# Selecting Parameters in Performance-Based Ground Delay Program Planning

Alexander Estes[1,2], David J. Lovell[1,2,3], and Michael O. Ball[1,2,4]
1. Applied Mathematics & Statistics, and Scientific Computation Program
2. Institute for Systems Engineering
3. Department of Civil and Environmental Engineering
4. R. H. Smith School of Business
University of Maryland
College Park, MD, USA
aestes@math.umd.edu, lovell@umd.edu, mball@rhsmith.umd.edu

***Abstract*—In this paper, we consider the problem of selecting a set of parameters for a Ground Delay Program so that the program achieves a vector of performance objectives that is similar to a target vector. This could be used to support consensus-based ground delay program planning. We propose a method that selects several potential candidates of vectors, and we compare our method with a simple greedy algorithm. Our results indicate that our proposed method is able to provide multiple solutions that are closer to the efficient frontier than the greedy solution.**

***Keywords-Ground Delay Programs; multi-objective optimization, geographically-weighted random forests, parameter selection***

## I. INTRODUCTION

The amount of traffic scheduled to arrive at an airport can exceed the capacity of that airport to handle traffic. This situation often occurs when adverse weather decreases the capacity of an airport. When such an imbalance is anticipated, the Federal Aviation Administration (FAA) can issue a *Ground Delay Program* (GDP). This is a traffic management initiative that delays flights that are destined for the impacted airport, but that have not yet departed. These flights would then arrive at a later time in which the airport has sufficient capacity. In this way, delays are taken on the ground rather than in the air, which is more efficient and avoids unsafe situations. GDPs are implemented frequently. For example, according the National Traffic Management Logs (NTML), the FAA implemented more than 900 GDPs in 2014. Some airports are especially prone to this type of management initiative; Newark Liberty International Airport (EWR), LaGuardia Airport (LGA), John F. Kennedy International Airport (JFK) and San Francisco International Airport (SFO) each experienced more than 100 GDPs in 2014.

Planning a GDP involves selecting several parameters, such as the number of flights that are permitted to arrive at the airport, and the duration of time in which the restrictions are in effect. This is not a trivial task, as there are many sources of uncertainty, including some sources that are difficult to characterize or predict. Weather can deviate from the forecast, flights do not always arrive on schedule, and flight operators may alter their schedules in response to the actions taken by the FAA. There have been some proposed methods for planning GDPs [1, 2, 3, 4, 5, 6, 7, 8]. These existing methods tend to have an objective function that minimizes a weighted sum of assigned ground delays and expected air delays, and for most of these methods it would be difficult to incorporate other types of performance measures.

Recent work suggests that there are multiple criteria that may be used to evaluate the performance of traffic management initiatives. In [9], several criteria for evaluation of GDPs were proposed. There are usually trade-offs between these criteria, and the goals of flight operators can vary day-to-day, so a well-executed traffic management initiative should balance the performance criteria in a way that matches the goals of the flight operators on that day. There has been some work towards planning GDPs within this multi-criteria paradigm. A GDP planning model that incorporates a predictability criterion alongside the usual delay objectives was proposed in [10].

We propose a more flexible approach that can be used with any performance criteria with no alterations. A mechanism, called Consensus Service Expectation Level setting (COuNSEL), has been developed that would provide a "consensus vector" of performance objectives based on the input of flight operators [11]. This consensus vector is simply a vector of multiple performance objectives, which reflect different aspects of performance. In the existing work on the COuNSEL mechanism, the consensus vector was three-

dimensional, and its components consisted of measures of efficiency, throughput, and predictability. The intent of the COuNSEL mechanism was that decision-makers at the FAA would then attempt to implement a ground delay program that achieves, as closely as possible, the specified vector of performance objectives. The main contribution in this paper is to provide a method for selecting the parameters of a GDP in order to achieve a specified balance of performance objectives under a given set of weather and traffic conditions. This would allow such a consensus mechanism to be implemented. More broadly, the work in this paper can be used to find a set of GDP parameters whose expected performance is close to some target vector of performance measures, regardless of the source of the vector.

