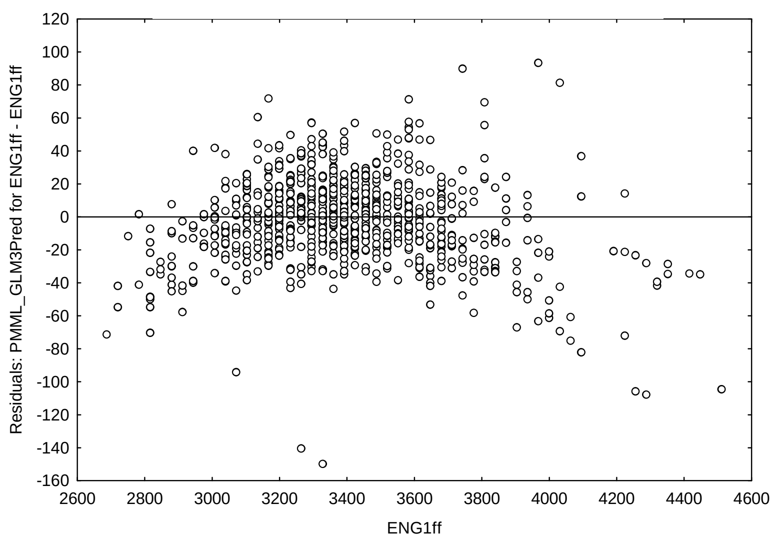


Figure 5. Multilayer Perceptron for Engine 1 Fuel Flow: Predicted versus Observed

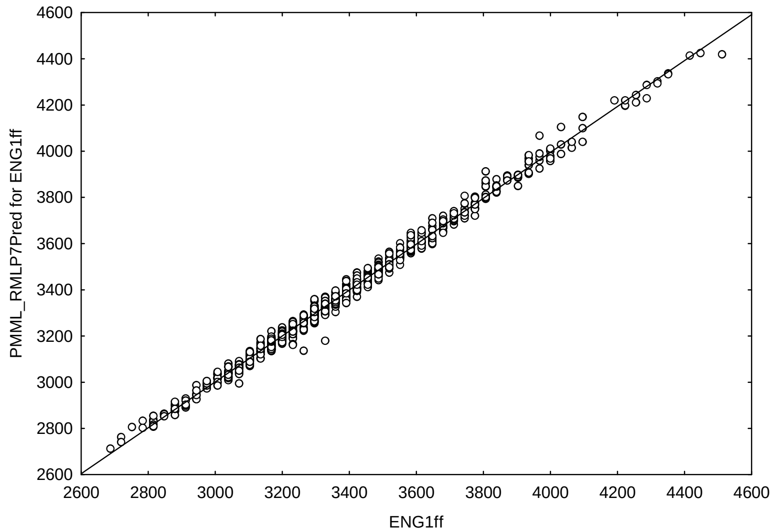


Figure 6. Multilayer Perceptron for Engine 1 Fuel Flow: Residuals versus Observed

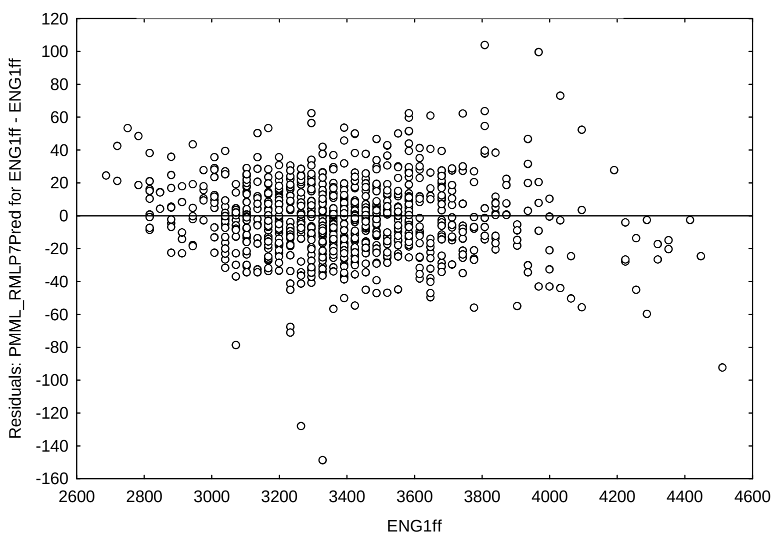


Figure 6 depicts a plot of the predicted variable versus the observed variable, and Figure 7 depicts a plot of the residuals versus the observed for the MLP.

Table 5 contains the summary output from goodness of fit tests on the various methods explored by the ACRM tool on the ENG2ff model. As with the ENG1ff model it can be concluded that the GLM and the MLP models provided the best predictive capability for ENG2ff of the models tested.

