Predicting the Operational Acceptance of Route Advisories

Antony D. Evans*
Crown Consulting Inc., NASA Ames Research Center, Moffett Field, CA, 94035

Paul Lee†
NASA Ames Research Center, Moffett Field, CA, 94035

NASA envisions a future Air Traffic Management system that allows safe, efficient growth in global operations, enabled by increasing levels of automation and autonomy. In a safety-critical system, the introduction of increasing automation and autonomy has to be done in stages, making human-system integrated concepts critical in the foreseeable future. One example where this is relevant is for tools that generate more efficient flight routings or reroute advisories. If these routes are not operationally acceptable, they will be rejected by human operators, and the associated benefits will not be realized. Operational acceptance is therefore required to enable the increased efficiency and reduced workload benefits associated with these tools. In this paper, the authors develop a predictor of operational acceptability for reroute advisories. Such a capability has applications in tools that identify more efficient routings around weather and congestion and that better meet airline preferences. The capability is based on applying data mining techniques to flight plan amendment data reported by the Federal Aviation Administration and data on requested reroutes collected from a field trial of the NASA developed Dynamic Weather Routes tool, which advised efficient route changes to American Airlines dispatchers in 2014. 10-Fold cross validation was used for feature, model and parameter selection, while nested cross validation was used to validate the model. The model performed well in predicting controller acceptance or rejection of a route change as indicated by chosen performance metrics. Features identified as relevant to controller acceptance included the historical usage of the advised route, the location of the maneuver start point relative to the boundaries of the airspace sector containing the maneuver start (the maneuver start sector), the reroute deviation from the original flight plan, and the demand level in the maneuver start sector. A random forest with forty trees was the best performing of the five models evaluated in this paper.

I. Introduction

THREE of six strategic thrusts outlined by the National Aeronautics and Space Administration (NASA) Aeronautics Research Mission Directorate (ARMD) in the NASA Aeronautics Strategic Implementation Plan[1] are safe, efficient growth in global operations (Thrust 1); real-time system-wide safety assurance (Thrust 5); and assured autonomy for aviation transformation (Thrust 6). In Air Traffic Management (ATM) particularly, as technologies move to more integrated systems, the three strategic thrusts described above may converge, *enabling autonomic system concepts that increase capacity and efficiency while maintaining or enhancing system safety*.[1]

In a safety-critical system such as ATM, system evolution and the introduction of increasing autonomy has to be done in stages. Technology is not yet at a point where a fully autonomous ATM system can be deployed. Hence human-system integrated concepts, with progressively increasing autonomy as the available technology develops and is introduced, will be critical in the foreseeable future. It is therefore essential for the success of future automation that automated support systems work effectively in the existing ATM system, and in particular, with the humans that operate in it. This requires that automation be designed to be understandable and acceptable to airline and air traffic personnel, including dispatchers, air traffic controllers and traffic managers. In the near term,

* Senior Scientist I, Aviation Systems Division, M/S 210-6, Senior Member.
† Senior Research Associate, Human-Systems Integration Division, M/S 262-4.

automation will primarily provide decision support to these human users rather than act autonomously. Lessons learned from the harmonious introduction of automation can also inform the design of future autonomous systems.

For automation tools to be successfully implemented in a mixed human-automation environment, there is a need to ensure that the outputs of the tools, including decision support tools, have high operational acceptance, i.e., have a high probability of being implemented by controllers. One example is for tools that generate alternative flight routings or reroute advisories. If the reroute advisories generated by decision support tools were more often operationally acceptable, a higher number of them would be accepted and implemented by controllers, increasing efficiency. Increased operational acceptance would also reduce the workload burden of the tool on the controller per requested route, because fewer route modifications would have to be made before a Center route amendment (a flight plan amendment entered into the host computer by an Air Route Traffic Control Center controller) could be implemented. Ultimately it could also free up controller time to work on other problems, with further benefits to the system.

One application for which the generation of operationally acceptable routings is particularly important is Trajectory Based Operations (TBO). TBO is an air traffic management paradigm that will enable all vehicle classes, including manned and unmanned systems, with energy-efficient, user-negotiated routing by sharing 4-dimensional trajectory information between decision-makers and their supporting automation so that aircraft and National Airspace System (NAS) resources are flexibly managed based on reliable predictions of future states. While some of this coordination will be done at the level of traffic flows, flights must ultimately be allocated specific 4-dimensional trajectories, which will be updated dynamically as conditions change. For success of the concept in a mixed human-automation environment, these routings should be generated in such a way as to maximize operational acceptance.

A number of tools and concepts are under development by NASA and the Federal Aviation Administration (FAA) to enable TBO by increasing the capacity, efficiency and accessibility of the NAS. This can be accomplished through changes in flight routing that avoid adverse weather and congestion, or by allowing increased opportunities for airlines and operators of future aviation concepts, such as unmanned aerial systems and on-demand mobility, to request routes that meet their preferences. Tools under development include tactical tools such as the Dynamic Weather Routes (DWR) tool[2,3], Multi-Flight Common Routes (MFCR)[4], Dynamic Routes for Arrivals in Weather (DRAW)[5], the Optimized Route Capability (ORC)[6], and the Traffic Aware Strategic Aircrew Requests (TASAR) tool[7,8]; and strategic tools such as the Collaborative Trajectory Options Program (CTOP)[9]. The effectiveness of each of these tools could be improved by generating routing advisories that have increased operational acceptance. These tools would therefore benefit from a functionality that quantifies route acceptability to Air Traffic Control (ATC). The contribution of this paper is to develop a predictor of a proposed route's operational acceptability. This allows identification of the most important features for route operational acceptance, and could be used to inform route selection by decision support tools, ultimately contributing to the development of increasingly autonomous systems that are capable of routing aircraft with less human input than is currently the case.

Background literature is presented in Section II. The approach for predicting the operational acceptability of route advisories is described in Section III, followed by results from training the algorithm in Section IV. Conclusions and recommendations for future work are presented in Section V.

II. Background

A number of studies have examined elements of route acceptability to ATC. Ref. 10 analyzed the historical usage of different flight routings in order to improve route acceptance for the NASA developed DWR tool[2], which has been used to advise American Airlines dispatchers of en-route weather avoidance reroute opportunities in an operational trial.[3] The results suggest that historical usage (i.e., whether a route has been used before) is a key requirement for a route's acceptance by ATC – 96% of Center route amendments made in response to dispatcher-accepted DWR reroute advisories from 29 days in the summer of 2014 also appeared in data from April 2015. However, the opposite is not true. Requesting a reroute that is observed in historical data does not guarantee ATC acceptance. In fact, 66% of dispatcher-accepted DWR reroute advisories that were rejected by ATC[‡] in the 2014 DWR data also appeared in the data from April 2015.[10] It is clear that, while historical usage is generally indicative of route acceptance, there are also other reasons for a controller's acceptance or rejection of a reroute request.

