Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATM2017)

Predicting Performance of Ground Delay Programs

Alexander Estes
Applied Mathematics and Scientific Computation, Institute for Systems Research
University of Maryland, College Park, MD, USA
aestes@math.umd.edu

Michael O. Ball
R.H. Smith School of Business and Institute for Systems Research
University of Maryland, College Park, MD, USA
mball@umd.edu

David J. Lovell
Department of Civil and Environmental Engineering, Institute for Systems Research
University of Maryland, College Park, MD, USA
lovell@umd.edu

***Abstract*— Models are proposed to estimate the performance of Ground Delay Programs as air traffic management initiatives. We apply Random Forest and Gradient-Boosted Forest regression techniques within the context of Geographically Weighted Regression. We estimate both the mean and 90th percentile responses for two performance indicators: average arrival delay and the number of cancelled arrivals.**

Keywords-ground delay program; delay prediction; air traffic management

I. INTRODUCTION

When severe weather or other circumstances affect the ability of an airport to accept arriving flights then the scheduled arrivals can exceed the capacity of the airport. In order to prevent an unsafe situation from occurring, the Federal Aviation Administration will issue a Ground Delay Program (GDP). This issues delays to flights that are still on the ground, so that these flights may arrive at a time when there is sufficient capacity.

There is a large body of existing research on planning GDPs, optimizing them, determining appropriate air carrier responses, and other issues related to this particular type of traffic management initiative. A useful piece of that puzzle, that this paper aims to address, is being able to predict the performance of a particular candidate GDP, given a set of traffic and weather conditions. A number of performance metrics could be interesting, and here we focus on delays and flight cancellations.

The main purpose of producing these estimates is to help decision-makers plan GDPs. The work described here is part of a larger effort to build a decision support tool for planning traffic management initiatives. While the work will be described in that context, it could also be applied as a module for different situations where GDP performance predictions would be useful.

The tool we are pursuing will identify historical days with conditions similar to the current day, and then produce a set of GDPs that would be representative of the variety of GDPs that had previously been deployed on those historical days. The work in this paper is intended to be used to estimate the performance that the representative GDPs would experience if re-applied on the current day. This would allow the decision-maker to evaluate which historical GDP had the best prospects for good performance. There are other ways this could be used to aid in the planning or evaluation of GDPs. For example, these methods could be used in a what-if analysis tool where GDP decision-makers could propose GDPs (rather than mining them from the historical record) and receive estimates of the predicted performance.

There is a substantial body of work that seeks to predict delays in the national airspace system. Much of this work makes its predictions primarily based on weather and traffic, and makes little or no use of GDP features. See for example [1], [2], [3], [4], [5], or [6]. This work would have limited direct use in the prediction of GDP performance. There are also some existing queue-based or network-flow-based methods that estimate the amount of delay that a GDP would produce, such as [7], [8], and [9].

If we were solely interested in delay as a measure of GDP performance, then these methods would perhaps be sufficient. However, recent research in air traffic flow management has advanced the idea that GDPs and other traffic management initiatives are multi-objective problems. Five performance criteria measuring different aspects of GDPs were provided by [10] and a model for GDP planning that includes two metrics of performance was examined in [11]. A mechanism was proposed in [12] and [13] that considers three performance metrics. This mechanism would allow flight operators to express their preferences for combinations of these metrics, and would then produce a GDP plan.

Our proposed methods are black-box methods. That is, the methods do not involve explicitly modelling or simulating elements of the national airspace system. Instead, our methods build estimates of performance purely using historical data of the GDPs that were taken, the conditions that these actions were taken in, and the resulting performance metrics. Thus, the method can be used to estimate any metric with no modification while the aforementioned queue-like models would require substantial modification. As far as we are aware, our methods are the first black-box methods that have been proposed for this purpose. Thus, our methods would be more readily applied to the implementation of a mechanism such as in [12] and [13], and could generally be more informative to decision-makers who are interested in objectives other than delay. There is a further advantage to our models in that they make use of observed performance outcomes from GDPs and are able to learn from these

outcomes, while the queue-like or network-flows-like models do not.

Our paper proceeds as follows. In section II, we describe our proposed methods for estimating GDP performance measures and in section III we present computational results that evaluate the quality of these methods. Conclusions and possible avenues of future research are discussed in section IV.

