A Tree-Based Ensemble Method for the Prediction and Uncertainty Quantification of Aircraft Landing Times*

2.2

Yan Glina†, Richard Jordan, Mariya Ishutkina
MIT Lincoln Laboratory
Lexington, MA, USA
yglina@ll.mit.edu, rjordan@ll.mit.edu, ishutkina@ll.mit.edu

***Abstract* – Accurate aircraft landing time predictions provide situational awareness for air traffic controllers, enable decision support algorithms and gate management planning. This paper presents a new approach for estimation of landing times using a tree-based ensemble method, namely Quantile Regression Forests. This method is suitable for real-time applications, provides robust and accurate predictions of landing times, and yields prediction intervals for individual flights, which provide a natural way of quantifying uncertainty. The approach was tested for arrivals at Dallas/Fort Worth International Airport over a range of days with a variety of operational conditions.**

Keywords – Air Traffic Control; Decision Support; Random Forests; Quantile Regression Forests

1. INTRODUCTION

The United States has the largest and most complex air transportation system in the world: air traffic controllers are responsible for handling safely and efficiently more than 5,000 flights at peak times. While traffic growth has been low in the past few years, the official 2011 forecast from the Federal Aviation Administration (FAA) states that domestic enplanements are projected to grow on average 2.5% per year during the next twenty years (FAA Fact Sheet, 2011). To meet future demand and avoid gridlock in the sky and on airport surfaces, the FAA is currently in the process of modernizing the National Airspace System (NAS) through several NextGen initiatives. As part of these initiatives, decision support tools (DSTs) are being developed to improve airport surface operations. In this paper we will discuss the development of a model for prediction of arrival landing times (also known as wheels-on time or Estimatd Time of Arrival [ETA]). Accurate predictions of ETA will enable aircraft sequence optimization DSTs aimed at decreasing delays, fuel burn and emissions. In addition, accurate aircraft arrival time predictions, when displayed on sequence timelines, provide valuable situational awareness to air traffic controllers.

NASA's Traffic Management Advisor (TMA) provides ETAs along with many other capabilities (NASA Traffic Management Advisor, 2011). However, this tool, to date, has limited deployment and relies on knowledge of routing decisions made by controllers to make accurate predictions. TMA ETA predictions are based on detailed deterministic physics-based models that incorporate airport configuration, winds aloft, aircraft types and separation and/or flow rate constraints. However, TMA does not explicitly account for uncertainties inherent in real operations, such as deviations from standard arrival routes. In addition, aircraft seldom fly at exactly the modeled (deterministic) speeds. Because of these and other sources of stochasticity, ETAs should really be thought of as random variables. With that in mind, we present herein a modeling framework that provides predictions of the probability distribution for each individual ETA. From the probability distributions, one can determine the expected (mean) or median ETA for a given flight. However, other valuable information can be extracted, such as quantiles of the distribution and prediction intervals, which provide a natural estimate of the degree confidence that should be attached to the mean or median ETA. We expect that explicit quantification of the uncertainties associated with individual ETAs will prove valuable for the development and implementation of advanced DSTs that have been envisioned for NextGen. It has been shown, for example, that deterministic aircraft sequence optimization algorithms may provide only limited benefits when confronted with realistic uncertainties in wheels-on times and aircraft taxi times (MIT Lincoln Laboratory, 2012). While it may be possible to design sequencing algorithms that are robust to the uncertainties, it is first necessary to have models that can accurately represent the uncertainties.

The approach presented in this paper is based upon a regression tree ensemble method, Quantile Regression Forests (QRF) (Meinshausen, 2006) which is an extension of Random Forests (RF) (Breiman, 2001). The

* This work was sponsored by the National Aeronautics and Space Administration (NASA) under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

† *Corresponding author address:* Yan Glina, MIT Lincoln Laboratory, Lexington, MA 02420; e-mail: yglina@ll.mit.edu

method provides accurate point predictions of aircraft landing times and, in addition, conditional probability distributions for the ETA of each individual flight.

The paper is organized as follows. In Section II, we provide a brief introduction of the RF and QRF algorithms. In Section III the wheels-on model specifics are provided, with focus on a case study of Dallas/Fort Worth International Airport. Section IV details the results of the computational experiments performed to validate our approach. Conclusions are presented in Section V.