Ref. 11 defines an optimization approach that generates operationally acceptable reroutes for flights predicted to request deviations from their current routes for weather. The method considers many factors including route

[‡] A reroute was considered to be rejected by ATC when no Center route amendment was made in response to the dispatcher-accepted DWR route advisory, or a Center route amendment was made, but it was different to the DWR route advisory accepted by the dispatcher.

deviation distance, conformance of the reroute to historically flown routes, weather impact on the current route, sector congestion, and factors including required point-outs[§] and inter-facility coordination. The routing network used for the optimization was generated by segmenting historically flown routes into fix-pair segments. Thus, all arcs in the modeled network consist of previously-flown connections between fixes, so each individual arc in the network has some level, depending on usage, of operational acceptability. Reroutes are constructed from these arcs using an optimization algorithm, and the reroutes that best meet a set of metrics of operational acceptability are presented as potential alternatives to users.[12]

Other related research includes Ref. 13, which describes a concept for tactical reroutes around convective weather that leverages new technologies to automate the necessary coordination between traffic managers and controllers. The concept assumes the use of the Corridor Integrated Weather System (CIWS) and Convective Weather Avoidance Model (CWAM) for weather detection and in reroute generation to avoid the weather, and incorporates route acceptability factors described in Ref. 11. Ref. 7 presents the TASAR concept, which combines Automatic Dependent Surveillance-Broadcast (ADS-B) and airborne automation to enable user-optimal in-flight trajectory re-planning and to increase the likelihood of ATC approval for the resulting trajectory change request, incorporating traffic, weather, and airspace information in the optimization process.

The approach described in the current paper differs from that used in Ref. 7, 11, 12, and 13 in that it uses historical data – particularly issued Center route amendments – to explicitly extract the drivers of operational acceptance, and builds a predictor that can be used to evaluate the likely operational acceptability of any generated route advisories. While it is very important to elicit subject matter expert (SME) feedback on what features are likely to drive ATC acceptance of route requests, because of the nature of how controllers make decisions about route requests, relying on SME input alone may have limitations. This is because controller decision-making is highly complex, based on many different factors and how these different factors impact each other, and it is difficult, even for SMEs, to explain all the nuances of their decisions to researchers. Operational data, however, if mined effectively, has the potential to yield many of these complexities and nuances, and it is these relationships that the current paper aims to extract. SME feedback can, however, be very useful in feature identification, and is also essential to better understand how the features that are identified as being most important drive operational acceptability.[8]

III. Approach

Data mining and machine learning techniques have been applied to learn trends and correlations from historical data in many fields, and ATM is no exception.[14,15,16,17,18,19,20] Given suitable historical data, data mining and machine learning techniques provide an effective tool to identify the degree to which the operational acceptance of reroute advisories is predictable, and to build a predictor that would estimate the probability that a reroute advisory would be rejected by ATC or accepted and implemented in the form of a Center route amendment. They also provide the opportunity to identify the dominant drivers of ATC acceptability from the predictor feature set.

A. Data Availability

Data availability is a key constraint, and one of the biggest challenges when applying data mining and machine learning techniques to problems in ATM, because of the broad range of information used by controllers and traffic managers to make decisions. Ideally two classes of observations are desired. In the context of predicting the operational acceptability of reroute advisories, this means identifying which reroute advisories were rejected by ATC, with associated features, and which were accepted. Accepted reroute advisories are easily identified in the form of recorded Center route amendments, but rejected reroute advisories are much more difficult to identify, as they are not typically recorded. Single class classifiers, which apply techniques in anomaly detection, may be used to build predictors based on a single class of observations only (e.g., accepted advisories), but their accuracy is generally poorer than two-class classifiers, so the latter is preferable.

During the DWR trials at American Airlines[3], all proposed- and dispatcher-accepted reroute advisories generated by DWR were recorded. The ATC response to these dispatcher-accepted reroute advisories was then surmised based on whether or not the flight had a Center route amendment implemented within 30 minutes of its DWR reroute advisory being accepted by the dispatcher. This means that dispatcher-accepted DWR reroute advisories that resulted in no Center route amendment, or resulted in Center route amendments that differed from the dispatcher-accepted reroute requested, were also recorded. These can be considered to represent ATC rejected reroute

[§] A point-out refers to the need for one controller to request that the controller of an adjacent sector also monitors a flight that is close to the sector boundary.

advisories. This data therefore provides a useful dataset for which two-class classifiers can be developed to predict ATC route acceptability.

The approach used in this paper is therefore to use two-class classification to train and test data-mining algorithms on the ATC rejected and accepted reroute advisories recorded by DWR over 5 months from May 9 through September 30, 2014. The features used are described in the following section.

B. Feature Identification

A number of different features have been identified in the literature as being important to controller acceptance or rejection of reroute requests. In Ref. 11, 12 and 13 features considered in designing reroutes included route deviation distance, conformance of the reroute to historically flown routes, weather impact on the current route, sector congestion, and factors including required point-outs and inter-facility coordination. Ref. 8 identified similar features impacting controller acceptability of TASAR trajectory change requests based on observations at Atlanta Air Route Traffic Control Center (ZTL) and Jacksonville Air Route Traffic Control Center (ZJX), during TASAR trials from June 8 to June 20, 2015. The most common reasons for rejecting requests during the trials were conflicts with other traffic, violating letters of agreement (LOAs) and negatively impacting neighboring sector workload, major arrival and departure flows and flow restrictions. Requests were also rejected due to the aircraft already being handed off to another sector, intruding into an active special use airspace (SUA), intruding into another center, weather, and unfamiliarity with the requested waypoint.

In this paper, four sets of features impacting controller acceptance are included:

1. Features describing historical route usage;
2. Features describing demand;
3. Features describing reroute deviation; and
4. Features describing the maneuver starting point.

These features are described in detail in the subsections below.

Features Describing Historical Route Usage

Features describing historical usage capture elements of many of the features listed in the literature, including conformance of reroutes to historically flown routes, the requirement for point-outs and inter-facility coordination, LOAs, major arrival and departure flows, SUA, Center boundary crossings, and familiarity with the requested waypoints. In this paper, historically observed routings were extracted from flight plans and Center route amendments in historical Aircraft Situation Display to Industry (ASDI) data from June to August 2015 for all sectors in Fort Worth Center (ZFW) and its adjacent Centers, generating tables of historically observed routings, with counts of historical usage, as described in detail in Ref. 10. These are referred to as common routing tables. June to August 2015 was chosen as the training set because it captures periods of high traffic and convective weather activity in the region studied, both of which can lead to the rerouting of flights. There is also no overlap with the period in 2014 for which DWR reroute advisories and Center route amendments were collected (May to September 2014), which are the routes for which historical usage is to be quantified. Routings were only extracted for local times between 05h00 and 00h00 (midnight) to ensure that the routings identified did not represent nighttime operations when unusual routings may be widely allowed because of very low traffic.

Two approaches, illustrated in Figure 1, were used to identify how often the observed DWR reroute advisories and Center route amendments from 2014 appeared in the historical data from 2015. Full routing historical counts were identified by directly comparing the full reroute, from the maneuver start point to the return point on the original flight plan (the return capture fix), to the historically observed routes listed in the common routing tables, and, when identified, recording the historical count for that routing from June to August 2015. However, because full routings often appear relatively infrequently in the data, the DWR reroute advisories and Center route amendments from 2014 were also broken up into the component segments linking all the waypoints in the routing. These were each compared to the common routing tables. For each given DWR reroute advisory or Center route amendment, a new historical routing count by segment was then generated, based on the counts of the route's component segments from the common routing tables. Specifically, the routing count by segment for a reroute is taken to be the minimum of the counts of all the component segments of that reroute. Routing counts by segment are generally higher than full routing counts. Further details of how the full and segment routing counts were generated can be found in Ref. 10.