## II. METHODOLOGY

### A. Background: Geographically-weighted Random Forests

We make use of the Geographically-Weighted Random Forest (GWRF) method, which was proposed in [12] as a method for estimating the expected performance that would result from implementing a GDP with a given set of parameters in a given set of weather and traffic conditions. The method is a supervised method, so it requires a training set of observations, where each observation describes a set of GDP parameters and the resulting performance that these parameters achieved. We would expect that these observations occurred under varying weather and traffic conditions, and the GWRF method requires a numerical measure of distance between each observed set of conditions and those conditions in which a prediction is desired. At least one such measure of distance has been proposed [13].

The GWRF method is built from collections of decision trees, which are rooted binary trees. Each non-leaf node has an associated decision rule, which can be written in the form '$x < a$' where $x$ is a variable and $a$ is a value. Each leaf node has an associated prediction for the target variables. Predictions for a specific set of values taken by the explanatory variables are produced by traversing the tree, starting at the root node. At each non-leaf node, if the decision rule is satisfied, then the traversal continues at the left node, while if the decision rule is not satisfied, then the traversal continues at the right node. Once a leaf node is reached, the corresponding prediction is returned. See [14] for details on how these trees are fit.

Random Forest (RF) estimators are formed by *bagging* decision trees. Bagging is a technique for improving the accuracy of an estimator in which copies of a data set are created by a bootstrap procedure. For each resampled data set, a corresponding copy of the estimator is fit. Some randomness is introduced into this fitting procedure to reduce correlation between the resulting estimates. The final estimator is produced by averaging these estimators. The averaging procedure improves the accuracy of the estimator by reducing the variance. For more details about RF estimators, see [15, 16], and for more details about bagging see [17].

A GWRF is a RF estimator in combination with a geographical weighting scheme, similar to that used in Geographically-Weighted Regression [18, 19]. In GWRF, this weighting works as follows. A different RF estimator is fit for each set of weather and traffic conditions for which we wish to provide estimates of GDP performance. The explanatory variables are the specified parameters of the GDP, while the target variables are the performance criteria. When this estimator is fit, higher weight is given to observations that occurred in similar conditions, while lower weight is given to observations that occurred in dissimilar conditions. This weight is generated by transforming the distance between the conditions (as provided by the aforementioned measure of distance) with a kernel function, which transforms large distances into small weights and small distances into large weights. The GWRF method is central to our method for identifying the set of parameters that would achieve the correct balance of performance criteria.

### B. Using GWRF in GDP Planning

The most straightforward manner of using the GWRF estimation method to produce a GDP given a set of conditions is as follows. First, fit a GWRF model for the given set of conditions using the available data on GDP parameters. Next, identify the set of GDP parameters whose predicted performance under the GWRF is closest to the consensus vector. The structure of the GWRF estimator is complicated, and it is not an easy problem to assign values for the explanatory variables in such a way that the predicted value is as close as possible to a given value. One possible heuristic would be to iterate through each observed set of GDP parameters, calculate the corresponding prediction, and select the set of GDP parameters whose predicted performance is closest to the consensus vector. As long as the training set is sufficiently complete and the consensus vector is attainable under the given conditions, then this procedure is likely to produce a GDP plan whose estimated performance is close to the consensus vector. Once the GWRF is fit, predictions for a given set of GDP parameters can be produced very quickly, so this approach is also computationally tractable for any realistically-sized data set of GDPs. We will refer to this procedure as the *naïve greedy method.*

The naïve greedy method can be improved upon. One potential downside of the procedure is that the estimate of the GDP performance is not perfectly accurate, and furthermore, the accuracy may vary depending on the conditions and the selected GDP parameters. For this reason, it may be desirable to limit the GDP parameters under consideration to those whose performance can be estimated with relatively high confidence. We propose a method for implementing a constraint on the GDP parameters in section II.C. It also may be possible to find parameters whose estimated performance is

Research supported by NASA grant no. NNX14AJ79A

better than the consensus vector, which we discuss in section II.D.