2. METHOD DESCRIPTION

The method upon which our wheels-on prediction algorithm is based is the Quantile Regression Forests algorithm (QRF) proposed by Meinshausen (Meinshausen, 2006). QRF is an extension of Random Forests (RF), which is an ensemble of classification and regression trees (CART) (Breiman et al., 1984). The basic strategy of CART is to partition a sample of data using binary rules to split parent nodes in such a way that the child nodes are more homogenous than the parent nodes. CART can be applied in a classification or a regression context and can handle very high-dimensional data sets. CART models have the advantage of interpretability, especially for relatively small trees. The major disadvantage of CART is its instability. A small change in the training sample can result in substantial changes in the predictor tree. This can result in poor predictive accuracy when the model is applied to new data that is independent of the training data on which the model was constructed. RF has been designed to overcome this fundamental limitation. It uses randomly generated CART predictors as weak learners in an ensemble learner.

Let (x_i, y_i), $i = 1, \ldots, N$ be the training data. The predictor variable vector x_i can be comprised of real-valued and/or categorical variables. We will assume that the response y_i is real-valued, as we are concerned herein with regression problems. In CART, the prediction of a tree given the new predictor variable vector $X=x$ is

$$T(x,\theta) = \sum_{i=1}^{N} w_i(x,\theta) y_i,$$

where θ represents the parameters (split points) defining how the tree is constructed and $w_i(x,\theta)$ are weights such that $w_i(x,\theta) > 0$ if the observation x_i is in the same terminal node as x and $w_i(x,\theta) = 0$ otherwise. The weights are normalized so that they sum to 1. Specifically, if $L(x,\theta)$ is the leaf (i.e., terminal node) in which x lands, then

$$w_i(x,\theta) = \frac{I\{x_i \in L(x,\theta)\}}{\#\{j \mid x_j \in L(x,\theta)\}}. \qquad (1)$$

In equation (1), $I\{x_i \in L(x,\theta)\} = 1$ if an only if x_i is in the leaf $L(x,\theta)$ and the denominator is the total number of training points that are in this leaf.

RF consists of a collection of CART predictors $T(x, \theta_k)$, $k = 1, \ldots K$, where the parameters θ_k are independent, identically distributed random vectors that determine how a tree is constructed and K is the number of trees. The RF approach employs two levels of randomization in the construction of individual trees: bagging (or bootstrap aggregation) (Breiman, 1996) and random selection of a subset of predictor variables to be considered for the splitting of nodes. The size of the random subset, denoted by *mtry*, is a tuning parameter of the model, though results are generally nearly optimal over a wide range of this parameter.

It is important to note that RF has good computational performance. Its complexity is of O(K*m*log(m)) during training and O(N*K*log(m)) during testing, where K = number of trees, N = number of instances. m = number of instances per node (Carrasquilla, 2010).

For RF the conditional mean $E(Y \mid X = x)$ is estimated as the average prediction over the K trees. Define

$$w_i(x) = \frac{1}{K}\sum_{k=1}^{K} w_i(x,\theta_k)$$

so w_i is the average of the weights associated with the individual trees. Then the (deterministic) RF prediction for $E(Y \mid X = x)$ is

$$T(x) = \sum_{i=1}^{N} w_i(x) y_i \,. \qquad (2)$$

Thus, the prediction is a weighted average over all observations and the weights depend on the covariate $X = x$. As shown in (Lin et al., 2006) the weights $w_i(x)$ are largest for those i where the conditional distribution of Y given $X = x_i$ is most similar to the distribution of Y given $X = x$.

The idea of Meinshausen (Meinshausen, 2006) is that one could expect that the weighted observations can be used to approximate not only the conditional mean, but also the entire conditional distribution

$$F(y \mid X = x) = Prob(Y \le y \mid X = x) = E(I_{\{Y \le y\}} \mid X = x)$$

where $I_{\{Y \le y\}}$ is the indicator function that is equal to 1 if $Y \le y$ and 0 if $Y > y$. Indeed, a natural estimate of $F(y \mid X = x)$ is the following weighted mean over the observations:

$$\hat{F}(y \mid X = x) = \sum_{i=1}^{N} w_i(x) I_{\{y_i \le y\}}$$

The expression above is simply an analog of equation (2) for the estimate of the conditional mean. This is the QRF prediction for conditional distributions.

Given the estimate of the conditional distribution, it is then straightforward to extract estimates of the conditional median, higher order moments, conditional quantiles and prediction intervals of the form $[Q_{\square}(x), Q_{\square}(x)]$ can readily be extracted, where $Q_{\square}(x)$ is the $\square$-quantile for Y given $X = x$. For example, a 90% prediction interval is $[Q_{0.05}(x), Q_{0.95}(x)]$.