Two features (with abbreviations underlined) are therefore included in the feature set that describe historical usage on the proposed reroute:

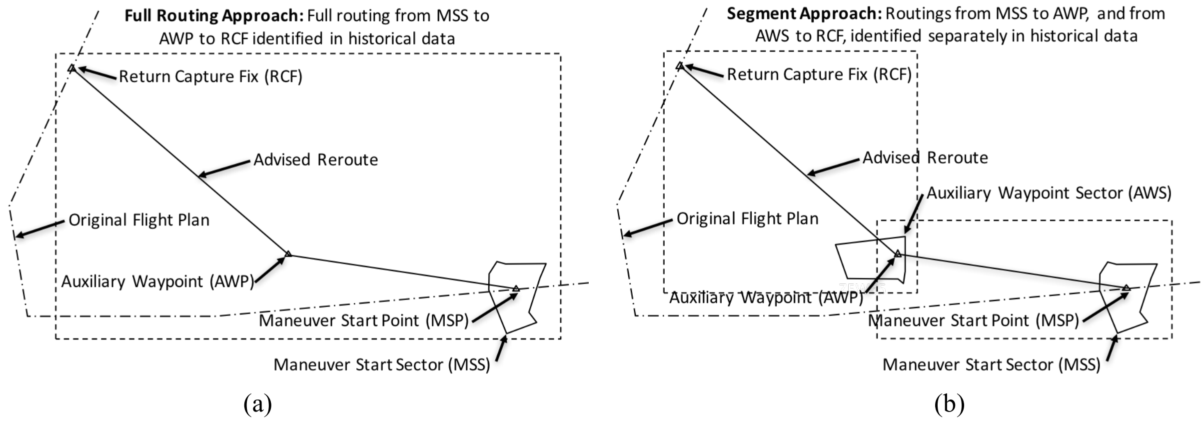

Figure 1. Comparison of (a) full routing approach and (b) segment approach to identifying routings in historical data.[10]

1. Hist. Count (Full Routing): The full routing count from historical data (June to August 2015);
2. Hist. Routing Count by Segment: The routing count by segment from historical data (June to August 2015).

Features Describing Demand

Features describing demand capture elements of some of the other features described in the literature, including sector congestion, conflicts with other traffic, sector workload, and major arrival and departure flows. Some of these features may be explicitly included in future work. Particularly, one feature that may be of importance is whether or not the requested route would require close monitoring to avoid a potential conflict. This would be the case, for example, if the route took the flight towards another aircraft downstream. Acceptance of such a route would require monitoring by the controller to ensure that no potential conflict with the other aircraft arises. In this case, a more likely response would be a modification to the requested reroute to route the aircraft away from the other aircraft, which would require less monitoring. Because of limitations in the data available, a feature explicitly capturing this effect is not included in this paper. Instead four features are included that describe demand levels on the proposed reroute:

3. MSS D/C Ratio: The ratio of demand to capacity (as defined by the Monitor Alert Parameter) (D/C) for the sector in which the reroute would start (the maneuver start sector, MSS), were it accepted by ATC (not all reroutes in the data ultimately resulted in Center route amendments). The maneuver start sector demand that was projected to be in the maneuver start sector when the flight in question started the maneuver was modeled explicitly. This demand was projected with data from the time when the controller made the decision to accept or reject the reroute, instead of using actual observed traffic counts from when the maneuver was to start. This provides a more realistic representation of the information available to the controller when making his/her decision to accept or reject the route request;
4. MSS Over Capacity: Explicitly whether the maneuver start sector would be over capacity or not when the maneuver started;
5. No. Sectors Over: The number of sectors between the maneuver start point and return capture fix in which demand would exceed capacity were the reroute to be accepted;
6. Max D/C Ratio: The maximum ratio of demand to capacity across all sectors between the maneuver start point and return capture fix, were the reroute to be accepted.

The ratio of demand to capacity for the maneuver start sector (MSS D/C Ratio), and whether or not it exceeds capacity (MSS Over Capacity), give an indication of the workload level for the controller that has to make the decision concerning whether to accept or reject the requested reroute. If the ratio of demand to capacity is high, there is also more likely to be a potential conflict that the controller must resolve if he or she were to accept the reroute. These features therefore represent proxies for the feature described above identifying whether or not the requested route would require close monitoring to avoid a potential conflict.

The other two features described above (No. Sectors Over and Max D/C Ratio) account for the congestion and workload levels for downstream sectors, including neighboring sector workload, identified as a feature in Ref. 8. The controller taking the request does not have access to explicit demand and capacity information for the downstream sectors (such as through the flight schedule monitor, which is only available to the traffic manager). However, the controller is likely able to observe and project how busy other controllers are in their specific area, and hence these features are included. If it was a traffic manager that was making the decision, it is expected that these features would be more important. Again, if the ratio of demand to capacity is high downstream, there is also more likely to be a potential conflict that a downstream controller must resolve if the reroute is accepted.

All demand related features were generated for each dispatcher-accepted DWR route advisory and flight plan amendment in the DWR trial data from 2015 using the Future ATM Concepts Evaluation Tool (FACET), which computes sector counts. Sector counts were projected based on active flight plans at the time when the DWR reroute advisory was accepted by the dispatcher, or when the flight plan amendment was made, and the resulting traffic in each sector at the times when the flight under question was projected to be in that sector. Projected sector counts were, however, based on actual departure times, which were extracted from the ASDI data, and not scheduled departure times, as would be the case in the flight schedule monitor. This is because scheduled departure times were not available in the data used. This will be adjusted in future work.

Note that one key feature described in Ref. 11 that is relevant to the operational acceptability of reroute advisories is weather impact. This is not included in the feature set in this paper because all reroute advisories generated by DWR already avoid the forecast weather. However, it is important to recognize that this is also a critical feature. In future work, the minimum distance of the reroute from forecast convective weather may also be included as a feature. This data was not available for this paper.

Features Describing the Reroute Deviation

Reroute deviation distance was identified as important in designing reroutes in Ref. 11 and 12. While lateral deviation from the original route is used in those studies, in this paper we quantify the route deviation simply by the number of sectors traversed by the reroute, prior to returning to the original flight plan at the return capture fix. This represents how many sectors are likely to be impacted by the reroute were it to be accepted, and makes no reference to the number of sectors impacted by the original flight plan. The nature of the route deviation is also quantified by including a feature indicating whether or not the reroute requested is a direct routing to the return capture fix, or includes one or more auxiliary waypoints. This feature is included because it was observed in Ref. 10 that direct routings saw higher ATC acceptance than routings via auxiliary waypoints.

The features included in the feature set describing reroute deviation are therefore as follows:

7. <u>No. Downstream Sectors</u>: The number of downstream sectors between the maneuver start point and return capture fix;
8. <u>Direct Routing</u>: Whether the routing is direct to the return capture fix, or via one or more auxiliary waypoints.

Features Describing the Maneuver Start Point

As described above, one cause for the rejection of reroute requests was that aircraft were close to being handed off to another sector.[8] This suggests that the location of an aircraft in the maneuver start sector when the reroute is requested may be relevant to its acceptance or rejection. Two features are therefore included describing the aircraft location in the maneuver start sector:

9. <u>Time to Exit MSS</u>: The time between when the reroute advisory is accepted by the airline dispatcher, and when the flight exits the maneuver start sector;
10. <u>Dist. to Exit MSS</u>: The distance between the maneuver start point and the maneuver start sector boundary, along the requested reroute.