### C. Prediction-Weighted Similarity Measure

Since GWRF develops estimates of GDP performance from the available data, the reliability of the resulting predictions is dependent on the quantity of available data. Given a set of GDP parameters and a set of conditions, if many similar GDPs have been conducted in similar conditions, then we expect the estimation to be reliable. We can develop a measure of this as follows. Given a set of GDP parameters *x* and a set of weather and traffic conditions *z*, then the estimate produced by the GWRF can be expressed as:

$$\hat{g}(x;z) = \sum w_i(x;z) y_i \tag{1}$$

where the values $w_1, \ldots, w_n$ are non-negative weights that sum to one, and the values $y_1, \ldots, y_n$ are the performance measures in the observations of the data set. In other words, the estimated performance is given by a weighted sum of the performance of the observations in the data set, but the weights depend on the given situation and set of GDP parameters. If similar GDPs have been run in similar situations, then the method would assign high weight to those GDPs. However, if such similar observations do not exist, then the method must assign weights to less similar observations. Following this intuition, we propose a measure called the *prediction-weighted similarity* measure:

$$\hat{s}(x;z) = \sum w_i(x;z) s_i(z) \tag{2}$$

where the value $s_i(z)$ is a measure of similarity between weather and traffic conditions of the $i^{\text{th}}$ observation and those of *z*. The values $s_i(z)$ are already used in the construction of the GWRF prediction, so no extra work is required to define or compute these values. If the predictor primarily makes use of observations that occurred in similar situations, then the prediction-weighted similarity measure will be high, while if the predictor makes use of observations from less similar situations, this measure will be lower.

### D. Moving Closer to the Efficient Frontier

When the COuNSEL mechanism identifies a consensus performance vector, it attempts to identify one that is close to the efficient frontier of attainable vectors. However, since this frontier is generally not known with certainty, it is possible that the mechanism can generate a dominated performance vector. Also, performance vectors could be generated by other means, possibly also resulting in dominated vectors. Here, when we say the performance vector is dominated, we mean that there is a GDP whose expected performance is better in all performance measures than the dominated vector. In this case, we would like to be able to find a GDP whose performance dominates the consensus vector. This would likely be more desirable.

However, we must keep in mind that our estimates are not certain, and even if the estimated performance of some GDP dominates the estimated performance of another GDP, the true performance may not follow the same dominance relation. In order to account for this uncertainty, we use a scoring scheme similar to that described in [20]. Suppose there is some set of parameters *S* to which we wish to assign scores, and a set of weather and traffic conditions *z*. For each point in *S*, the dominance score is given by:

$$r(s;z) = \sum_{t \in S} P(t \leq s;z) + \frac{1}{2} P(t \sim s;z) - \frac{1}{2}. \tag{3}$$

In this equation, $P(t \leq s;z)$ is the probability that the performance of a GDP with parameters *s* dominates a GDP with parameters *t* under the conditions *z*. The value $P(t \sim s;z)$ is the probability that the performance of a GDP with parameters *s* neither dominates nor is dominated by that of a GDP with parameters *t* under the conditions *z*. This scheme assigns lower scores to points that dominate most other points with high probability and gives higher scores to points that are likely to be dominated by other points. Naturally, the probabilities used in this scheme are not known, so they must be estimated.

We make use of the bagged estimators in GWRF in order to estimate these probabilities. As discussed in Section II.A., the GWRF is formed by independently fitting many decision tree models, each of which is fit on a different resampled data set. If all of these models predict that the GDP parameters *s* dominate the GDP parameters *t*, then we take that to indicate that *s* would dominate *t* with high probability. Thus, we define our estimate $\bar{P}(t \leq s;z)$ to be the proportion of the decision trees that predict that GDP *s* would dominate *t* under conditions *z*. We similarly define the estimate $\bar{P}(t \sim s;z)$ to be the proportion of the decision trees that predict that neither set of parameters would dominate the other under the weather and traffic conditions *z*.

### E. A Constrained Greedy Selection Process

We propose a method for selecting GDP parameters that combines a greedy selection process with the prediction-weighted similarity measure (described in Section II.C) and the dominance-scoring scheme (described in Section II.D). We will refer to this method as the constrained greedy selection (CGS) method. The CGS method assumes that GDP parameters are desired for weather and traffic conditions *z*, that a GWRF for these conditions has been fit to the observed data, and that there is a known, desired threshold *s** for the prediction-weighted similarity measure. In our method, we only allow selections of GDP parameters whose prediction-weighted similarity (defined in Section II.C) is less than *s**.

The effect that this parameter has on results is discussed in Section III.B.