II. METHODOLOGY

Our methodology was inspired by local spatial regression techniques. In spatial regression, the goal is the same as in a standard regression problem, that is, to predict one or several dependent variables using some explanatory variables. However, in spatial regression the observations of the dependent variable and explanatory variables take place at some geographic locations, and the relationship between the variables may vary among locations. In particular, our method uses a weighting scheme similar to one that has been used in Geographically Weighted Regression (GWR) ([14] and [15]). In this section, we first provide a brief description of GWR, and then we describe how we altered this approach to make it suitable for predicting the performance of GDPs.

A. Geographically Weighted Regression

In GWR, a separate regression model is fit for each geographic location of interest. Given observations of the explanatory variables in some location, a prediction of the dependent variable is produced by applying the regression model for that location. The regression model for a particular location is constructed under the assumption that observations that occurred at nearby locations are more relevant than observations that occurred at more distant locations. This is achieved by the following weighting scheme. Let there be some known distance measure d so that $d(l_1, l_2)$ gives the distances between locations l_1 and l_2. Then, a function k, known as the kernel, is chosen. The kernel takes a distance as an input and returns a weighting between zero and one, where higher distances are given a lower weight and smaller distances are given a higher weight. A typical choice for this function would be a Gaussian kernel,

$$k(d;\beta) = \exp\left(\frac{-d^2}{\beta}\right). \quad (1)$$

Here, β is a tunable parameter, which is known as the bandwidth. The kernel is used to assign a weight to each observation. Let l^p be the location where the prediction will be made, and let l_i be the location of the i^{th} observation. Then the weight assigned to the i^{th} observation is given by

$$k\left(d\left(l^p, l_i\right);\beta\right). \quad (2)$$

In GWR, regression models are built by using these weights in weighted least-squares linear regression, but any regression technique that admits the use of sample weights can be converted into a spatial regression technique by applying this weighting scheme.

In order to implement this method, the bandwidth parameter must be given a value. If a low value is chosen, then nearby observations are given much higher weight than further observations. As the bandwidth increases, then the assigned weights become more even, with all weights approaching one as the bandwidth approaches infinity. Thus, smaller bandwidth values lead to more localized regression models, while higher values lead to more global regression models.

B. Weighting Observations for GDP Performance Prediction

We approached the problem of predicting the performance of a planned GDP similarly to a spatial regression problem. Our explanatory variables describe the planned GDP and our independent variables are measures of the performance of the GDP, such as cancellations and average arrival delay over the day. Our observations do not vary in their geographic origin. All of the GDPs are assumed to have been run at one airport, and the observed performance measures are specific to that airport. However, each observed GDP was implemented in response to some weather and traffic conditions. These conditions could include the weather present at the time the GDP was being planned, as well as the weather forecasts, the traffic present in the system, and the forecast demand at that time. The weather and traffic conditions present when the GDP was being planned serve the same role that the geographic location serves in a typical spatial regression problem. That is, the weather and traffic conditions affect the relationship between GDP parameters and the performance of the GDP, and conditions that are "closer" to those associated with the subject GDP have more impact than those that are "farther away". In order to provide predicted performance of different GDP parameters in some weather and traffic conditions, we fit a regression model for those specific weather and traffic conditions.

We applied the same weighting scheme as in GWR, so historical observations that took place under similar weather and traffic conditions received higher weight and those that took place under dissimilar weather and traffic conditions received lower weight. This required a measure of distance between sets of weather and traffic conditions. We used a distance measure that is under development as part of an ongoing project ([16] and [17]). Demand and terminal weather features were used to estimate the distribution of capacity at the airport. Then, the distance measure between each pair of days was produced by comparing the estimated capacity distributions for those days. Weights for the observations were produced using a Gaussian kernel as described in Section II.A. The demand and weather features used in the construction of this distance measure are real observations that were recorded on the corresponding day. These features would not be available for the current day. In practice, the distances used by this method should be produced by comparing forecast weather and traffic conditions on the current day with a combination of forecast and actual conditions in previous days. We leave the problem of producing such a distance measure as an avenue of future research.