A schematic of reroute events is shown in Figure 2. Two features are included to describe the aircraft location in the maneuver start sector because of limitations in the data used. The feature of most interest is the time between when the pilot requested the reroute from the controller (event 4 in Figure 2), and when the flight exited the maneuver start sector (event 7 in Figure 2), because this would give an indication of how long the controller would have to handle the request. However, no data is available on when the pilot requested the reroute from the controller. The only times that are available are the time when the airline dispatcher-accepted the reroute request (event 2

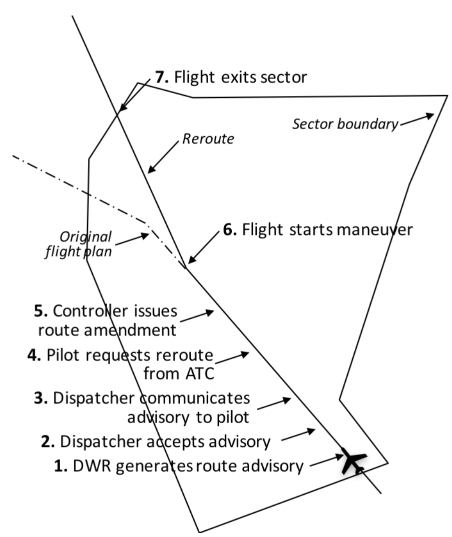

Figure 2. Reroute Event Schematic

in Figure 2), before communicating it to the pilot, and when the aircraft exited the maneuver start sector. The difference between these times is calculated, and included as a feature (Time to Exit MSS). However, the location of the maneuver start point is set by DWR so as to be far enough downstream to allow for communication between the dispatcher and pilot, communication between the pilot and controller, and controller decision making. The location of the maneuver start point can therefore be considered a limit on when the controller is expected to be ready to implement a Center route amendment. The distance between the maneuver start point and sector boundary, along the requested reroute, is therefore also calculated, and included as a feature (Dist. to Exit MSS). In the case of observed Center route amendments, the maneuver start point is classed as the location at which the Center route amendment departs from the original flight plan. Note that the dispatcher can accept a reroute advisory when the maneuver start point is in a future sector. In this case, feature 9 is still calculated relative to the time to exit the maneuver start sector, and not the current sector. This is to account for the fact that, if the controller of the current sector were to reject the request because the aircraft is too close to hand-off, the pilot still has the opportunity to make the request of the controller in the following sector, for which this is less likely to be a problem. With more complete data, these features will be adjusted to more accurately capture the desired effects.

The 10 features listed above do not capture all components of operational acceptability for requested reroutes, but cover the majority of those identified in the literature. Other features will be added in future work as they are identified.

C. Data Mining

Identification of the best model for predicting the operational acceptability of reroute advisories requires a number of development steps. These include (1) feature selection, (2) model selection and (3) parameter selection, in which the subset of features, the data mining algorithm, and parameters required by the data mining algorithm, respectively, are identified to maximize the chosen performance metric.

1. Feature selection involves identifying which combination of features provide best model performance. This is done using *k*-fold cross-validation[**] to estimate the model performance with different combinations of features. For the results presented in this paper, a greedy forward search[21] is used for feature selection, reducing the number of feature combinations for which the model performance is calculated. The greedy forward search procedure begins with an empty feature set and, in each step, adds the feature that results in the largest improvement in model performance to the current feature set, building up the feature set one feature at a time, until all features have been added. The feature set with highest model performance of all those tested is then adopted. In this paper, 10-fold cross validation is used for feature set evaluation, applying a random forest. A random forest[22] is an ensemble learning method for classification that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes of the individual trees.[23]
2. Model selection involves identifying which data mining algorithm provides the best model performance. The model is a two-class classifier, with ATC rejection of a reroute defined as the 'positive' class, and ATC acceptance of a reroute defined as the 'negative' class. The positive class is typically labeled as the class for which it is most important to correctly predict results. This is dependent on how the predictor is to be used. Misclassification of a reroute advisory rejection could lead to an increase in controller workload, because the predictor would incorrectly predict that the advisory would be accepted by ATC, prompting the airline to request the route. ATC would then be required to reject or modify the proposed route. Misclassification of a reroute advisory acceptance could lead to no reroute being requested, or a less efficient reroute being requested, because the predictor would incorrectly predict that the advisory would be rejected by ATC. This could lead to a loss in efficiency benefits. In this paper higher importance is given to reducing controller workload, and hence ATC rejection is classed as the 'positive' outcome.

   A number of algorithms are applied to train the model, using the *R* statistical computing environment. These include (1) logistic regression; (2) a support vector machine using a sigmoid kernel; (3) a decision tree; and two ensemble methods – (4) a random forest; and (5) adaptive boosting (AdaBoost)[24]. Again, *k*-fold cross validation, with k=10, is used to calculate model performance using each algorithm.

   Imbalances in the number of observations in each class were also considered. For the full dataset 40% of routes are classed as rejected by ATC, while 60% of routes are classed as accepted. This does not represent a large data imbalance. The impact of class balancing on model performance (in terms of F-score, described

[**] In *k*-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data.

below) was still explored, using a variety of over- and under-sampling approaches[25,26,27,28], but none were found to improve performance, so are not presented in this paper.

3. Finally, parameter selection involves identifying key parameters in the chosen data mining algorithm that provide the best model performance, e.g., the number of trees or weak leaners in a random forest or AdaBoost. In this paper, k-fold cross validation, with k=10, is used to calculate model performance across a range of each of the key parameters of the chosen algorithm.

The choice of an appropriate metric for the model development process described above is critical. A number of metrics exist. *Accuracy*, which measures the fraction of correct predictions from all predictions made, is the most intuitive, but can be misleading when datasets are imbalanced. An alternative metric is *F-Score* (also called F_1 *Score* or *F-measure*), which is the harmonic mean of precision and recall, calculated as in Equation (1). Here *precision* refers to the number of elements correctly labeled by the model as belonging to the positive class divided by the total number of elements labeled by the model as belonging to the positive class (i.e., the fraction of retrieved instances that are relevant). *Recall* refers to the number of elements correctly labeled by the model as belonging to the positive class divided by the total number of elements that actually belong to the positive class in the data (i.e., the fraction of relevant instances that are retrieved by the model) (also called *Sensitivity* or *True Positive Rate*). F-Score varies from 0 to 1, with 1 being best.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1}$$

One issue with both accuracy and F-Score as metrics for model performance is that they require setting of a discrimination threshold. Most binary classifiers output the probability of an observation falling within one of two classes. In order to classify the observation, this probability must be compared to a discrimination threshold. If the output probability is greater than the threshold, the observation is classed as positive. Otherwise, it is classed as negative. The discrimination threshold can be set to 0.5, giving equal importance to both classes. However, if one class is considered less desirable than the other, the threshold can be increased or decreased so as to decrease the likelihood of a false prediction in that class. If costs are known for misclassification in each class, the optimal threshold can be set accordingly. However, in many cases such costs are not known, making the setting of the threshold difficult. In the prediction of route operational acceptability, misclassification of route rejection could be considered costlier than misclassification of route acceptance, as discussed above, but specific costs have not been identified. In this paper, a discrimination threshold of 0.5 is therefore applied, for simplicity.

Additional performance metrics can be calculated by varying the discrimination threshold and plotting the resulting performance metrics. The Receiver Operating Characteristic (ROC) curve plots true positive rate against false positive rate, and the area under it provides a metric of model performance ranging from 0 to 1 (1 being best) that is not a function of the chosen discrimination threshold. When using normalized units, this area (typically referred to as Area Under the Curve, *AUC*) indicates the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming positive ranks higher than negative).[29] AUC varies from 0 to 1, with 1 being best.

For the purposes of this paper, model development is based on maximizing F-Score, with the discrimination threshold set to 0.5. Accuracy and AUC are, however, also quoted. ROC curves are also plotted.