The width of this prediction interval can vary considerably with x, and the narrower the interval, the greater the reliability of the prediction. Thus, QRF offers a meaningful way in which to attach a measure of confidence to individual predictions.

3. DFW CASE STUDY

The case study presented in this paper is for Dallas/Fort Worth International Airport (DFW). The airport layout is shown schematically in Figure 1.

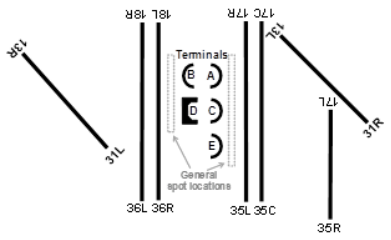


Figure 1: Dallas/Fort Worth airport operates multiple runways serving 19 airlines across 5 terminals.

DFW handles about 1,800 operations per day and is the third busiest airport in the United States when ranked by scheduled enplanements on U.S. airlines (U.S. Bureau of Transportation Statistics, 2011). The majority of the flights land or depart from the four parallel runways located around the central terminal area. The airport usually operates either in southflow or northflow configurations, depending on the prevalent winds. In southflow operations, the inner runways (17R and 18L) are typically used for departures, while the outer runways (17C, 18R, 13R, 17L and 13L) are typically used for arrivals. The surveillance tracks for southflow operations at DFW are shown in the statistical heat map in Figure 2. There are clear statistical flight path patterns observable in the figure, along with examples of behaviors that are quite difficult to predict. For example, there are parallel arrival tracks from the north for those flights landing on runways 18R, 17C and 17L and northwest tracks for arrivals landing on 13R. However, even these characteristic patterns have a large amount of variability. For example, there are elliptical tracks that are indicative of aircraft that were required to wait in a holding pattern before landing.

Several data sources were used for analysis of DFW operations. ASDI (Aircraft Situation Display to Industry) data was used to obtain latitude, longitude and altitude for each aircraft. ASDI has two sampling rates: within 60 nautical miles (NM) of the airport the aircraft are observed every 20 seconds, while outside of the range, locations are reported every 60 seconds. Spline interpolation and low-pass filtering were used to post-process the data to reduce noise. High resolution ASDE-X (Airport Surface Detection Equipment, Model X) data with update frequency of 1 second was used to derive wheels-on times for individual arrivals. ASPM (Aviation System Performance Metrics) throughput data was used to determine the runways used for operations.

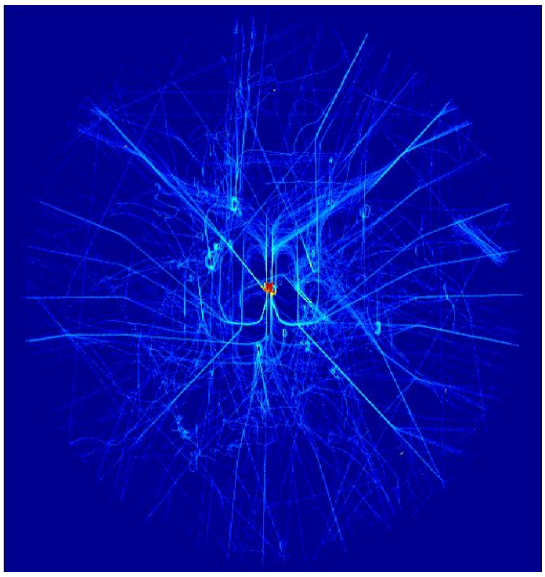


Figure 2: ASDE-X heat map of surveillance tracks for DFW operations for 1/3 day (log color scale) shows complexity of arrival trajectories, vectoring, and holding patterns.

The target variable used in the analysis was the Estimated Time of Arrival which is defined as the time when the aircraft crosses the landing runway threshold. Aircraft were observed at 60NM, 50NM, 40NM, 30NM, 20NM, 10NM and 3NM with a +/-1NM tolerance. The following predictor variables associated with the aircraft track data were used to construct training/testing exemplars: Euclidean distance from aircraft to the center of airport, Latitude/Longitude, Heading, Altitude, and Speed, as well as Track Start Location (latitude/longitude of the first observed point in ASDI track, most frequently at the airport of origin) and Sample Times (time past since/before aircraft is observed at 60NM from airport center, sampled at regular intervals of 15 seconds, at 10 reverse-chronological samples from each observation distance). The following additional predictor variables were used since they affect airport operations: Time of Day, Weather (Visual vs. Instrument Meteorological Conditions flag), Runway Availability (per-runway indicator functions, derived from ASPM throughput data). We did not use aerial congestion information in the construction of our data set for simplicity and independence of each data point; however, we realize that we may be able to improve our estimate further by including this information.