The selection process in the CGS is defined as follows. Let $z$ be the set of weather and traffic conditions in which we wish to plan a GDP. We take the set $S$ of all observed GDPs in the data set to be the initial set of choices for the GDP parameters. Next, we calculate the prediction-weighted similarity for each choice of GDP parameters $x$, and any GDPs such that $\hat{s}(x; z)$ is greater than $s^*$ are removed from $S$. From the reduced set $S$, we select a list $L$ of promising choices of GDP parameters. This list is constructed so that the expected performance of the GDPs in the list increases in distance from the target performance and decreases in dominance score (as defined in section II.D). In other words, elements earlier in the list have performance that is closer to the target performance, but elements later in the list have better expected performance. In this way, we can provide a variety of options to the GDP planner, and the planner can select the GDP parameters that balance the overall performance of the method with the distance from the target vector.

The first element that we place in the list $L$ is the GDP $x^1$ from $S$ whose estimated performance in the GWRF is closest to the target performance vector. If there are multiple such GDPs, then we select the one with the lowest dominance score. After this first element has been selected, we proceed as follows. We let $x^{k+1}$ be the element whose estimated distance from the target vector is lowest amongst those elements whose dominance score is less than $x^k$, if such an element exists. If no such element exists, then we stop.

## III. Computational Experiments

### A. Data and Setup

For these experiments, we used data concerning ground delay programs that were implemented at Newark Liberty International Airport (EWR). Observed GDP parameters were taken from the FAA's National Traffic Management Log (NTML) data, and the distances between the corresponding weather and traffic conditions were generated by the procedure described in [13]. Our data set includes 480 observations that occurred between 2011 and 2014. From these, 80% of the observations were randomly selected to constitute a training set, while the remaining 20% of observations were placed in a test set.

We used a set of five GDP parameters, which are roughly consistent with work such as [12]. These are as follows:

- *Entry Time*: the time at which the GDP is announced, which we express in the number of minutes after 4:00 a.m. local time.
- *Earliest ETA*: this is the earliest time at which flights receive restrictions. Any flight whose estimated arrival time is before the earliest ETA will not be controlled by the GDP. As with entry time, we express this time in minutes after 4:00 a.m. local time.
- *Duration*: the length of time in which the GDP is planned to be in effect, expressed in minutes.
- *AAR Average*: the averaged planned number of flights that are allowed to arrive at the airport for each hour in which the GDP is in effect. The units for this feature are flights/hour.
- *Number of Core 30 Airports Within Scope*: every GDP has a declared scope, which describes a geographic region. Flights departing from outside this region are not controlled by the GDP. This feature describes the number of Core 30 airports that fall within this scope, where Core 30 is a list of the 30 airports in the U.S. that have the most traffic.

We use two measures of GDP performance. The first measure is the average arrival delay experienced by flights arriving at the impacted airport in the day. This measure was calculated from Aggregate Demand List (ADL) data. The ADL is maintained by the FAA and provides information about flights, such as estimated and actual arrival and departure times. The second measure that we use is the total number of holding events in the day, which is the number of times that a flight had to wait in the air to land at the impacted airport. This information is present in the Aviation System Performance Metrics Database, which is also maintained by the FAA. These two performance measures were chosen because they reflect two aspects of GDP performance that cannot be simultaneously optimized. A strict GDP will generally lead to fewer holding events but will cause higher arrival delays. Conversely, if the restrictions placed by the GDP are less strict, then there will be more holding events and lower arrival delays. While we believe these are reasonable measures of performance, there is ongoing research and discussion as to which performance measures are the most pertinent for GDP planning (see for example [9] or [11]). Since this method is data-driven, any performance measure could be substituted for these measures.

The GWRF model that we use is fit and tuned as described in [12]. In all of the experiments discussed in this section, the training set is used to tune the model and serves as the set of observations for the CGS method and the naïve greedy method. Each observation in the test set provides a weather and traffic situation, a set of GDP parameters, and the actual performance achieved on that day. However, we do not use the GDP parameters in the test set in our computational experiments. Instead, we make use of the weather and traffic and the performance observations. We treat each observation of the performance in the test set as if it were the target performance vector. In this way, the test set provides a set of weather and traffic conditions and corresponding target performance vectors.

