Predicting Performance of Ground Delay Programs

Alexander Estes, David Lovell and Michael Ball
University of Maryland – College Park

Research funded by NASA grant no. NNX14AJ79A
Ground Delay Programs (GDPs)

• Scheduled traffic can exceed capacity of airport to handle arrivals.
• To prevent unsafe situations, FAA will delay flights on the ground.
Predicting GDP Performance

• Goal: Given some weather and traffic conditions, and given a GDP, predict the expected performance of the GDP.

• Purpose: Help decision-makers plan GDPs.
  • Can be combined with decision support tools.
  • Can be used with post-operations analysis.
  • What-if analysis.
  • Developed for Similar Days project.
Existing Work

• A large body of work involves predicting delays, but much of this work makes little or no use of GDP (or other TMI) features.

• Existing methods for predicting GDP performance focus on delays and are largely queue-like or network-flow-like models.

• Recent research uses multiple criteria to measure GDP performance
  • Liu and Hansen 2010, Swaroop and Ball 2012, Ball et al. 2014, Liu and Hansen 2016
  • This suggests a need for more flexible models.

• We propose two methods that can be used to produce estimates of arbitrary performance measures with no modifications.
Proposed Methods

• The performance measures that we predict:
  • Delay averaged across all arriving flights on the given day.
  • Total number of departures that were cancelled on the given day.

• We can also produce estimates of quantiles of these measures.
  • E.g.: an estimate of the 90% quantile of the distribution of average arrival delays that could result from running a TMI on a given day.
  • We expect this to be useful because there are many sources of uncertainty that are difficult to account for.
  • Estimates come directly from model.
Overview

• Methodology – Estimating expected values
  • Geographically Weighted Regression
  • Forest-Based machine-learning methods
  • Parameter Tuning

• Computational Results

• Methodology - Estimating Quantiles

• More Computational Results

• Similar Days Project

• Conclusions
Proposed Methods

• We propose two methods.

• Both of these methods use a weighting scheme similar to that used in Geographically Weighted Regression (Brunsdon, Fotheringham and Charlton 1998).

• This weighting scheme is combined with two existing machine learning methods.
Geographically-Weighted Regression

• Typical setting:
  • Explanatory variables, e.g. population density, unemployment levels.
  • Dependent variable, e.g. levels of illness.
  • Use explanatory variables to predict dependent variable.
  • Relationship between variables depends on geographic location.

Coefficient of population density. Figure from Brunsdon, Fotheringham and Charlton 1998
Geographically-Weighted Regression (2)

• How GWR works:
  • A separate linear regression model is fit for each location where a prediction is desired.
  • Observations are weighted by their distance from the location.
Geographically Weighted Regression (3)

• Weights are generated by a kernel that transforms distances into similarities.

• A typical choice of kernel is a Gaussian kernel:

\[ K(d; \beta) = \exp\left(\frac{-d^2}{\beta}\right) \]
Geographically Weighted Regression (4)

- The parameter $\beta$ is called the bandwidth, and is tunable.
- Low values of $\beta$ assign very high weight to the closest values.
- Higher values of $\beta$ assign weights more evenly.

\[ K(d; \beta) = \exp\left(\frac{-d^2}{\beta}\right) \]
Application to GDP Performance Prediction

• Weather/traffic conditions serve the role that geographic location serves in GWR.

• We fit a new model for each weather/traffic condition.

• The explanatory variables are GDP parameters.
  • Our choice of parameters will be discussed later.

• The dependent variable is a measure of performance.
  • In our case, arrival delay or cancelled departures

• This requires a measure of distance between weather/traffic conditions (to be discussed later).
Decision Trees

- Rooted Binary tree.
- Each non-leaf node has an associated feature and an associated threshold.
- Each leaf node has a prediction.

- **Duration > 4 hr.**
  - **Start > 1500Z**
    - No: Est. Delay: 20 min
    - Yes: Est. Delay: 40 min
  - **Duration > 8 hr.**
    - No: Est. Delay: 30 min
    - Yes: Est. Delay: 50 min
Decision Trees (2)

- To get a prediction, traverse the tree.
- E.g. Duration= 3 hrs, Start = 1600Z.
- See Breiman et al. 1984 for details on constructing these trees.
Random Forests

• Random forests build `forests’ of decision trees randomly.
• Each tree is built independently.
• Randomness is introduced into fitting procedure to reduce correlation between trees of forest.
Gradient-Boosted Forests

• Builds a `forest’ of decision trees iteratively.
• In each stage, a new tree is fit to residuals from the previous iteration.
• Predictions from new tree are multiplied by a learning rate and added to predictions.
• This is analogous to gradient descent in continuous optimization.
Forest-based Methods

• For more details on these methods, see:
  • Breiman 2001
  • Friedman 2002
  • Hastie, Tibshirani and Friedman 2009
  • Elith, Leathwick and Hastie 2008
  • Zhou 2012
Incorporating GWR Weights

• Weights can be incorporated into these methods.