D. Validation

The purpose of validation is to demonstrate that the approach described above, including the processes of feature, model and parameter selection, produces a predictor that can be effectively used to predict whether a new reroute advisory would be operationally acceptable to ATC or not. This is typically done by dividing the original dataset into two sets: one 'development' set for developing the model, and one 'evaluation' or 'test' set, for testing the developed model on as-yet-unused data. This ensures that the data used to train the model and compare model parameters is not used to evaluate the model performance. Once the approach has been validated, a final model can be trained on all the available data.

If the available dataset is small, it is desirable to calculate performance metrics based on all the data, which can be done using k-fold cross validation. Because k-fold cross validation is already used in the selection of the features, model and parameters, 'nested' k-fold cross validation is needed for validation. This means that the data is split into k_1 folds, which are used, in different combinations, to compile k_1 different evaluation sets and k_1 different development sets. Each development set is then split into k_2 folds (k_1 may differ from k_2), which are each used for feature, model and parameter selection using k_2-cross validation. With the feature set, model and parameters selected, the resulting models developed using each of the k_1 development sets are then re-trained on their full development set, before being tested on their respective k_1 evaluation sets. The model performance across all k_1 evaluation sets, which represents the whole dataset, can then be calculated. Note, however, that the feature, model and parameter selection across all k_1 models may not be consistent. Results applying this approach are presented in Section IV-D for the prediction of route operational acceptability to ATC.

Once suitable validation results have been generated, the model development can be repeated on the full dataset, using k-fold cross validation across the full dataset for feature, model and parameter selection. The features, model and parameters selected for the prediction of route operational acceptability to ATC are presented in Section IV-A, B, and C respectively. Once features, model and parameters are selected, the final predictor can be trained on all the available data.

IV. Results

Results differ depending on how ATC reroute advisory acceptance and rejection are defined. All those accepted DWR reroute advisories that had no Center route amendment implemented within 30 minutes of it being accepted by the dispatcher are classed as rejected by the controller. Similarly, all those observed Center route amendments that are made within 30 minutes of a DWR route advisory being accepted by a dispatcher are classed as accepted by the controller. This includes observed Center route amendments that match the DWR reroute advisory accepted by the dispatcher, and observed Center route amendments that are a modification of the accepted DWR route advisory requested. In this latter case, the original dispatcher-accepted reroute advisory that was modified by the controller before being implemented can be classed as either accepted or rejected by the controller, depending on how the predictor is to be used. From the airline perspective, these reroute advisories did lead to delay savings in the form of a Center route amendment, even if the actual routing ultimately differed from the accepted advisory. This would suggest that, were the predictor to be used by airlines, it may make sense to class these original dispatcher-accepted reroute advisories as ultimately accepted by the controller. However, from a system perspective, there could be workload benefits from generating reroute advisories that are not only acceptable to controllers after modification, but acceptable to controllers before any modification. In this case, for reroute advisories that are modified, only the implemented Center route amendment is classed as ATC accepted, while both the original dispatcher-accepted reroute advisory that is later modified by the controller, and reroute advisories that do not result in any Center route amendment, are classed as ATC rejected. This is how ATC acceptance and rejection of reroute advisories is defined in Ref. 10. For the results presented in this paper, we apply this latter definition. Feature, model and parameter selection results, applied to the full dataset, are presented below, followed by model validation results using nested cross-validation. Finally, results applying a single class classifier are also presented for comparison.

A. Feature Selection

Feature selection results, calculated for the full dataset with the positive class defined for routes that are rejected or modified by ATC, and the negative class defined for routes that are accepted by ATC, are presented in Table 1 and Figure 3. The maximum number of observations included in all steps of the feature selection is 544, of which 40% are positive and 60% negative. Because some observations have missing data for some features, each step of the feature selection has a different number of observations, ranging from 317 to 544, depending on the feature set used. The first column of Table 1 shows a list of all features considered. The second column shows the F-Scores for a random forest, trained on each of these features separately. The random forest is trained with 40 decision trees, no constraint on the tree depth, and a minimum terminal node size of 1. The number of features randomly sampled as candidates at each split, with replacement, is the floor of the square root of the total number of features (i.e., $\lfloor\sqrt{p}\rfloor$, where p is the total number of features). The highest F-Score in column 2 (0.674) is for the feature *Hist. Routing Count by Segment*. This indicates that historical usage is the strongest single contributor to reroute operational acceptance tested, as suggested in Ref. 10. Using the greedy forward search approach to feature selection, *Hist. Routing Count by Segment* is therefore retained as a feature, and the model performance reevaluated with each remaining feature added to the feature set separately. Results are shown in column 3 of Table 1. Including *Time to Exit MSS* in the feature set with *Hist. Routing Count by Segment* yields the highest F-Score for 2 features (0.719), indicating that the location of the aircraft in the maneuver start sector when the reroute is requested is relevant to the reroute operational acceptability, as suggested in Ref. 8. Forward search is used to add all remaining features, with results shown in Table 1.

The feature yielding the highest F-Score (0.815) contains 7 features, as follows:

1. Hist. Routing Count by Segment;
2. Time to Exit MSS;
3. No. Downstream Sectors;
4. Direct Routing;
5. Dist. to Exit MSS;

Table 1. Forward search feature selection results (F-Score), applying a random forest for the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014. Order of features selected in parentheses in column 1. (MSS: Maneuver Start Sector; D/C: Demand over Capacity).

	1 Feature	2 Features*	3 Features*	4 Features*	5 Features*	6 Features*	7 Features*	8 Features*	9 Features*	10 Features*
Hist. Count (Full Routing) **(10)**	0.648	0.695	0.753	0.771	0.764	0.766	0.801	0.775	0.767	**0.780**
Hist. Routing Count by Segment **(1)**	**0.674**	-	-	-	-	-	-	-	-	-
Direct Routing **(4)**	0.387	0.597	0.705	**0.775**	-	-	-	-	-	-
No. Sectors Over **(9)**	NA	0.599	0.693	0.743	0.746	0.766	0.809	0.783	**0.797**	-
Max D/C Ratio **(8)**	0.255	0.664	0.751	0.773	0.769	0.789	0.772	**0.784**	-	-
MSS Over Capacity **(7)**	NA	0.583	0.674	0.744	0.758	0.782	**0.815**	-	-	-
MSS D/C Ratio **(6)**	0.381	0.660	0.749	0.758	0.773	**0.796**	-	-	-	-
No. Downstream Sectors **(3)**	0.484	0.667	**0.755**	-	-	-	-	-	-	-
Time to Exit MSS **(2)**	0.497	**0.719**	-	-	-	-	-	-	-	-
Dist. to Exit MSS **(5)**	0.467	0.665	0.719	0.761	**0.789**	-	-	-	-	-

* Includes feature set with highest F-Score from previous column

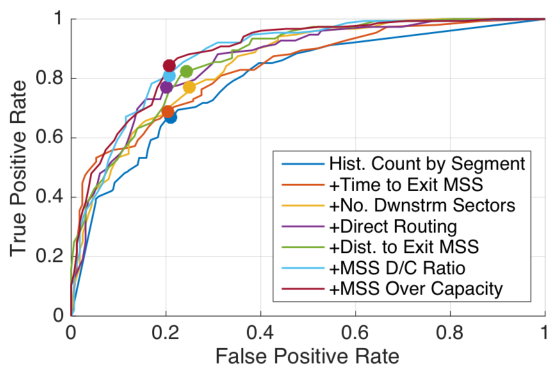

Figure 3. ROC curves for forward search feature selection, applying a random forest in the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014. Filled circles represent the points with 0.5 discrimination threshold presented in Table 1. (MSS: Maneuver Start Sector; D/C: Demand over Capacity)

6. MSS D/C Ratio; and
7. MSS Over Capacity.

The inclusion of any more features than those listed above did not improve the model performance. Note that this feature set includes features describing historical usage (1), location of the aircraft in the maneuver start sector (2 and 5), reroute deviation (3 and 4), and demand (6 and 7). Significantly, the selected features describing demand only describe demand in the maneuver start sector, and not in downstream sectors. This means that downstream sector demand did not manifest in the data, suggesting that downstream demand may not be a significant factor in the operational acceptability of reroutes. This may be because controllers have limited information on downstream sector demand when making decisions.