Table 5. Summary of Goodness of Fit for Engine 2 Fuel Flow: Advanced Comprehensive Regression Model

Factor	GLM Predicted	Trees Predicted	CHAID Predicted	ECHAID Predicted	MLP Predicted	RBF Predicted
Mean Square Error	633.783	8899.214	42906.560	38836.210	786.319	32815.580
Mean Absolute Error	18.734	68.991	159.980	150.560	19.877	129.160
Mean Relative Squared	0.000	0.001	0.000	0.000	0.000	0.000
Mean Relative Absolute Error	0.006	0.020	0.050	0.040	0.006	0.040
Correlation Coefficient	0.996	0.945	0.690	0.720	0.995	0.770

GLM – Generalized Linear Model
CHAID - Chi-squared Automatic Interaction Detection Model
ECHAID - Exhaustive Chi-square Automatic Interaction Detection Model
MLP - Multilayer Perceptron Model
RBF - Radial Basis Function Model

The final procedure used was STATISTICA's IPS. The IPS creates and tests several neural networks for data analysis and prediction problems. Tables 6 and 7 are summaries of a goodness of fit analyses for the five models retained for ENG1ff and ENG2ff, respectively.

Table 6. Summary of Goodness of Fit for Engine 1 Fuel Flow: Intelligent Problem Solver

Factor	ENG1ff Model 1 GLM	ENG1ff Model 2 MLP	ENG1ff Model 3 MLP	ENG1ff Model 4 RBF	ENG1ff Model 5 RBF
Mean Square Error	683.411	690.712	711.707	6043.053	4424.089
Mean Absolute Error	19.061	19.030	20.130	48.525	50.813
Mean Relative Squared Error	0.000	0.000	0.000	0.000	0.000
Mean Relative Absolute Error	0.006	0.006	0.006	0.014	0.015
Correlation Coefficient	0.996	0.996	0.995	0.961	0.971

ENG1ff – Engine 1 Fuel Flow
GLM – General Linear Model
MLP – Multilayer Perceptron Model
RBF - Radial Basis Function Model

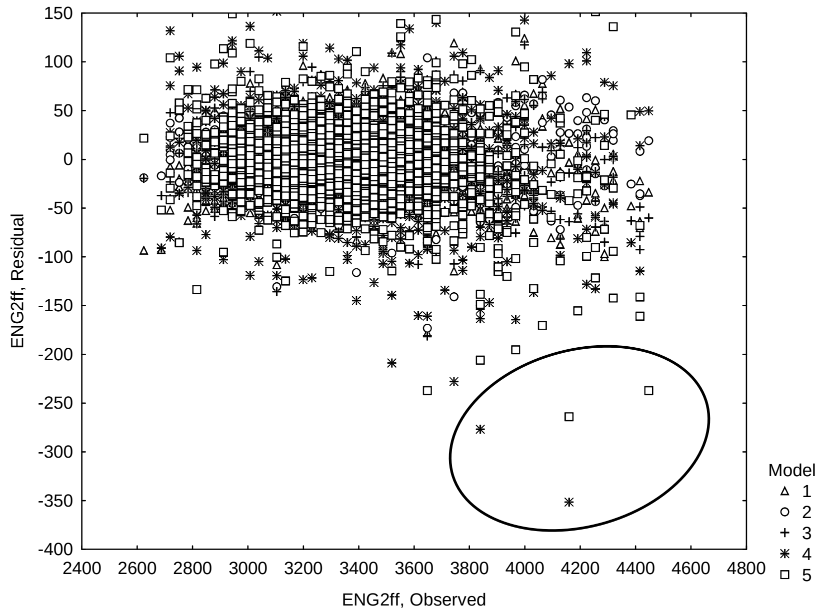
Table 7. Summary of Goodness of Fit for Engine 2 Fuel Flow: Intelligent Problem Solver

Factor	ENG2ff Model 1 Linear	ENG2ff Model 2 MLP	ENG2ff Model 3 MLP	ENG2ff Model 4 RBF	ENG2ff Model 5 RBF
Mean Square Error	736.102	600.180	660.567	1802.759	1706.794
Mean Absolute Error	20.319	18.778	19.273	29.733	28.654
Mean Relative Squared Error	0.000	0.000	0.000	0.000	0.000
Mean Relative Absolute Error	0.006	0.006	0.006	0.009	0.008
Correlation Coefficient	0.995	0.996	0.996	0.988	0.989

ENG2ff – Engine 2 Fuel Flow
GLM – General Linear Model
MLP – Multilayer Perceptron Model
RBF - Radial Basis Function Model

Figure 7 presents a composite graph of all five models evaluated depicting observed versus residuals for the ENG2ff model. This graph shows a fairly tight pattern of observations with only few possible outliers, which are mostly found in Models 4 and 5 - the two Radial Basis Function (RBF) models.

Figure 7. Composite Graph of all Five Models Evaluated Depicting Observed versus Residuals for the Engine 2 Fuel Flow Model



DISCUSSION

An earlier study was performed using multiple regression methods to predict fuel consumption on an air carrier's Boeing 757 fleet of aircraft. It was determined that some of the data generated by the FOQA system violated assumptions of regression methods, and attempts to transform the data were minimally successful. To ensure a high level of confidence in the results, those data were removed from further consideration. The remaining data produced models with excellent predictive capability. Specifically, the ENG1ff and ENG2ff models had correlation coefficients of .853 and .872 respectively, and tested on new data at approximately these values.