A natural alternative approach would be to build a regression model that includes variables describing the GDP as well as variables describing the weather and traffic conditions. There are advantages and disadvantages to either approach. The largest disadvantage to our approach is that a model must be fit for each set of weather and traffic

conditions. This means that it will be much more computationally expensive to use our approach if performance estimates are desired in a large variety of weather and traffic conditions. On the other hand, if the goal is to produce performance estimates for many different GDP choices under a single set of weather and traffic conditions, then our method will not cause any additional computational burden. In the global approach, the objective is to minimize a total loss function over all observations. This can result in a model that fits some locations better than others, especially if the relationship between GDP variables and the performance changes greatly in different weather and traffic conditions. In our approach, we form a regression model specifically for the location of interest, so this is not as much of a problem. Indeed, existing research suggests that GWR produces residuals that are more even in magnitude across geographic locations as compared to several other regression techniques, including ordinary least squares and two types of neural network models [18].

C. Loss Functions in Regression and Quantile Regression

For some observed dependent variable y and some prediction $\hat{y}$ a *loss function* is a function that takes a predicted value and an actual value as inputs and whose output is a number representing a penalty for misestimating the value. Typical loss functions include squared error $(\hat{y}-y)^2$ or absolute error $|\hat{y}-y|$. These functions are used to evaluate the quality of the predictions made by a regression model.

We chose to use absolute error as the loss function for estimates of GDP performance because this leads to models that are more robust to outliers, and the loss function is in the same scale as the original data. Appropriate choice of loss function can also provide a way to estimate quantiles of the dependent variable. It is known that the loss function f defined by

$$f(y,\hat{y}) = \begin{cases} (\alpha-1)(y-\hat{y}) & \text{if } y < \hat{y} \\ \alpha(y-\hat{y}) & \text{if } y > \hat{y} \end{cases} \tag{3}$$

is minimized when $\hat{y}$ is equal to the α-quantile of y. Thus, a regression method that attempts to minimize this loss function will provide an estimate of the α-quantile of the dependent variable. This is the basis of quantile linear regression [19], and any regression that admits a flexible loss function can make use of this observation to provide quantile estimates.

D. Forest Methods with a Spatial Weighting Scheme

We propose two methods that involve applying the weighting scheme described in section II.A to an existing regression method. The regression methods that we used are Random Forest and Gradient-Boosted Forest, both of which provide predictions using a collection of decision trees. Decision trees are rooted binary trees, so each node has either two children or no children. A node with no children is called a leaf node. Each non-leaf node has an associated rule, which can be expressed in the form '$x < a$' where x is an explanatory variable and a is a constant. Each leaf node has an associated predicted value for the dependent variable. For an observation of the explanatory variables, a prediction is produced in the following manner. At each node, if the observation satisfies the rule then the next node to be examined will be the left child. Otherwise, the next node will be the right child. This continues until a leaf node is reached, at which point the prediction is given by the value associated with that node. Decision trees are usually constructed in an iterative procedure, which begins with a single root node with no children. In each iteration, one leaf node is chosen and is 'split', which means that it is given two children and an associated rule. This proceeds until a stopping condition is met, at which time a 'pruning' procedure may be employed to remove some branches from the tree to simplify the tree and prevent overfitting. One well-known such procedure is described in [20].

The Random Forest regression method was introduced in [21], and is a method that builds a forest of decision trees. Each decision tree is built independently of all other trees, and randomness is introduced into the fitting procedure in order to decrease the correlation between the predictions of the trees in the forest. This in turn reduces the variance of the predictions. Random Forest techniques already allow the use of weighting of sample observations. This has previously been used to improve the performance of the method for classification problems in which some classes of data appear much more often than other classes [22]. The weights of the observations are incorporated in the splitting criteria in the formation of the tree and in the resulting predictions once the forests have been formed.

The Gradient-Boosted Forest method builds a collection of decision trees iteratively. Each new tree is fit in a way that attempts to correct the errors of the previous tree. This is accomplished by fitting the new tree to the gradient of the loss function. This improvement procedure, known as gradient boosting and introduced in [23], is analogous to the gradient descent method in continuous optimization. In a similar manner to the Random Forest method, weights on sample observations can be added to Gradient-Boosted Forests by including these weights in the splitting criteria for each tree, and in the predictions produced by each tree. In Gradient-Boosted Forest methods, the weights can also be incorporated directly into the loss function. Since Gradient-Boosted Forests support any loss function, this method can be applied to the problem of estimating quantiles with no modifications. The Random Forest method cannot be applied to arbitrary loss functions without developing new methods for fitting the trees, so we omitted the Random Forest method when we estimated quantiles. For both methods, we used the implementation in the Python `scikit-learn` package. There are many resources that explain the Random Forest and Gradient-Boosted Forest methods in more detail. See for example [24], [25] or [26].