Figure 3 shows the ROC curve for the random forest, for each of the feature sets identified to maximize F-Score in Table 1, from 1 to 7 features. There is a clear progression as features are added, shifting the ROC curve towards the upper left (a perfect model would be in the upper left corner).

With the feature set selected, the data-mining algorithm that maximizes F-Score can be selected.

B. Model Selection

Model selection results are presented in Table 2 and Figure 4. The features used are the 7 features listed in Section IV-A above. The number of observations included is 317, of which 48% are positive and 52% negative. The SVM applies a sigmoid kernel; the random forest is as described in Section IV-A; and AdaBoost is applied with 40 decision trees as weak learners. The best performing model in terms of F-Score is the random forest (0.815), with AdaBoost the second best (0.766). The random forest also performs best under all other metrics.

Table 2. Model selection results (F-Score) for the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014.

	Logistic Regression	Decision Tree	SVM	Random Forest	AdaBoost
Accuracy	0.732	0.735	0.685	0.817	0.776
Misclassification Error	0.268	0.265	0.315	0.183	0.224
True Positive Rate	0.711	0.750	0.632	0.842	0.763
True Negative Rate	0.752	0.721	0.733	0.794	0.788
Precision	0.725	0.713	0.686	0.790	0.768
F-score	**0.718**	**0.731**	**0.658**	**0.815**	**0.766**
Area Under ROC	0.818	0.767	0.770	0.886	0.864
Average Precision	0.776	0.687	0.735	0.870	0.826

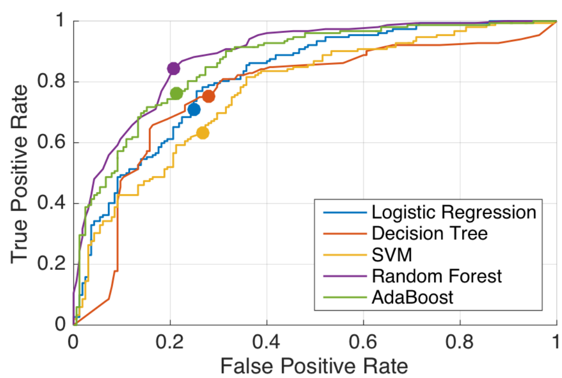

Figure 4. ROC curves for model selection, in the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014. Filled circles represent the points with 0.5 discrimination threshold presented in Table 2.

Figure 4 shows the ROC curve for each of the models, where it is clear that the random forest provides the best performance across a range of discrimination thresholds.

C. Parameter Selection

In a random forest, a key parameter that must be set is the number of decision trees, i.e., the weak learners used. It is important that the number of trees is large to ensure that all observations are used, preferably more than once, and that all features are used, again preferably more than once. If the number of trees is too small the predictive power of the random forest decreases. Because the problem described in this paper has a small number of features (7), and the number of observations is also not large (317), a very large number of trees is unnecessary. 10-fold cross-validation is used to compare the performance of the model with a range of values for this parameter, from 20 to 100, as shown in Table 3 and Figure 5. In Table 3, the number of trees that maximizes the model F-Score is 40. This value also maximizes all other metrics. The ROC curve for the different parameter values in Figure 5 shows that the results vary very little with number of trees.

Three other key parameters that require setting in random forests are the maximum tree depth for the weak learners, the minimum terminal node size, and the number of features randomly sampled as candidates at each split. Because of the small data size and small number of features, however, no constraint is imposed on the tree depth in this paper, and the minimum terminal node size is fixed at 1. The number of features randomly sampled as candidates at each split, with replacement, is also fixed, at $\lfloor\sqrt{p}\rfloor$, where p is the total number of features. Given p=7, 2 features are randomly sampled at each split.

In summary, the model development suggests that the best performing model for the prediction of ATC operational acceptability of reroute requests in ZFW and its adjacent Centers is as follows:

- Features: Hist. Routing Count by Segment; Time to Exit MSS; No. of Downstream Sectors; Direct Routing; Dist. to Exit MSS; MSS D/C Ratio; and MSS Over Capacity
- Model: Random Forest
- Number of Trees: 40

Table 3. Parameter selection results (F-Score) for a random forest with the specified number of trees, for the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014.

	Number of Trees								
	20	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Accuracy	0.798	0.801	0.817	0.798	0.798	0.795	0.808	0.792	0.785
Misclassification Error	0.202	0.199	0.183	0.202	0.202	0.205	0.192	0.208	0.215
True Positive Rate	0.829	0.816	0.842	0.809	0.822	0.816	0.829	0.803	0.796
True Negative Rate	0.770	0.788	0.794	0.788	0.776	0.776	0.788	0.782	0.776
Precision	0.768	0.780	0.790	0.778	0.772	0.770	0.783	0.772	0.766
F-score	**0.797**	**0.797**	**0.815**	**0.794**	**0.796**	**0.792**	**0.805**	**0.787**	**0.781**
Area Under ROC	0.877	0.871	0.886	0.875	0.870	0.878	0.883	0.874	0.867
Average Precision	0.860	0.820	0.870	0.833	0.844	0.854	0.863	0.840	0.835

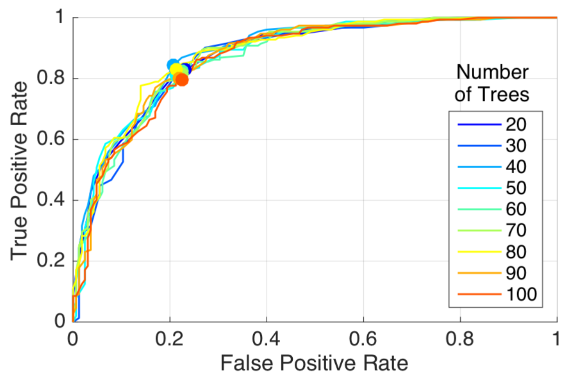

Figure 5. ROC curve for a random forest with the specified number of trees, in the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014. Filled circles represent the points with 0.5 discrimination threshold presented in Table 3.

D. Model Validation

Table 4 and Figure 6 present the results of the nested 10-fold cross validation, including feature, model and parameter selection as described in Section III-D. The validation dataset is reasonably well balanced (out of 317 observations, 40% are positive, and 60% negative). *Accuracy* measures the fraction of correct predictions from all predictions made. *Misclassification Error* measures the fraction of incorrect predictions from all predictions made (i.e., 1 – *Accuracy*). *True Positive Rate* refers to *recall*, defined in Section III-C. *True Negative Rate* refers to the number of elements correctly labeled by the model as belonging to the negative class divided by the total number of elements that actually belong to the negative class (also called *Specificity*). *Precision*, *F-Score* and *AUC* are as defined in Section III-C.