The goal of the present study was to evaluate various DM techniques on the same dataset used in the previous study. A recursive partitioning method, C&RT, and *STATISTICA*'s ACRM and IPS algorithms were deployed on the data. Since DM methods are generally robust to data condition problems, no additional analysis was performed on the data.

The recursive partitioning method, C&RT, produced excellent results, that is, correlation coefficients of .92 and .91. Further, the dendrograms produced by the C&RT are easy to interpret (these graphics are difficult to extract from the software in a readable format and, thus, are not included in this manuscript). For example, it can easily be determined that the first node generated in the ENG2ff dendrogram is based on variable CAS, the bivariate nodes from CAS are GWeight and CAS, the nodes from GWeight are ENG2n1 and GWeight, and so on. This information enables the analyst to better understand the classifications being determined by the algorithm.

The ACRMs also produced excellent results on the data. The correlation coefficients reported by each of the models were very high. The GLM reported correlation coefficients of .996 for both ENG1ff and ENG2ff, and the MLP reported correlation coefficients of .997 and .995 for ENG1ff and ENG2ff, respectively. These values significantly exceed those obtained by standard multiple regression methods. The error values for the GLM and the MLP models were also low relative to the other models examined.

The IPS model produced five models with no correlation coefficients less than .961. As with the ACRM results, the GLM and MLP models were the best performers, with all correlation coefficients exceeding .995.

CONCLUSIONS AND NEXT STEPS

The purpose of the study was to compare DM methods against standard multiple regression methods using FOQA data on a fuel consumption study. The study examined several DM methods, and several performed very well in predicting fuel consumption. In general, CR&T, GLM, MLP, and RBF methods performed much better than standard multiple regression methods in predicting the dependent variable. As with other neural networks, interpretation of results is more difficult than with traditional statistical tools, and would require knowledge of the underlying theory.

It was determined that DM holds great potential for exploring large datasets, such as are generated in a FOQA program, and learning more from the data than can be accomplished using standard statistical tools alone. Further, this project suggests that DM techniques might be utilized effectively on air carrier-generated datasets to improve operational efficiency and safety.

The broader goal of this work is the creation of a practical tool that can be used by airlines to quickly identify aircraft with outlier fuel burns. This is not a trivial problem. While aircraft manufacturers provide detailed performance information and airlines routinely compute fuel consumption statistics for their fleets, the factors that contribute to any one flight's fuel consumption are quite variable. Differences between flights in load, cruise altitude, temperature and chosen cruise airspeed cause noticeable changes in fuel flow, making the identification of anomalous rates of fuel consumption difficult.

The accuracy with which the GLM and MLP neural network models predict fuel flow give encouragement that these models, coupled with other statistical tools such as process control charts, will enable the analyst to sensitively detect adverse trends, caused perhaps by out of trim conditions, improper loading, or engine foreign object damage. Testing whether such a fuel consumption anomaly detector can be constructed is the next project in this research effort.

REFERENCES

BaFail, O. (2004). Applying data mining techniques to forecast number of airline passengers in Saudi Arabia (domestic and international travels). *Journal of Air Transportation 9*(1), 100-116.

Global Aviation Information Network. (2004). *Application of Smiths Aerospace data mining algorithms to British Airways 777 and 747 FDM data: A technology demonstration in partnership with the Federal Aviation Administration and the Global Aviation Information Network.* Retrieved July 1, 2006, from http://www.gainweb.org/

Global Aviation Information Network. (2005). *Application of Insightful Corporation's data mining algorithms to FOQA data at JetBlue Airways: A technology demonstration in partnership with the Federal Aviation Administration and the Global Aviation Information Network.* Retrieved July 1, 2006, from http://www.gainweb.org/

Knight, P., Cook, J., & Azzam, H. (2005). Intelligent management of helicopter health and usage management systems data. *Proceedings of the Institution of Mechanical Engineers 219* (Part G: Journal of Aerospace Engineering), 507-524.

Regalado, A. (1999). Mining the genome. *Technology Review 102*(5), 56-62.

Solomon, S., Nguyen, H., Liebowitz, J., & Agresti, W. (2006). Using data mining to improve traffic safety programs. *Industrial Management & Data Systems 106*(5), 621-643.

StatSoft. (2003). *STATISTICA data miner (version 7.1)* [computer software]. Tulsa, OK.

Stolzer, A. (2003). Fuel consumption modeling of a transport category aircraft: A flight operations quality assurance (FOQA) analysis. *Journal of Air Transportation 8*(2), 3-18.

Strickland, L. S., & Willard, J. (2002). Re-engineering the immigration system: A case for data mining and information assurance to enhance homeland security—part I: Identifying the current problems. *Bulletin of the American Society for Information Science and Technology 29*(1), 16-22.

Wang, J. (2003). *Data mining: Challenges and opportunities*. Hershey, PA: Idea Group Publishing.