E. Baseline Methods

We used two additional methods to establish a baseline that we could compare to our proposed methods. The first baseline method estimates the performance of a GDP in some weather and traffic conditions by taking a weighted average of all the historical observations. Observations that were similar to the GDP and that occurred in similar weather and traffic

conditions received higher weights than observations that were dissimilar in either aspect.

We used a measure of distance that is similar to the one that we used for the spatial Random Forest and Gradient-Boosted Tree models, but that considers GDP features along with the traffic and weather conditions. Again, we used a Gaussian kernel to convert these distances into weights. The resulting prediction is given by taking the weighted average of the observations $\sum_{i=1}^{n} w_i y_i$.

Our second baseline method was a *k*-nearest neighbors method. For a GDP in some weather and traffic conditions, the prediction for this method is the average of the *k* closest observations. We used the same distance measure here that was used for the weighted average method.

These baseline methods required some slight modifications to produce estimates of quantiles rather than estimates of expected values. In this case, we took a weighted quantile instead of taking a weighted average. For the quantile α, this would be

$$\min\left\{q : \alpha \le \sum_{i:y_i \le q} w_i\right\} . \quad (4)$$

For a large quantile α, then the maximum of the *k* nearest neighbors can be used as a baseline rather than the average of the *k* nearest neighbors. Alternatively, if one were interested in estimating a low quantile α, the minimum of the *k* nearest neighbors could serve as a baseline. We were more interested in worst-case performance than best-case performance, so we only implemented the maximum *k*-nearest neighbors method.

F. Choice of Explanatory Variables

The relevant variables chosen in the planning of a GDP have been discussed in [27] and [28]. Our choice of explanatory variables was mostly consistent with these previous works, although the exact variables we used are slightly different because those authors made stronger simplifying assumptions than we did. We used three variables that determine when the GDP takes place:

- *Entry Time*: the time that the GDP was declared and put into effect. This was measured in minutes after 4:00 a.m. local time.
- *Earliest ETA*: the earliest arrival time for flights that were controlled by the GDP, measured in minutes after 4:00 a.m. local time.
- *Duration*: the difference in time between the earliest and latest arrival times that were controlled by the GDP, measured in minutes.

There were several variables that determine how severe the GDP is:

- *AAR in time period t*: For each fifteen minute time period *t* starting at 4:00 a.m. local time, we recorded the planned AAR if this time period was included within the time period in which the GDP was planned to be in effect. If not, this field was left undefined. All of the regression methods that we considered can handle fields with missing values, so there was no need to define this field in time periods when the GDP was not active.
- *Average AAR*: the average of the AARs in the time interval that the GDP was defined.

The scope, or geographic area controlled by the GDP, can either be described by a set of pre-defined geographic regions or by a radius surrounding the affected airport. A scope defined by a certain set of pre-defined geographic regions would be similar to some choice of radius. Ideally, the variables should be chosen so that our regression models would treat these scopes similarly. With this intent, we defined a variable,

- *Number of Core 30 airports*: the number of core 30 airports that fell within the scope of the GDP.

The Core 30 airports are a set of 30 airports that the Federal Aviation Administration has identified as those having the largest volume of traffic. This variable gives a measure of the magnitude of the scope of the GDP that is valid for either manner of representing the scope. Occasionally, there may be a ground stop immediately preceding the implementation of a ground delay program. In these cases, we included a feature:

- *Ground Stop Duration:* the duration of the ground stop that led into the GDP, in minutes. If there was no such ground stop, then we defined this field to be zero.

All of these explanatory variables were taken from National Traffic Management Log (NTML) data. We considered two dependent variables. The first was the average gate delay of arriving and departing flights, which was taken from the FAA's Aviation System Performance Metrics (ASPM) database. The second was the number of flights that were scheduled to arrive but were cancelled, which was taken from the NASA/FAA Performance Data Analysis and Reporting System (PDARS) database. We estimated the expected value of these variables, which is the usual goal in regression. However, there is generally uncertainty in weather and traffic conditions, and there are some variables that we have not included such as the response taken by the airlines. We expect that if the same GDP were to be implemented multiple times in similar traffic and weather conditions then the resulting performance may vary significantly. For this reason, it would also be useful to provide some estimate of the range of performance that may be expected from the GDP. An upper bound for the delay and cancellations would be more useful than a lower bound, as decision-makers generally would like to avoid GDPs that could result in very poor performance. Therefore, we estimated the 90% quantiles of these variables.