One of the outputs of the RF algorithm is its estimate of relative importance for each predictor variable. Figure 3 shows the RF-based feature importance ranking. As can be seen in the figure, the shortest-path distance between the aircraft's current position and the airport center is the dominant predictor variable. Altitude is also found to be important, but time of day (ToD) is relatively unimportant.

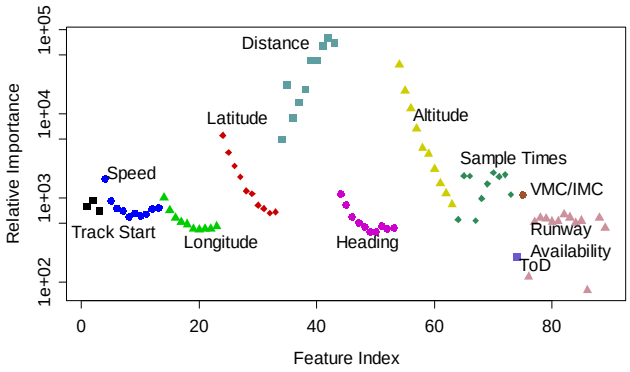


Figure 3: RF-based relative importance feature ranking for DFW ETA model shows variable groupings (by sample times) and individual variable utility.

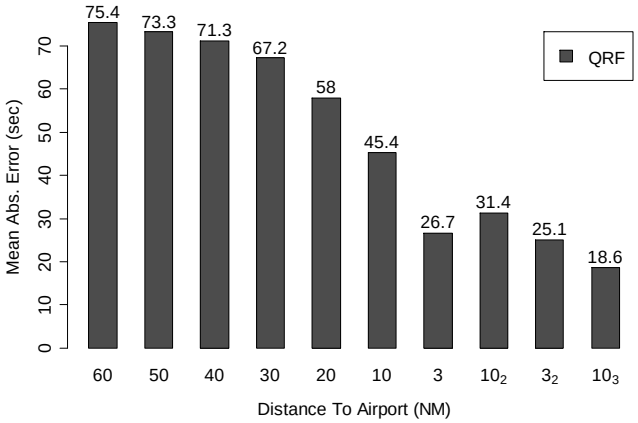


Figure 4: QRF Mean Absolute Errors of time-to-wheels-on estimates as function of distance to center of airport show overall performance and its improvement as aircraft approach runway.

Our data analysis showed that for some flights distance to the airport does not always decrease monotonically. This is the effect of tromboning, where aircraft must loop around the airport to make their final approach. The impact of this effect on prediction uncertainty is observable in the uncertainty histogram for 10NM (Figure 8). Such cases will appear more than once at particular distances (10NM, 3NM) and are indexed with a subscript (*e.g.,* 10_2) in this paper. This data processing approach is amenable to on-line implementation, as it uses a fixed number of features for the regression task, regardless of the position of the aircraft in the track. However, we expect that the inclusion of such data, usually considered outliers in other approaches, will somewhat reduce the accuracy of our predictions.

4. RESULTS

The data used for the computational experiments presented in this section spanned the period of five days (04/06/2011, 04/07/2011, 04/09/2011, 04/27/2011, and 06/02/2011). During those days, the airport operated in southflow runway configuration and the operating conditions were visual except on 04/07/2011 which had instrument meteorological conditions early in the morning (before 9am) and between noon and 1pm.

A total of 4011 unique cases were identified and separated randomly into 67% (2674 cases) training and 33% (1337 cases) testing data points. The Quantile Regression Forests algorithm was tested against this data set. The QRF algorithm was evaluated with the quantiles (0.05, 0.5, 0.95) to extract 90% prediction intervals, as well as an estimate of the median time until wheels-on.

Figures 4 and 5 summarize the performance of the QRF algorithm in predicting the ETA. Although we do not possess a large amount of TMA performance data, QRF point prediction accuracies appear comparable to those of TMA.