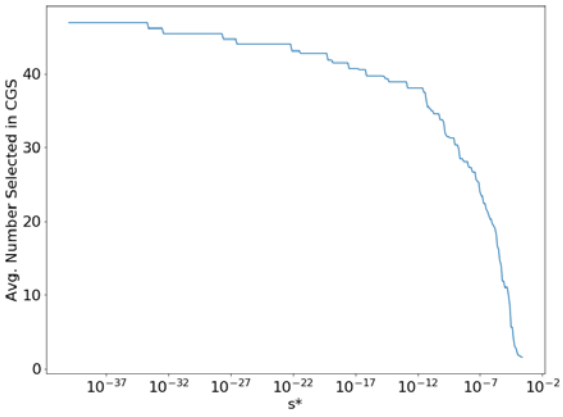


Figure 1. Number selected by CGS, plotted against the threshold $s^*$.

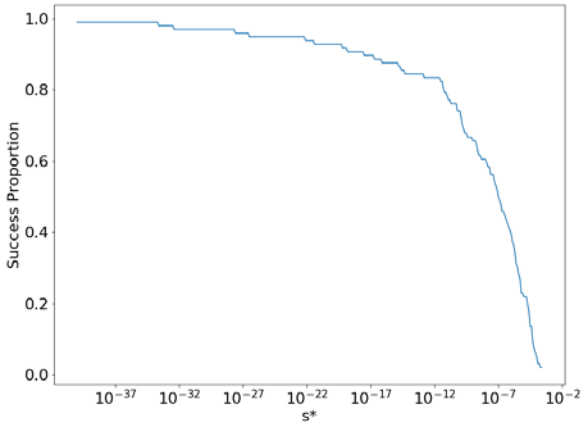


Figure 2. Proportion of tests for which the CGS selects at least one element, plotted against the threshold $s^*$.

### *B. Evaluation of Performance of CGS method.*

The results of the CGS method depend on the selected threshold for the probability-weighted similarity. A lower threshold means there is a tighter constraint on the parameters that are accepted while a higher threshold means that there is a looser constraint. If the threshold $s^*$ is high enough, then the first element selected by the CGS will be exactly the same as the naïve greedy method. As the threshold $s^*$ increases, the number of elements selected by the CGS decreases, and if the threshold $s^*$ is small enough then the constrained greedy procedure will not return any solution. The CGS procedure was run for each of the target vectors in the test set and for a variety of values of $s^*$. The measure of similarity used in these experiments is that defined in [12], which is produced by taking an exponential function of the negative squared distance. This tends to make the resulting predicted-weighted similarities quite small, and makes it sensible to use a log scale for the similarity threshold.

The average number of elements produced by the procedure is shown for each choice of threshold in Figure 1, and the proportion of test vectors for which the CGS successfully returns at least one element is shown for each threshold in Figure 2. When the threshold $s^*$ is set to values between $10^{-40}$ and $10^{-35}$, then the CGS almost always produces at least one element, and on average returns nearly 50 elements. As the threshold increases from $10^{-40}$ to $10^{-10}$ the number of elements selected by the CGS decreases relatively slowly. When the threshold is greater than $10^{-10}$, this decrease happens at an increasingly fast pace. The proportion of tests for which the CGS selects at least one element follows a similar trend. These results indicate that as long as the threshold is set less than $10^{-11.5}$, then the CGS will be able to provide more options to the planner in most cases.

The relative quality of the options provided by the CGS is demonstrated in Figures 3 and 4. By definition, any element in the results produced by the CGS must have expected performance that is further from the target performance vector than the naïve greedy method. For each threshold $s^*$, Figure 3 shows the percent increase in distance from the target vector when comparing the naïve greedy result with the element of the CGS whose expected performance is closest to the target (for those test instances in which the CGS selected at least one element). This trend is similar to that of the proportion of tests in which the CGS selects at least one element and the average number of elements selected by the CGS. When the threshold $s^*$ is less than $10^{-11}$, then the closest element selected by the CGS is not much further than the one selected by the naïve greedy result. For threshold values greater than $10^{-11}$, the distance between the CGS element and the target vector increases sharply. Figure 4 shows the relative change in dominance score between the naïve greedy solution and the CGS element with the lowest score, plotted against the threshold $s^*$. When the similarity threshold is relatively low, the CGS method is able to identify solutions whose score is on average less than half of the score of the naïve greedy solution. As the threshold increases, this improvement in score tends to get smaller, but even for relatively large thresholds (i.e. when $s^*$ is less than $10^{-5.5}$), the CGS method is able to identify GDP parameters whose estimated performance has a lower dominance score than the naïve greedy solution.