G. Parameter Tuning Procedure

The Random Forest and Gradient-Boosted Forest methods each have parameters that require tuning. For the Random Forest method, the only parameter that we tuned was the number of trees in the forest and the remaining parameters we left to the defaults as implemented in the Python `scikit-learn` package. For the Gradient-Boosted Forest methods, we tuned the number of trees in the forest, the maximum depth

of the tree, and the learning rate. Our proposed spatial variants of the Random Forest and Gradient-Boosted Forest also require a choice of bandwidth, as described in section II.A. The bandwidth must also be chosen for the weighted average baseline method. The k-nearest neighbors method requires a choice for the value of k, i.e, the number of neighbors used to create the prediction.

Our criteria for choosing these parameters was the average leave-one-out loss. That is, we fit the model to the data set with one observation (x_i, y_i) excluded. We used the resulting model to provide a prediction $\hat{y}_i$ for the dependent variable y_i given the observation x_i of the explanatory variables. Then we calculated the value of the loss function for the prediction and the actual value. This procedure was repeated for each observation in the training set, and the loss was averaged. We considered one choice of parameters to be better than another if the former produced a lower average leave-one-out loss.

In order to reduce the computational burden, we first tuned the number of trees in a Random Forest model that attempts to predict performance measures from the GDP features without using information about weather or traffic states. We will refer to this model as the global Random Forest model. We assumed that a choice of parameters that works well for this model would also work well for the spatial Random Forest model. Similarly, we tuned the parameters of a global Gradient-Boosted Forest model and assumed that this choice of parameters would also work well for the spatial Gradient-Boosted Forest model. A Random Forest model does not tend to overfit as the number of trees increases. As the number of trees increases the average leave-one-out loss tends to decrease, but the marginal benefit of adding another tree becomes increasingly small. We therefore plotted the average leave-one-out loss against the number of trees, and chose a number of trees where the plot appears to become flat. For the Gradient-Boosted Forest, we performed a grid search to choose the best number of parameters. We allowed five different values of the learning rate, specifically 0.2, 0.1, 0.05, 0.01 and 0.005, we allowed maximum tree depth to take each value between 2 and 7 and we allowed the number of trees to take any value between 1 and 300. The average leave-one-out loss was calculated for each choice of these parameters and we kept the set of parameters with the lowest average leave-one-out loss.

TABLE 1. RESULTS FROM ESTIMATION OF EXPECTED VALUE OF AVERAGE DELAY.

Method	Avg. Error	Improvement Over Unweighted Avg.
Unweighted Average	16.982	0.0%
Weighted Average	12.865	-24.2%
Average of *k*-Nearest Neighbors	14.139	-16.7%
Global Random Forest	11.759	-30.8%
Spatial Random Forest	11.612	-31.6%
Global Gradient-Boosted Forest	12.471	-26.6%
Spatial Gradient-Boosted Forest	12.381	-27.1%

After the parameters were tuned for the global models, we tuned the bandwidth for the spatial model. We calculated the leave-one-out loss for each bandwidth between 0.3 and 10 in increments of 0.05, and we selected the bandwidth that produces the lowest average leave-one-out loss. The bandwidth for the weighted average model was chosen in the same manner. For the *k*-nearest neighbors model, we computed the average leave-one-out loss for each value of k between 1 and 100, and we took the choice k that produced the smallest value.

III. COMPUTATIONAL RESULTS

Our dataset consisted of 369 days on which GDPs were planned at Newark Liberty International Airport. These days occurred between April 2011 and October 2014. We randomly selected 80% of the days (296 days) to form a training set, while the remaining days (73 days) were used for the test set. We estimated the expected value and 90% quantiles of average gate delay across the day, and the same metrics of the number of cancellations. In each case, the parameters for the models were chosen using the procedure described in section II.F. For each model, we used the training set observations to produce predictions for the test set and we calculated the average loss for the test set. For the expected values of the dependent variables, we used the average absolute error, while for the 90% quantiles we used the quantile loss function as described in section II.C. In addition to the proposed models and baseline models described in section II.D, we show the performance of the global Random Forest model and global Gradient-Boosted Forest model, which do not make use of weather or traffic information. We also show the loss that resulted from using the unweighted average of the dependent variable as the estimate for that variable regardless of the choice of explanatory features. The results for the estimation of the expected value of average delay are shown in Table 1.