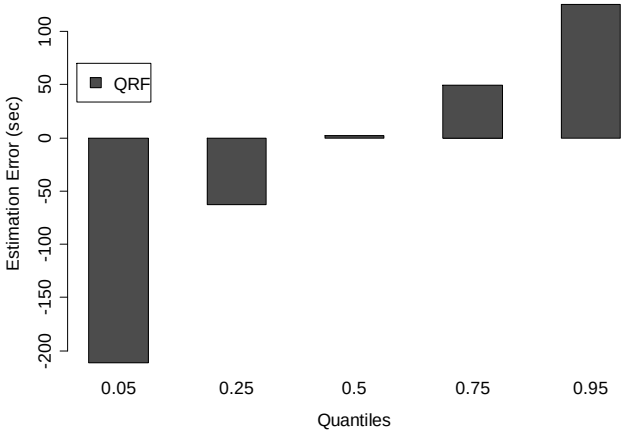
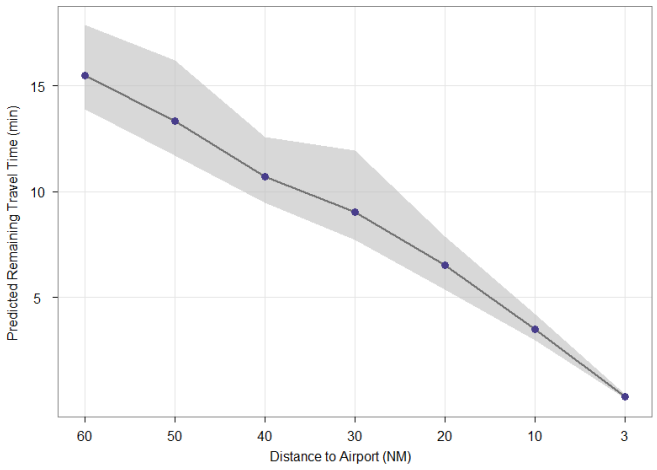


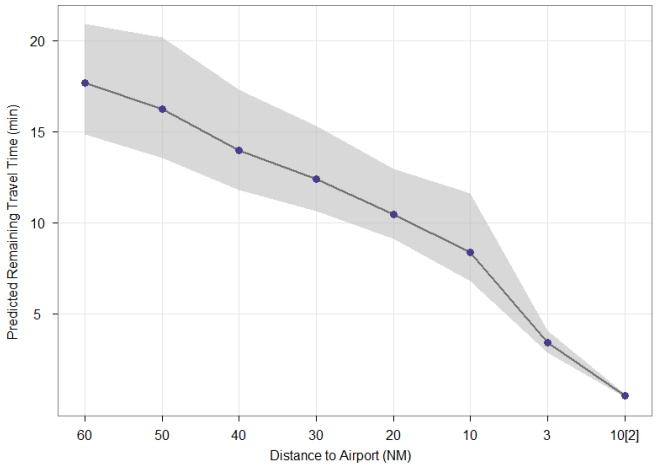
Figure 5: QRF error quantiles for 60-80NM range show excellent prediction characteristics even for aircraft at the edge of the surveillance region.

As described above, QRF also enables the estimation of prediction intervals. Figures 6 a) and b) show that the 90% prediction interval widths decrease as these arrivals near the landing runway. Note that the 90% prediction interval for the flight represented in Figure 6a) has a width of about 4 minutes 60 nm out, while at the same distance away, the 90% prediction interval for the flight represented by Figure 6b) is approximately 6 minutes. Thus the ETA for the first flight can be considered more predictable than that of the second flight at a distance of 60 nm from the runway. In fact, the prediction intervals for the first flight remain narrower than those for the second flight all the way into the runway, with the exception being at around 30 nm, where the prediction intervals have comparable widths. We believe that a closer examination of the differences

between prediction intervals for different flights could lead to valuable operational insights and suggest the inclusion of new predictor variables that could improve model accuracy.



a)



b)

Figure 6: Two distinct prediction interval estimates of arrival time. Both the arrival time estimates and the prediction interval envelopes vary significantly at various points on approach trajectory. Note the non-monotonic decrease of the width of the prediction interval in (a), inflecting at 40NM.

The distributions of the widths of the 90% prediction intervals provided by the QRF algorithm are shown in Figure 7 and Figure 8. As was explained in Section I, such uncertainty quantification is important for advanced decision support tools that have been envisioned for NextGen, such as aircraft sequence optimization, where failure to account for uncertainties can severely limit the benefits delivered.