In summary, the performance of the CGS method is dependent on the similarity threshold $s^*$ that is allowed. However, for a wide range of values for this parameter, the CGS method is able to identify GDP parameters whose estimated performance is close to the target vector and is able to identify GDP parameters whose dominance score is lower than those produced by the naïve greedy solution.

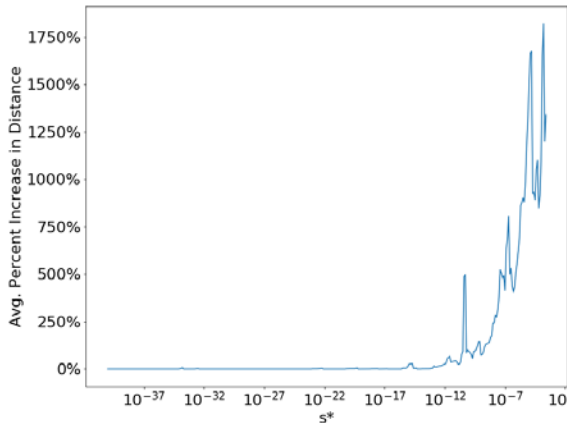


Figure 3. Average percent increase in distance from target performance of CGS method compared with naïve greedy method, plotted against the threshold *s**.

### *C. Example of Results From Selected Days*

In order to demonstrate the utility of this method, we display the results of this method when applied to two days, specifically November 14th, 2011 and May 28th, 2013. As before, the actual performance achieved on these days is used as the target vector. These days were chosen because on the latter day the achieved performance appears to be close to the efficient frontier, while on the former day the performance appears to be further from the frontier. Thus, examining these two days can provide some intuition about how the method performs in these two situations. For both of these days, we ran the CGS method with two values of the threshold *s**, specifically $10^{-5}$ and $10^{-10}$ respectively.

The results from the CGS method are plotted in Figure 5. In each plot, the target performance is displayed as a black `x', the expected performance vectors of the GDP parameters selected by the CGS are shown as larger orange circles, and the expected performance of the naïve greedy solution is shown as a magenta triangle. In order to provide some context, the expected performance vectors for the forty choices of GDP parameters with the highest prediction-weighted similarity scores are plotted as small blue circles. On November 11th, 2014 there are many potential choices of GDP parameters that seem to dominate the target prediction vector. When the threshold *s** is smaller, the CGS identifies many potential choices for GDP parameters. Even when the threshold is relatively large, the CGS is still able to identify some GDP parameters that dominate the target performance. On May 28th, 2013 the target performance seems to be closer to the boundary of attainable performance, and the CGS identifies fewer potential choices of parameters. For the smaller value of the threshold *s**, the CGS method is still able to identify an ample set of parameter choices. The CGS method does not produce any choices of parameters when the threshold is set to the larger value. That is, there are no historically-observed GDPs whose prediction-weighted similarity score is greater than $10^{-5}$. This indicates that there are more data available on conditions similar to those present on November 11th, 2014 than those present on May 28th 2013.

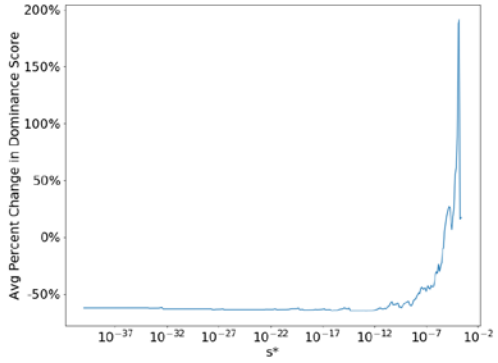


Figure 4. Average percent change in dominance score of CGS method compared with naïve greedy method, plotted against the threshold *s**.

## IV. Conclusion

We propose a new method to identify a set of GDP parameters whose estimated performance is close to a target vector of performance objectives. In addition, our method also ensures that the parameters selected have estimates that are well-supported by the available data, and is able to identify alternatives that may dominate the target vector. We compared our method against a greedy approach, and we demonstrated that our method is able to provide more options and can often identify solutions whose performance is likely to dominate that of the greedy method.