• Applying the methods to weighted data is equivalent to applying to an unweighted data set with duplicated observations, where the amount of duplications are proportional to the weights.
Parameter Tuning Procedure

• These methods have parameters that need tuning (e.g. number of trees)
• We also need to tune the bandwidth for GWR weights.
• In order to reduce computational burden, we tune the parameters in two stages:
  1. Tune all parameters other than the bandwidth, using unweighted data
  2. Tune the bandwidth used for the weights
• Use standard K-Fold cross-validation for first step.
• Use leave-one-out cross-validation for second step.
Weather/Traffic Distance Measure

• Need a distance measure between weather/traffic conditions.
• We use an existing distance measure (Gorripaty, Hansen, and Pozdnukhov 2016).
• Weather/traffic is used estimate distributions of capacity at airport.
• Distance is produced by comparing estimated distributions.

Estimate CDFs of capacity using weather/traffic
Compare CDFs to measure distance
Comparison Methods

• Unweighted Median – predicted value is median of all observations of dependent variable, regardless of explanatory variables.

• Weighted Median – predicted value is weighted median of observations, using GWR weighting scheme with combined distance.

• K-nearest neighbors – predicted value is the average of the K-nearest observations, according to combined distance measure.

• Unweighted Random Forest – random forest model is used, without GWR weighting scheme.

• Unweighted Gradient-Boosting – gradient-boosted forests are used, without GWR weighting scheme.
GDP Features

• All features reflect first planned GDP. Revisions are not included.
• File time - minutes after 4:00 a.m. local
• Start Time - minutes after 4:00 a.m. local
• Duration - minutes
• Called Rate - in 15 min. intervals
• Average called rate
• GS Duration - duration of GS preceding GDP, 0 if GDP is not preceded by GS
GDP Features (2)

- **Scope** - number of CORE30 airports

Scope declared as radius

Scope declared by center

Data Sources

• Weather/Traffic distance and combined distances generated by Sreeta Gorripaty, Yulin Liu, Alexei Pozdnukhov and Mark Hansen (U.C. Berkeley).

• Average arrival delays and departure cancellations generated from Aggregate Demand List (ADL) data.

• GDP Features taken from National Traffic Management Log (NTML) Data.

• Thanks to collaborators at UC Berkeley and ATAC for their assistance in procuring and processing this data.
Test Procedure

• Data is divided into a training set and test set.
  • 80% of observations placed in training set.
  • 20% of observations placed into test set.

• For each observation in test set, generate the GWR weights (if applicable).

• Fit the model with training set.

• Use the fit model to predict the test set observation.

• Compare with actual value.
## Results – Average Arrival Delays

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. Test Error</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted Average</td>
<td>12.91</td>
<td>0.0%</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>10.00</td>
<td>-22.5%</td>
</tr>
<tr>
<td>K Nearest Neighbors</td>
<td>9.32</td>
<td>-27.8%</td>
</tr>
<tr>
<td>Unweighted Random Forest</td>
<td>16.49</td>
<td>27.7%</td>
</tr>
<tr>
<td>Unweighted Gradient Boosting</td>
<td>9.33</td>
<td>-27.7%</td>
</tr>
<tr>
<td><strong>Weighted Random Forest</strong></td>
<td><strong>8.66</strong></td>
<td><strong>-32.9%</strong></td>
</tr>
<tr>
<td><strong>Weighted Gradient Boosting</strong></td>
<td><strong>8.62</strong></td>
<td><strong>-33.2%</strong></td>
</tr>
</tbody>
</table>

- Weighted RF and GB methods outperformed comparison methods.
- Weighted methods outperform unweighted methods.
Results – Average Arrival Delays

- Similar results for weighted GB and weighted RF.
- Except for a couple outliers, residuals are fairly even.
## Results – Departure Cancelations