The model performance is reasonable, with an F-Score of 0.767. Other performance metrics such as accuracy, precision and AUC are also reasonable. Accuracy particularly can be compared directly to the percentage of the majority class (60%). This can be considered to be a baseline accuracy which could be achieved with a model that simply specifies all outputs to be the majority class (i.e., ATC accepted), without consideration of any feature values. In comparison, the model developed has an accuracy of 74.4%, which is 14.4% better than this baseline. The true positive rate in Table 4 is particularly high (0.875), indicating that the approach is effective at accurately predicting whether a reroute is likely to be rejected or modified by ATC. As stated above, this is of particular importance in the application considered, because misclassification of route rejection by ATC could lead to increased controller workload. The true negative rate is lower (0.624), indicating that the approach is less effective at accurately predicting whether a reroute is likely to be accepted by ATC. If it is considered of higher importance to avoid misclassifying reroute acceptance, because of the associated losses in efficiency, the model can be retrained with reroute acceptance set as the positive class.

Choice of a different discrimination threshold could also alter the tradeoff between true positive rate and true negative rate, moving along the ROC curve shown in Figure 6 (an increase in discrimination threshold would lead to a shift down and left along the curve, while a decrease in discrimination threshold would lead to a shift up and right

Table 4. Nested cross validation results for the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014.

	Nested Cross-Validation
Accuracy	0.744
Misclassification Error	0.256
True Positive Rate/Recall	0.875
True Negative Rate	0.624
Precision	0.682
F-Score	**0.767**
AUC	0.814

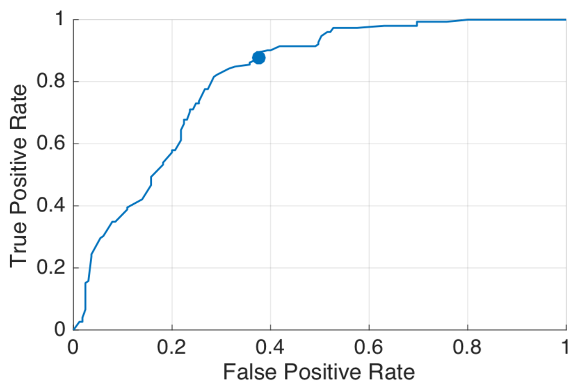

Figure 6. ROC curve for nested cross validation results for the prediction of ATC operational acceptability of DWR advisory reroutes from May to September 2014. Filled circle represents the point with 0.5 discrimination threshold presented in Table 4.

along the curve). The area under the ROC curve (AUC), however, is high (0.814), indicating good model performance across a range of discrimination thresholds.

As part of the model validation, reroute acceptability by ATC was also discussed with a retired traffic manager familiar with ZFW and the adjacent Center airspace. He suggested that the most important factors in determining the operational acceptability of a requested reroute in ZFW while airborne were (1) characteristics of the underlying airspace, such as location of airport arrival and departure flows; (2) the controller's individual workload levels; and (3) time to hand-off to the next sector. The importance of characteristics of the underlying airspace (1) is consistent with identifying *Hist. Routing Count by Segment* as a key feature in the feature selection results described in Section IV-A. Controller workload (2) would be impacted by demand levels in the maneuver start sector, which is consistent with identifying *MSS D/C Ratio* and *MSS Over Capacity* as key features in Section IV-A. Time to hand-off (3) would be impacted by the location of the aircraft in the maneuver start sector, which is consistent with identifying *Time to Exit MSS* and *Dist. to Exit MSS* as key features in Section IV-A.

Reroute deviation was considered by the retired traffic manager to be of lower importance. Two of the features identified in Section IV-A, *No. Downstream Sectors* and *Direct Routing*, describe reroute deviation. *No. Downstream Sectors* may be capturing the need for flight plan waypoints at center boundary crossings, or similar geographical airspace related requirements. It is more likely that reroutes traversing a high number of sectors would not meet this requirement, requiring reroute modification. *Direct Routing* may be capturing the historical usage of the reroute, since direct routings in this dataset saw higher historical usage than routings via an auxiliary waypoint. As described in Ref. 10 this is because auxiliary waypoints are primarily used to avoid weather, and weather is highly variable. In the version of DWR fielded, many of the routings used to avoid different weather cells via auxiliary waypoints are unique, not appearing in the historical data at all.

The retired traffic manager also confirmed that controllers do not have information about downstream demand, which would make features describing downstream demand, such as *No. Sectors Over* and *Max D/C Ratio*, of little relevance. This is consistent with the feature selection results in Section IV-A, where these features were not identified as being important.

E. Single Class Classifier

It is not obvious how the model developed in this paper can be extended to other regions and applications, because of the limited data for rejected reroute advisories on which to train a two-class classifier. Alternative approaches exist, such as single-class classification, for which only a single class of observations are required in the form of Center route amendments. These are widely available. The performance of single class classifiers is, however, significantly poorer than the two-class classifiers developed in this paper. Developing a single-class classifier for the problem considered in Sections IV-A to IV-D, using a Support Vector Machine (SVM) with sigmoid kernel, yields a very low F-Score of 0.314 when model results are compared to actual route acceptance. This is significantly lower than the 0.815 model development F-Score for the two-class classifier using a random forest (Table 2), or even the 0.658 F-Score for the two-class classifier using a comparable SVM. The rate at which the single-class classifier predicts reroute acceptance by ATC is actually high – 0.879 – but the prediction of reroute rejection by ATC – at a rate of only 0.211 – leads to the low F-Score. This is expected since it is route rejection for which no training data is used. This is a problem if it is the prediction of route rejection by ATC that is considered of highest importance, as discussed in Section III-C. However, in applications for which the prediction of route acceptance is most important, this could provide a viable alternative. Other algorithms will be considered in future work. However, ultimately, most beneficial to the building of an effective predictor of route operational acceptability would be better data for the two-class classification problem, both in terms of geographical coverage (beyond ZFW and adjacent Centers) and in terms of quality. This latter point is of particular importance. The data from the 2014 DWR trial does not include information describing exactly what route was requested of the controller, or when the route was requested. While proxies for this exist in the data used in this paper, discrepancies between these and what was really seen by the controller erode the performance of the model. Hence, future work should focus on identifying or generating more complete sources of data.

V. Conclusions

In this paper, the authors developed a predictor of operational acceptability for airborne reroute advisories. This was based on applying data mining techniques to flight plan amendment data and data from a trial of the NASA developed DWR tool at American Airlines in 2014. Routes implemented in the form of a Center route amendment within 30 minutes of the reroute advisory being accepted by the airline dispatcher were classed as operationally acceptable to ATC. Routes are classed as operationally unacceptable to ATC if it they are either modified by the controller before being implemented in the form of a Center route amendment, or do not result in any Center route amendment being implemented at all. Features used in the development of the predictor include data on historical usage of the advised route, demand levels in the maneuver start sector and downstream, route deviation from the original flight plan, and the location of the maneuver start point in the maneuver start sector. Features, models and key model parameters were selected using 10-fold cross validation, while nested cross validation was used to validate the model performance, giving a reasonable F-Score of 0.77. Importantly, the rate at which the classifier predicts whether a reroute is likely to be rejected or modified by ATC is high: 0.88. This is of particular importance in the application considered, because misclassification of route rejection by ATC could lead to increased controller workload.

Features identified as relevant include:

(1) Hist. Routing Count by Segment: A route's historical routing count by segment, based on the counts of the route's component segments from historical data;
(2) Time to Exit MSS: The time between when the reroute advisory is accepted by the airline dispatcher and when the flight exits the maneuver start sector;
(3) No. Downstream Sectors: The number of downstream sectors between the maneuver start point and return capture fix;
(4) Direct Routing: Whether the routing is direct to the return capture fix or via one or more auxiliary waypoints;
(5) Dist. to Exit MSS: The distance between the maneuver start point and the maneuver start sector boundary, along the requested reroute;
(6) MSS D/C Ratio: The ratio of demand to capacity (as defined by the Monitor Alert Parameter) for the maneuver start sector; and
(7) MSS Over Capacity: Explicitly whether the maneuver start sector is over capacity, or not, when the maneuver is to start.