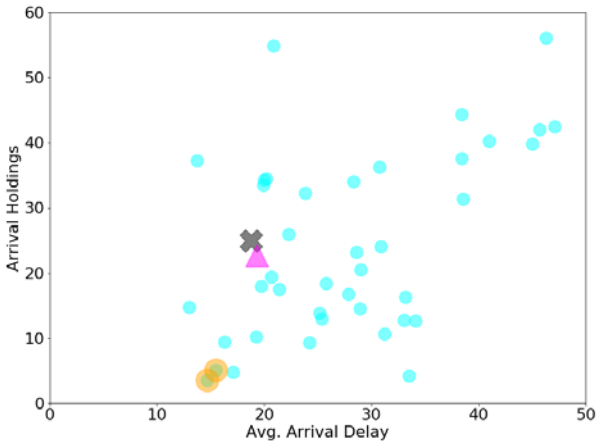


Figure 5a: Results from November 14th, 2011; $s^*$ set equal to $10^{-5}$

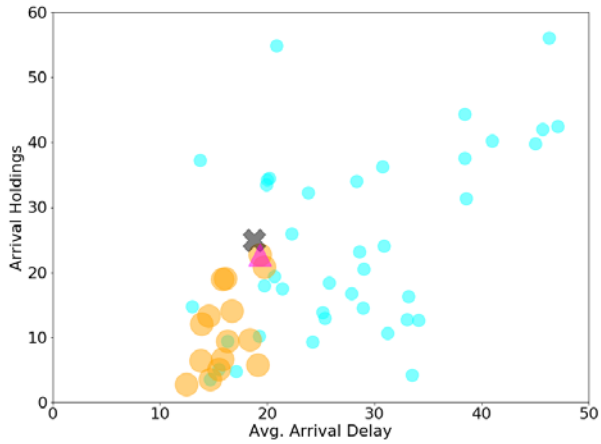


Figure 5b: Results from November 14th, 2011, $s^*$ set equal to $10^{-10}$

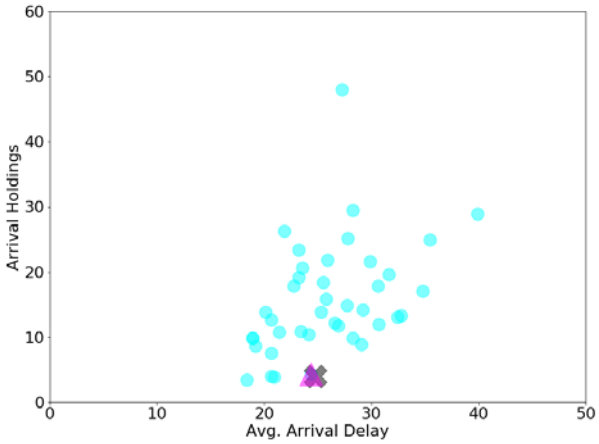


Figure 5c: Results from May 28th, 2013; $s^*$ set equal to $10^{-5}$

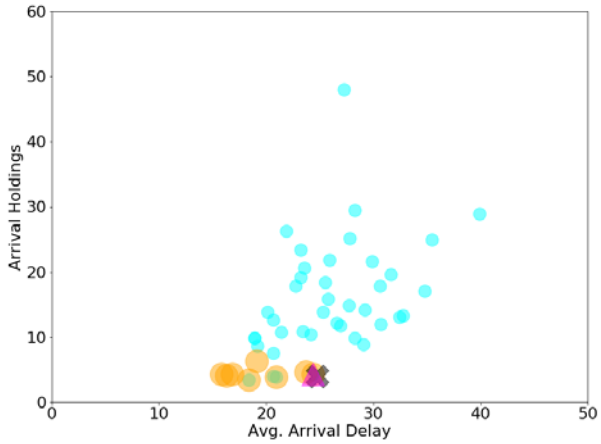


Figure 5d: Results from May 28th, 2013; $s^*$ set equal to $10^{-10}$

Figure 5. Results from CGS. Target performance vector is plotted as an `x'; greedy solution is plotted as a triangle; the expected performance of the GDP parameters selected by the CGS method are plotted as larger orange circles, while the expected performance of the choices of parameters with the highest similarity score are plotted as smaller blue circles.