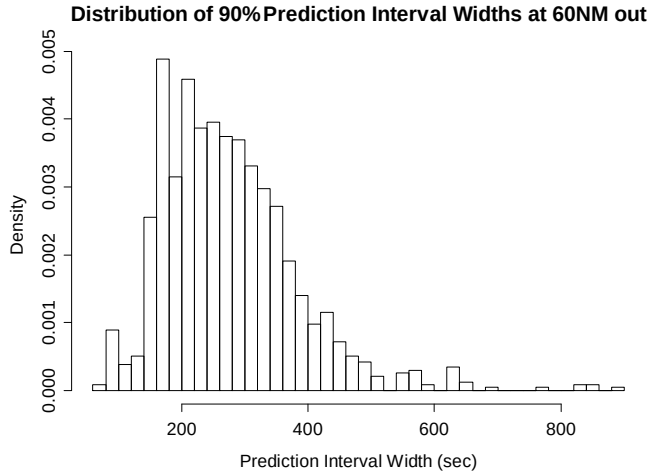


Figure 7: Prediction Interval Widths at 60NM encompass the entirety of approach trajectories. At this point the distribution of the 90% prediction interval widths is rather broad, reflecting primarily the diversity in approach trajectories and the resulting uncertainty in actual wheels-on times.

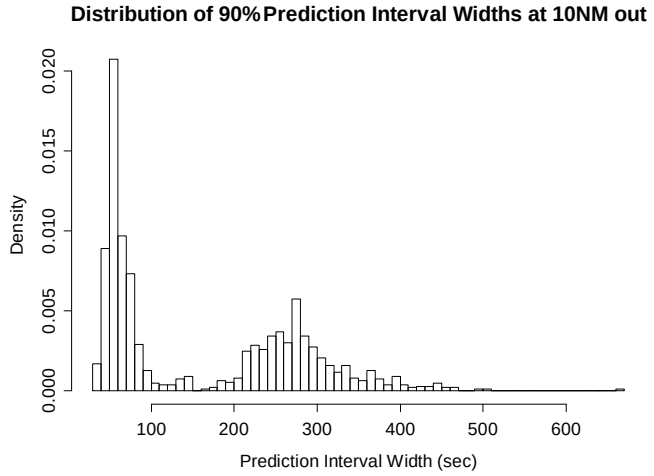


Figure 8: At 10NM, the prediction intervals are clearly separated into those aircraft taking a direct path into the runway (low uncertainty) vs. those tromboning / going around (more uncertainty).

5. Conclusions and Future Work

This paper presented recent work on the problem of Estimated Time of Arrival, which is an important sub-problem in the Air Traffic Control decision support domain. During our algorithmic trials, we have experimented with several Machine Learning approaches of the regression tree ensemble variety. In addition to producing high-fidelity predictions, these algorithms also deliver sufficiently high computational performance to be implemented in a real-time system. Such approaches

also benefit from being readily adaptable to operation in different environments; thus, we would expect comparable performance at different airports. As we have demonstrated, Quantile Regression Forests are especially promising, given the need for not only accurate point predictions of ETAs, but also a method for quantifying uncertainty. It is worth noting that the same technique that we propose here for ETA at the runway can be used to predict ETA for waypoints / fixes along the track, thereby improving en-route prediction of overall ETA, and allowing actions such as weather-related corridor reroutes to happen more efficiently.

We are currently investigating the use of additional features in the model, such as weather information and airspace/tarmac congestion variables; we expect that the inclusion of such variables will lead to increased accuracy of point predictions of ETAs, as well as tighter prediction intervals.

6. REFERENCES

Breiman, L., 1996: *Bagging Predictors*, Machine Learning 24: 123-140.

Breiman, L., 2001: "Random Forests," Machine Learning, 45(1):5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A., 1984: *Classification and Regression Trees*, Chapman & Hall/CRC.

Carrasquilla, U., 2010: "Benchmarking Algorithms for Detecting Anomalies in Large Datasets," CMG.

FAA Fact Sheet – FAA Forecast Fact Sheet –Fiscal Years 2011-31. Available from www.faa.gov. Last accessed on July 18, 2011.

Lin, Y. & Jeon, Y., 2006: *Random Forests and Adaptive Nearest Neighbors*, Journal of the American Statistical Association 101: 578-590.

Meinshausen, N., 2006: "Quantile Regression Forests," Journal of Machine Learning Research, 7 983-999.

MIT Lincoln Laboratory, 2012: "Tower Flight Data Manager Benefits Assessment: Initial Investment Decision Final Report", Project Report ATC-394

NASA Traffic Management Advisor, 2011. Available from http://www.aviationsystemsdivision.arc.nasa.gov/research/foundations/tma.shtml. Last accessed on July 18, 2011.

U.S. Bureau of Transportation Statistics, 2011. Available from www.transtats.bts.gov. Last accessed on July 18 2011.