Through discussions with a retired traffic manager, it was confirmed that the majority of these features would be expected to be relevant to the operational acceptability of requested airborne reroutes in ZFW and adjacent Centers. The model found to provide best performance was a random forest with 40 trees.

The application of data mining to the DWR trial data presented in this paper provides significant results indicating that reroute advisory operational acceptability is indeed predictable, and that, with suitable data, models can be trained to predict reroute advisory operational acceptability. Such models may ultimately be used to inform route selection by decision support tools, contributing to the development of increasingly autonomous systems that are capable of routing aircraft with less human input than is currently the case.

It is also important to highlight that one of the advantages of increasingly autonomous systems is that they can enable efficiency improvements that would not be possible without automation. This could include the enabling of control actions that would not be considered operationally acceptable to controllers today, but which are safe. One example would be to allow reroutes through regions of airspace, like arrival corridors, that are generally protected by controllers, if it can be confirmed that this would be safe because no arrival traffic is present at that time. In today's operations, controllers are unlikely to do this because they have limited information about demand levels in other sectors. The development of a tool solely using data mining based on past operations would miss this opportunity. Data mining should therefore be combined with other approaches, such as optimization, to take full advantage of opportunities for efficiency improvements.

Acknowledgements

The authors would like to acknowledge helpful comments from Heather Arneson, Deepak Kulkarni and Yao Wang of NASA Ames Research Center, and feedback on the identified features from retired ZFW traffic manager Danny Vincent.

References

[1] National Aeronautics and Space Administration, "NASA Aeronautics Strategic Implementation Plan," NASA Aeronautics Research Mission Directorate, Washington DC, 2016.

[2] McNally, D., Sheth, K., Gong, C., Love, J., Lee, C., Sahlman, S., Cheng, J., "Dynamic Weather Routes: A Weather Avoidance System for Near-Term Trajectory-Based Operations," *28th International Congress of the Aeronautical Sciences*, Brisbane Australia, September 2012.

[3] McNally, D., Sheth, K., Gong, C., Sterenchuk, M., Sahlman, S., Hinton, S., Lee, C., Shih, F-T., "Dynamic Weather Routes: Two Years of Operational Testing at American Airlines," *Air Traffic Control Quarterly*, Vol. 23, No. 1, pp. 55-81, 2015.

[4] Sheth, K., Clymer, A., Morando, A., and Shih, F., "Analysis of Multi-Flight Common Routes for Traffic Flow Management," *16th AIAA Aviation Technology, Integration, and Operations Conference*, Washington, DC, p. 13. 2016.

[5] Gong, C., McNally D., and Lee C-H., "Dynamic arrival routes: A trajectory-based weather avoidance system for merging arrivals and metering." In *AIAA Aviation Technology, Integration, and Operations Conference*, p. 3394. 2015.

[6] Zelinski, S., Xue, M., and Bassett P., "Optimized Route Capability (ORC) Intelligent Offloading of Congested Arrival Routes." In *16th AIAA Aviation Technology, Integration, and Operations Conference*, p. 4357. 2016.

[7] Ballin, M.G., and Wing, D.J., "Traffic Aware Strategic Aircrew Requests (TASAR)," *12th AIAA Aviation Technology, Integration, and Operations Conference,* Indianapolis, IN, 2012.

[8] Idris H., Enea G., Burke K., and Wing D., "Assessment of Air Traffic Controller Acceptability of Aircrew Route Change Requests", to be presented at the FAA Eurocontrol, ATM R&D Seminar, Seattle, WA 2017.

[9] Kim, A., and Hansen, M., "A framework for the assessment of collaborative en route resource allocation strategies," *Transportation Research Part C: Emerging Technologies* 33 (2013): 324-339.

[10] Evans, A.D., Sridhar, B., and McNally, D., "Improving Operational Acceptability of Dynamic Weather Routes Through Analysis of Commonly Used Routings," In *16th AIAA Aviation Technology, Integration, and Operations Conference*, p. 3600. 2016.

[11] Taylor, C., and Wanke, C., "Dynamically Generating Operationally Acceptable Route Alternatives Using Simulated Annealing," *Air Traffic Control Quarterly*, Vol. 20, No. 1, pp. 97-121, 2012.

[12] Taylor, C., Wanke, C., "Dynamic Generation of Operationally Acceptable Reroutes," *9th AIAA Aviation Technology, Integration, and Operations Conference*, Hilton Head, SC, 21-23 September 2009.

[13] Stewart, T., Askey, L., and Hokit, M., "A Concept for Tactical Reroute Generation, Evaluation, and Coordination," *12th AIAA Aviation Technology, Integration, and Operations Conference*, Indianapolis, IN, 17-19 September 2012.

[14] Smith, D.A., and Sherry, L., "Decision support tool for predicting aircraft arrival rates, Ground Delay Programs, and airport delays from weather forecasts," *Proc. of International Conference on Research in Air Transportation*, Fairfax, VA, February 2008.

[15] Pfeil, D.M., and Balakrishnan, H., "Identification of robust terminal-area routes in convective weather," *Transportation Science*, 2011.

[16] Buxi, G., and Hansen, M., "Generating probabilistic capacity profiles from weather forecast: A design-of-experiment approach," *Proc. of USA/Europe Air Traffic Management Research & Development Seminar*, Berlin, Germany, June 2011.

[17] Wang, Y., and Kulkarni, D., "Modeling weather impact on Ground Delay Programs," *SAE Journal of Aerospace*, 4(2):1207–1215, November 2011.

[18] Bloem, M., Hattaway, D., and Bambos, N., "Evaluation of algorithms for a miles-in-trail decision support tool," *International Conference on Research in Air Transportation*, Berkeley, CA, May 2012.

[19] Grabbe, S.R., Sridhar, B., Mukherjee, A., "Clustering Days with Similar Airport Weather Conditions," *14th AIAA Aviation Technology, Integration, and Operations Conference*, Atlanta, GA, June 2014.

[20] Kulkarni, D., Wang, Y., Sridhar, B., "Analysis of Airport Ground Delay Program Decisions Using Data Mining Techniques," *14th AIAA Aviation Technology, Integration, and Operations Conference*, Atlanta, GA, June 2014.

[21] Deng, K., "OMEGA: On-line Memory-based General Purpose System Classifier", The Robotics Institute, School of Computer Science, Carnegie Mellon University Thesis, CMU-RI-TR-98-33, 1998.

[22] Breiman, L., "Random Forests," *Machine Learning,* (2001) 45(1), 5-32.

[23] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, 2nd ed., Springer, ISBN 0-387-95284-5, 2008, pp. 587-588

[24] Freund, Y., and Schapire, R.E., "A decision-theoretic generalization of on-line learning and an application to boosting," In *European conference on computational learning theory* (pp. 23-37). Springer Berlin Heidelberg, 1995.

[25] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., and Napolitano, A., "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40, no. 1 (2010): 185-197.

[26] Menardi, G. and Torelli, N., "Training and assessing classification rules with imbalanced data," *Data Mining and Knowledge Discovery*, 28 (2014):92–122.

[27] Lunardon, N., Menardi, G., and Torelli, N., "ROSE: a Package for Binary Imbalanced Learning," *R Journal*, (2014) 6:82–92.

[28] Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research* 16 (2002): 321-357.

[29] Fawcett, T., "An introduction to ROC analysis," *Pattern Recognition Letters*, (2006) 27, 861–874.