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. Test Error</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted Average</td>
<td>9.41</td>
<td>0.0%</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>8.14</td>
<td>-13.5%</td>
</tr>
<tr>
<td>K Nearest Neighbors</td>
<td>8.14</td>
<td>-13.5%</td>
</tr>
<tr>
<td>Unweighted Random Forest</td>
<td>13.69</td>
<td>45.5%</td>
</tr>
<tr>
<td>Unweighted Gradient Boosting</td>
<td>7.47</td>
<td>-20.6%</td>
</tr>
<tr>
<td><strong>Weighted Random Forest</strong></td>
<td><strong>7.06</strong></td>
<td><strong>-25.0%</strong></td>
</tr>
<tr>
<td><strong>Weighted Gradient Boosting</strong></td>
<td><strong>6.53</strong></td>
<td><strong>-30.6%</strong></td>
</tr>
</tbody>
</table>

- Weighted RF and GB methods outperformed comparison methods.
- Weighted methods outperform unweighted methods.
Results – Departure Cancelations

• Similar results for weighted GB and weighted RF
• Less accurate when higher amounts of cancellations?
Gradient-Boosted Quantile Forests

• Gradient-Boosted Forests can be adapted to produce estimates of quantiles.

• Instead of fitting a new tree to the residuals of the previous stage, fit it to the negative gradient of a different loss function.

• Otherwise, method and intuition remains the same.
Computational Results

• We estimate 90% quantiles of arrival delays and departure cancellations.
• Testing procedure remains the same, but different baseline methods are used, and different loss function is used.
• Unweighted 90% quantile – prediction is 90% quantile of sample data regardless of weather/traffic and GDP conditions.
• Weighted 90% quantile – prediction is 90% quantile of sample data weighted by GWR weighting, using combined distance.
• Max of K nearest neighbors – take the maximum of the k closest points in the sample, using combined distance.
## Results – Average Arrival Delays

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. Loss</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted 90% Quantile</td>
<td>16.16</td>
<td>0.0%</td>
</tr>
<tr>
<td>Weighted 90% Quantile</td>
<td>12.78</td>
<td>-20.9%</td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>10.85</td>
<td>-32.9%</td>
</tr>
<tr>
<td>Unweighted Gradient Boosting</td>
<td>4.60</td>
<td>-71.5%</td>
</tr>
<tr>
<td><strong>Weighted Gradient Boosting</strong></td>
<td><strong>4.45</strong></td>
<td><strong>-72.5%</strong></td>
</tr>
</tbody>
</table>

- Weighted GB is still best, but no significant improvement over unweighted GB
Results – Departure Cancelations

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. Test Error</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted 90% Quantile</td>
<td>14.93</td>
<td>0.0%</td>
</tr>
<tr>
<td>Weighted 90% Quantile</td>
<td>12.98</td>
<td>-13.1%</td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>13.18</td>
<td>-11.7%</td>
</tr>
<tr>
<td>Unweighted Gradient Boosting</td>
<td>4.23</td>
<td>-71.7%</td>
</tr>
<tr>
<td>Weighted Gradient Boosting</td>
<td>4.56</td>
<td>-69.5%</td>
</tr>
</tbody>
</table>

• Weighted GB performs slightly worse than unweighted. Overfitting?
Use in Decision Support Tool

• Prototype decision support tool in development
  • Part of NASA-funded ‘Similar Days Project’ in collaboration with UC Berkeley and ATAC.

• Goal: support for planning and evaluation of Traffic Management Initiatives (TMIs).

• Approach:
  1. User inputs a reference day (e.g. current day or day under evaluation)
  2. Identify historical days similar to reference day, find the TMIs that occurred.
  3. Choose set of historical TMIs that are representative of TMIs from step 2.
  4. Provide estimated performance of TMIs from step 3.
Similar Days Engine

Days similar to reference day.

Representative set of TMIs on similar days

Estimated TMI Performance

Data visualization and presentation.

Reference Day

Weather & Traffic Data

TMI Data

Representative TMI Finder

KPI Data

TMI Performance Prediction Model
Reference Day Specification

Select Analysis Day: 06/12/2014

Select Scope:

Select Airport: EWR

Apply

Clear
## Similar Days

### EWR, 06/12/2014

<table>
<thead>
<tr>
<th>View</th>
<th>Include</th>
<th>Date</th>
<th>Similarity Rank</th>
<th>IFR Exposure</th>
<th>Local Thndstm. Exposure</th>
<th>Moderate Wind Exposure</th>
<th>High Wind Exposure</th>
<th>Weighted Avg. Wind Direction</th>
<th>Snowfall (inches)</th>
<th>Max 1hr Arrival Demand</th>
<th>Max 4hr Arrival Demand</th>
<th>Max 1hr Total Demand</th>
<th>Max 4hr Total Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>06/12/2014</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>02/05/2014</td>
<td>1</td>
<td>0.95</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>45</td>
<td>159</td>
<td>78</td>
<td>301</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>05/15/2014</td>
<td>2</td>
<td>0.96</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>120</td>
<td>0</td>
<td>47</td>
<td>154</td>
<td>79</td>
<td>293</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>04/30/2014</td>
<td>3</td>
<td>0.57</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>64</td>
<td>0</td>
<td>44</td>
<td>152</td>
<td>83</td>
<td>300</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>07/04/2014</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>344</td>
<td>0</td>
<td>44</td>
<td>156</td>
<td>76</td>
<td>264</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>12/23/2014</td>
<td>5</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>46</td>
<td>157</td>
<td>82</td>
<td>308</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>12/07/2011</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>43</td>
<td>162</td>
<td>80</td>
<td>296</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>05/09/2014</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>109</td>
<td>0</td>
<td>42</td>
<td>155</td>
<td>76</td>
<td>288</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>04/23/2011</td>
<td>8</td>
<td>0.95</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>186</td>
<td>0</td>
<td>40</td>
<td>148</td>
<td>71</td>
<td>264</td>
</tr>
</tbody>
</table>
### Representative TMI Options

<table>
<thead>
<tr>
<th>Day Similarity options:</th>
<th>30 Nearest</th>
<th>50 Nearest</th>
<th>100 Nearest</th>
<th>Suggested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Number of Representative TMI:</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of outliers:</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max distance in scope:</td>
<td>7 airports</td>
</tr>
<tr>
<td>Max distance in file time:</td>
<td>208 minutes</td>
</tr>
<tr>
<td>Max distance in GS Duration:</td>
<td>31 minutes</td>
</tr>
<tr>
<td>Max distance in proportion of flights allowed:</td>
<td>1.19</td>
</tr>
<tr>
<td>Number of days with TMI:</td>
<td>24</td>
</tr>
</tbody>
</table>

- User can select how many representative TMI and which set of days representatives are taken from
Representative TMI\text{s}

<table>
<thead>
<tr>
<th>Date</th>
<th>Prevalence</th>
<th>File Time GS</th>
<th>Start Time GS</th>
<th>Stop Time GS</th>
<th>Scope GS</th>
<th>File Time GDP</th>
<th>Start Time GDP</th>
<th>Stop Time GDP</th>
<th>Scope GDP</th>
<th>Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/24/2014</td>
<td>0.04</td>
<td>14:37</td>
<td>14:19</td>
<td>15:30</td>
<td>1st Tier</td>
<td>15:12</td>
<td>15:08</td>
<td>18:59</td>
<td>1000</td>
<td>33/34/34</td>
</tr>
<tr>
<td>5/15/2014</td>
<td>0.17</td>
<td></td>
<td>7.02</td>
<td>8.00</td>
<td>23:59</td>
<td></td>
<td>1425</td>
<td></td>
<td></td>
<td>32/32/32</td>
</tr>
<tr>
<td>1/16/2013</td>
<td>0.54</td>
<td>8.43</td>
<td>8.37</td>
<td>23:59</td>
<td></td>
<td></td>
<td>1400</td>
<td></td>
<td></td>
<td>28/28/28</td>
</tr>
<tr>
<td>9/25/2014</td>
<td>0.04</td>
<td>15:34</td>
<td>17.30</td>
<td>22:59</td>
<td></td>
<td></td>
<td>1200</td>
<td></td>
<td></td>
<td>34/34/34</td>
</tr>
<tr>
<td>6/7/2013</td>
<td>0.67</td>
<td>7.09</td>
<td>7.00</td>
<td>0.59</td>
<td>ALL+CZY_AP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
</tr>
</tbody>
</table>

• Our method would be used to provide additional information about these representatives.
Conclusions

• We provide two methods for estimating performance of GDPs in some weather/traffic conditions.

• These methods combine forest-based machine learning methods with a weighting scheme from GWR.

• The methods provided less test error than comparison methods when predicting average arrival delays and departure cancellations.

• The methods can produce estimates of any arbitrary performance measure.

• One of these methods can also be used to estimate quantiles of performance measures.

• This method produced less loss than comparison methods, although weighting scheme did not seem to improve the loss of this method.
Further Work

• Methods that incorporate relationships between dependent variables.
• Identifying when there is insufficient data to make an accurate estimate for some set of GDP parameters.
• Estimates for how revisions of GDPs will perform.
References


References (2)


• 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.
References (3)


References (4)