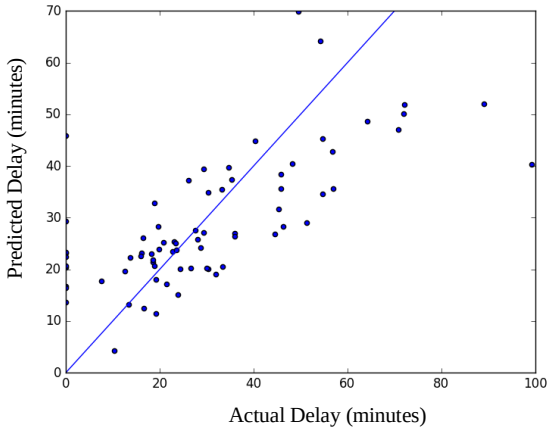

Figure 1. Actual delays vs. predicted delays from spatial Random Forest.

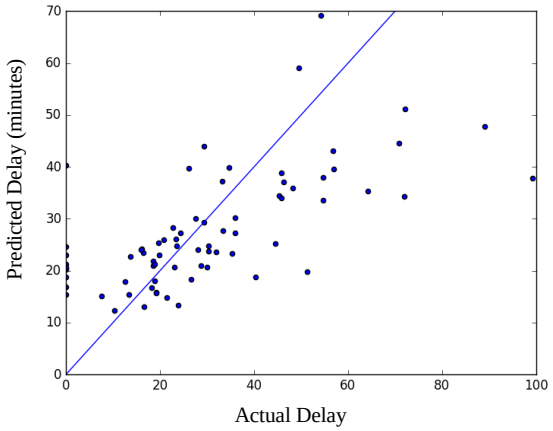

Figure 2. Actual delays vs. predicted delays from spatial Gradient Boosted Forest.

The methods that performed the best were the two variants of the Random Forest method, followed by the two Gradient Forest methods. The plots of predicted versus actual values for the spatial Random Forest and spatial Gradient-Boosted Forest methods are shown in Fig. 1. and Fig. 2., respectively. From these plots, it seems that the spatial Gradient-Boosted Forest was more accurate when the average delay was less than 25 minutes but the spatial Random Forest model was more accurate for higher levels of delay. In general, the variant of each method that included weather and traffic information outperformed the variants that did not. However, the spatial Random Forest and spatial Gradient-Boosted Forest methods only slightly outperformed their global variants. This may indicate that the relationship between GDP parameters and delays does not change greatly in different weather and traffic conditions.

The results from the estimation of the 90% quantile of average delay are shown in Table 2. Both variants of the Gradient-Boosted Forest method again outperformed the baseline methods, and the spatial model had almost exactly the same performance as the global model. The improvement of the Gradient-Boosted Forest model over the baseline models was much more pronounced in the estimation of the 90% quantile, as the best baseline method produced nearly twice the average loss than the Gradient-Boosted Forest models did.

TABLE 2. RESULTS FROM ESTIMATION OF 90% QUANTILE OF AVERAGE DELAY

Method	Average Loss	Improvement Over Unweighted Average
Unweighted Quantile	8.589	0.0%
Weighted Quantile	7.044	-18.0%
Maximum of *k*-Nearest Neighbors	6.255	-27.2%
Global Gradient-Boosted Forest	3.356	-60.9%
Spatial Gradient-Boosted Forest	3.391	-60.5%

The results for the estimation of the expected value of the cancelled arrivals are shown in Table 3. This time, the spatial Gradient-Boosted Forest method had the best performance, while the spatial Random Forest had the second-best performance. In contrast to the previous results, the spatial methods showed significant gains as compared to their respective global versions.

TABLE 3. RESULTS FROM ESTIMATION OF EXPECTED VALUE OF CANCELLED ARRIVALS.

Method	Avg. Error	Improvement Over Unweighted Avg.
Unweighted Average	16.381	0.0%
Weighted Average	10.511	-35.8%
Average of *k*-Nearest Neighbors	13.297	-18.8%
Global Random Forest	10.924	-33.3%
Spatial Random Forest	9.310	-43.2%
Global Gradient-Boosted Forest	10.437	-36.3%
Spatial Gradient-Boosted Forest	8.443	-48.5%

Table 4 shows results from estimating the 90% quantile of the cancelled arrivals. Similarly to the results for average delay, the Gradient-Boosted Forest methods greatly outperformed the baseline models, and the improvement is much greater than in the estimation of expected cancelled arrivals. The global method performed slightly better than the spatial method. This could indicate that the spatial model is over-fitting, although the difference is slight enough that it could simply be sampling error.