This method could be used in conjunction with a collaborative decision-making mechanism such as COuNSEL. More broadly, our method could be incorporated into other types of decision-support systems. Further study is necessary for this method to be implementable in practice. The results from this method are dependent on the threshold $s^*$, so some work would be required to identify an appropriate value for this parameter. As we discussed, in some cases the CGS fails to produce a solution. This indicates that the observations occurring in situations similar to the situation in which the method is run are not sufficient to provide stable estimates of GDP performance. In this case, an alternative method is required. There is also more work that could be done in the presentation of these methods, so that the results can be displayed in an easier-to-interpret manner.

In this work, we considered the problem of planning a single GDP at a single airport. There are other types of traffic management initiatives that this method could be applied to, such as airspace flow programs. Similar methods could also be developed to coordinate multiple traffic management initiatives, or to help produce more comprehensive types of air traffic management plans.

## ACKNOWLEDGMENT

We thank our collaborators Mark Hansen, Alexei Pozdnukhov, Sreeta Gorripaty, Yulin Liu, and Yi Liu at U.C. Berkeley and Kennis Chan, Corey Warner, and John Schade at Airborne Tactical Advantage Company for their assistance in procuring and processing data.

## REFERENCES


[1] A. Mukherjee and M. Hansen, "A Dynamic Stochastic Model for the Single Airport Ground Holding Problem," Transportation Science, vol. 41, no. 1, pp. 444–456, 2007.

[2] M.O. Ball, R. Hoffman, A.R. Odoni, and R. Rifkin, "A Stochastic Integer Program with Dual Network Structure and its Application to the Ground-Holding Problem," Operations Research, vol. 51, no. 1, pp. 167–171, 2003.

[3] O. Richetta and A.R. Odoni, "Dynamic Solution to the Ground-Holding Problem in Air Traffic Control," Transportation Research Part A: Policy and Practice, vol. 28, no. 3, pp. 167–185, 1994.

[4] O. Richetta and A.R. Odoni, "Solving Optimally the Static Ground-Holding Policy Problem in Air Traffic Control," Transportation Science, vol. 27, no. 3, pp. 228–238, 1993.

[5] L.S. 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, 2010.

[6] J. Cox and M. J. Kochenderfer, "Optimization Approaches to the Single Airport Ground-Holding Problem," Journal of Guidance, Control, and Dynamics, vol. 38, no. 12, pp. 2399-2406, 2015.

[7] P.-c.B. Liu and M. Hansen, "Scenario-Free Sequential Decision Model for the Single Airport Ground Holding Problem," in Proceedings of the 7th USA/Europe Air Traffic Management R&D seminar. Barcelona, 2007.

[8] M.O. Ball, R. Hoffman, and A. Mukherjee, "Ground Delay Program Planning Under Uncertainty Based on the Ration-by-Distance Principle," Transportation Science, vol. 44, no. 1, pp. 1–14, October 2009.

[9] Y. Liu and M. Hansen, "Evaluation of the Performance of Ground Delay Programs," Transportation Research Record, vol. 2400, pp. 54–64, 2013.

[10] Y. Liu and M. Hansen, "Incorporating Predictability Into Cost Optimization for Ground Delay Programs," Transportation Science, vol. 50, no. 1, pp. 132–149, June 2015.

[11] M. Ball, P. Swaroop, M. Hansen, L. Kang, Y. Liu, C. Barnhart, C. Yan, and V. Vaze, "Service level expectation setting for air traffic flow management: Practical challenges and benefits assessment," in Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar, Seattle, WA, 2017.

[12] A. Estes, M. Ball and D. Lovell, "Predicting Performance of Ground Delay Programs," in Proceedings of the 12th USA/Europe Air Traffic Management Research and Development Seminar, Seattle, WA, 2017.

[13] S. Gorripaty, Y. Liu, M. Hansen, and A. Pozdnukhob, "Identifying similar days for air traffic management," Journal of Air Transport Management, vol. 65, pp. 144-155, October 2017.

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

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

[16] C. Chen, A. Liaw, and L. Breiman, "Using random forests to learn imbalanced data," Technical Report no. 666, U. C. Berkeley, July 2004.

[17] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, January 1995.

[18] 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.

[19] A.S. Fotheringham , C. Brunsdon, and M. Charlton, Geographically Weighted Regression. John Wiley & Sons: Chichester, 2002.

[20] E.J. Hughes, "Multi-objective Probabilistic Selection Evolutionary Algorithm (MOPSEA)," Technical Report no. DAPS/EJH/56/2000, Department of Aerospace, Power & Sensors, Cranfield University, 2000.