TABLE 4. RESULTS FROM ESTIMATION OF 90% QUANTILE OF CANCELLED ARRIVALS

Method	Avg. Error	Improvement Over Unweighted Avg.
Unweighted Quantile	9.164	0.0%
Weighted Quantile	7.800	-14.9%
Maximum of *k*-Nearest Neighbors	6.892	-24.7%
Global Gradient-Boosted Forest	3.488	-61.9%
Spatial Gradient-Boosted Forest	3.697	-60.0%

Overall, the spatial Gradient-Boosted Forest and spatial Random Forest methods both seem to be viable models for estimating expected values of the dependent variables, outperforming all of the baseline models in all tests. For quantile estimates, the Gradient-Boosted Forest models also performed much better than the baseline methods, although the spatial model performed very slightly worse than the global model.

IV. CONCLUSIONS

We presented methods for predicting how well a GDP will perform in some given weather and traffic conditions. Our methods are based on spatial regression techniques, where a regression model is fit for each set of weather and traffic conditions. The explanatory variables consisted of GDP features, while our dependent variables were the average delay and the number of cancelled arrivals. These models were

compared against each other and against baseline methods. Both of our proposed methods outperformed all of the baseline methods. The spatial Random Forest method performed better than the Gradient Boosted method when predicting the expected value of average delay, while the reverse was true when predicting the expected number of cancelled arrivals.

We observed from our tests that the methods that used information about the weather and traffic did not seem to predict the expected value of delay much better than the methods that did not use this information. Furthermore, in the estimation of quantiles of the dependent variables, the proposed methods that used the weather and traffic information performed very slightly worse than the variants that did not use this information. This could indicate that these values depend highly on the GDP and not as much as on the weather and traffic. This would have implications for the planning of GDPs. On the other hand, this could indicate that there is room for improvement in our methods. Either way, it would be of value to determine exactly why this occurs.

In our tests, we predicted average delays and cancelled arrivals separately. In practice, there is likely a relationship between these dependent variables. Cancelled flights do not contribute to the delay measure, and in fact a reasonable expectation of high delay can be a good reason to cancel a flight. Thus, we expect that higher cancellations would coincide with lower average delays. Further work could produce methods that estimate trade-offs between several measures of delay performance. This could either be presented to decision-makers to help them make decisions directly, or this could be included in a GDP-planning procedure such as the one described in [12] and [13].

We only estimated the performance of the initial GDP plan. In practice, it is sometimes beneficial to revise a GDP after traffic and weather has developed. Our methods could be extended to estimate the performance of proposed revisions. As with the initial GDP plan, this could be presented to decision-makers to help them plan revisions, or could be incorporated into an automated tool for GDP planning.

ACKNOWLEDGMENT

The data for this study were processed and graciously provided to us by Sreeta Gorripaty, Yulin Liu, Mark Hansen, and Alexei Pozdnukhov of the University of California at Berkeley and Kennis Chan and John Schade of ATAC, Inc. This work was funded under NASA grant no. NNX14AJ79A.

REFERENCES

[1] G. Chatterji and B. Sridhar, "National airspace system delay estimation using weather weighted traffic counts," AIAA Guidance, Navigation, and Control Conference and Exhibit, August 2005.

[2] D. A. Smith and L. Sherry, "Decision support tool for predicting aircraft arrival rates, ground delay programs, and airport delays from weather forecasts," Proc. of the International Conference for Research in Air Transportation, 2008.

[3] N. Xu, L. Sherry, and K. Laskey, "Multifactor model for predicting delays at US airports," Transportation Research Record, no. 2052, pp. 62–71, 2008.

[4] A. Klein, "Airport delay prediction using weather-impacted traffic index (WITI) model," Digital Avionics System Conference, pp. 2.B.1-1–2.B.1-13, October 2010.

[5] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," Transportation Research Part C: Emerging Technologies, vol. 44, pp. 231–241, July 2014.

[6] S. Choi, Y. Kim, S. Briceno, and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," Digital Avionics System Conference, pp. 1–6, September 2016.

[7] Y. Wan and S. Roy, "A scalable methodology for evaluating and designing coordinated air traffic flow management strategies under uncertainty," IEEE Transactions on Intelligent Transportation Systems, vol. 9, no. 4, pp. 644–656, December 2008.

[8] C. Wanke and C. Taylor, "Exploring design trade-offs for strategic flow planning," Aviation Technology, Integration, and Operations Conference, August 2013.

[9] J. Rebollo and C. Brinton, "Brownian motion delay model for the integration of multiple traffic management initiatives," USA/Europe Air Traffic Management Research and Development Seminar, June 2015.

[10] Y. Liu and M. Hansen, "Evaluation of the performance of ground delay programs," Transportation Research Record: Journal of the Transportation Research Board, no. 2400, pp. 54–64, 2013.

[11] Y. Liu and M. Hansen, "Incorporating predictability into cost optimization for ground delay programs," Transportation Science, vol. 50, no. 1, pp. 132–149, November 2016.

[12] P. Swaroop and M. Ball, "Consensus building mechanism for setting service expectations in air traffic management", Transportation Research Record: Journal of the Transportation Research Board, no. 2325, pp. 87–96, 2012.

[13] M. Ball, C. Barnhart, M. Hansen, L. Kang, Y. Liu, P. Swaroop, V. Vaze, and C. Yan, "Distributed mechanisms for determining NAS-wide service level expectations: final report," delivered to FAA, October, 2014.

[14] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically weighted regression," Journal of the Royal Statistical Society: Series D (The Statistician), vol. 47, no. 3, pp. 431–443, 1998.

[15] J. Clerk Maxwell, Geographically Weighted Regression. John Wiley & Sons: Chichester, 2002.

[16] S. Gorripaty, A. Pozdnukhov, M. Hansen, and Y. Liu, "Identifying similar days for air traffic management," World Conference on Transport Research, July 2016.

[17] M. Hansen, "Similar historical days and air traffic management response strategies – progress report: August 1, 2015 – July 31, 2016," delivered to NASA Ames Research Center, July 2016.

[18] L. Zhang, J. H. Gove, and L. S. Heath, "Spatial residual analysis of six modeling techniques," Ecological Modelling, vol. 186, no. 2, pp. 154–177, August 2005.

[19] R. Koenker and G. Bassett Jr., "Regression quantiles," Econometrica, vol. 46, no. 1, pp. 33–50, January 1978.

[20] L. Breiman, J. Friedman, R. A. Olshen and C. Stone, Classification and Regression Trees. Wadsworth & Brooks: Monterey, 1984.

[21] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, October 2001.

[22] C. Chen, A. Liaw, and L. Breiman, "Using random forests to learn imbalanced data," Technical Report 666, U. C. Berkeley, July 2004.

[23] J. Friedman, "Stochastic gradient boosting," Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367–378, February 2002

[24] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, 2009.

[25] J. Elith, J. R. Leathwick,and T. Hastie. "A working guide to boosted regression trees," Journal of Animal Ecology, vol. 77, no. 4, pp. 802–813, April 2008.

[26] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Bacon Raton, FL: University Science, 2012.

[27] M. O. Ball and G. Lulli, "Ground delay programs: optimizing over the included flight set based on distance," Air Traffic Control Quarterly, vol. 12, no. 1, pp. 1–25, January 2004.

[28] L. Cook and B. Wood, "A model for determining ground delay program parameters using a probabilistic forecast of stratus clearing," Air Traffic Control Quarterly, vol. 18, no. 1, pp. 85–108, January 2010.

AUTHOR BIOGRAPHIES

Alexander Estes is a Ph.D. candidate in the Applied Mathematics & Statistics, and Scientific Computation program at the University of Maryland. He holds a B.S. in Mathematics from the University of Nebraska, Lincoln.

Michael Ball holds the Dean's Chair in Management Science in the Robert H. Smith School of Business at the University of Maryland. He also has a joint appointment within the Institute for Systems Research in the Clark School of Engineering, and is co-Director of NEXTOR-II, an FAA consortium in aviation operations research. Dr. Ball received his Ph.D. in Operations Research in 1977 from Cornell University.

David Lovell is a Professor of Civil and Environmental Engineering at the University of Maryland. He holds a joint appointment with the Institute for Systems Research. Dr. Lovell received his Ph.D. in Civil Engineering in 1997 from the University of California, Berkeley.