Ground Delay Program Analytics with Behavioral Cloning and Inverse Reinforcement Learning

Michael Bloem[1]
NASA Ames Research Center, Moffett Field, CA, 94035

Nicholas Bambos[2]
Stanford University, Stanford, CA 94305

We used historical data to build two types of model that predict Ground Delay Program implementation decisions and also produce insights into how and why those decisions are made. More specifically, we built behavioral cloning and inverse reinforcement learning models that predict hourly Ground Delay Program implementation at Newark Liberty International and San Francisco International airports. Data available to the models include actual and scheduled air traffic metrics and observed and forecasted weather conditions. We found that the random forest behavioral cloning models we developed are substantially better at predicting hourly Ground Delay Program implementation for these airports than the inverse reinforcement learning models we developed. However, all of the models struggle to predict the initialization and cancellation of Ground Delay Programs. We also investigated the structure of the models in order to gain insights into Ground Delay Program implementation decision making. Notably, characteristics of both types of model suggest that GDP implementation decisions are more tactical than strategic: they are made primarily based on conditions now or conditions anticipated in only the next couple of hours.

---

[1] Research Aerospace Engineer, Systems Modeling and Optimization Branch, MS 210-15. Member, AIAA. michael.bloem@nasa.gov
[2] Professor, Departments of Management Science & Engineering and Electrical Engineering. bambos@stanford.edu

Nomenclature

$\mathcal{A}$	set of all GDP implemented actions (GDP and no GDP)
a_t	GDP implemented action in t (GDP or no GDP)
AAR_t	airport arrival rate in t
b_t^{a}	number of flights in air buffer at the start of t
b_t^{g}	number of flights in ground buffer at the start of t
controlled_t	true if the arrival rate is controlled by a GDP in t and false otherwise
$f(s,a)$	$M \times 1$ vector of reward features for state-action pair (s,a)
GDP_t	true if a GDP is implemented in t and false otherwise
M	number of reward features
$q(s,a)$	score for state-action pair (s,a); used by π_{C}
$R(s,a)$	reward for using action a in state s
$R_{\text{C}}(s,a)$	reward relative to π_{C}
$\hat{R}_{\text{C}}(s,a)$	regressor that approximates $R_{\text{C}}(s,a)$
rate_t	controlled arrival rate for a GDP in t
$\mathcal{S}$	set of all states
s_t	system state in t
scheduled_t	scheduled number of arrivals in t
t	time step
γ	discount factor
π_{C}	deterministic score function-based multi-class classifier policy
$\hat{\pi}^{\star}_{\hat{R}_{\text{C}}}$	rollouts policy that approximately optimizes with respect to reward function $\hat{R}_{\text{C}}$
$\hat{\theta}$	$M \times 1$ vector of estimates of parameters in $\hat{R}_{\text{C}}$

I. Introduction

When predictions of air traffic and capacity suggest that an excessive number of flights will arrive at an airport at some future time, air traffic flow management (TFM) actions like Ground Delay Programs (GDPs) can be used to delay flights on the ground, where delay is relatively inexpensive. These actions are typically used when a weather event such as high winds or low ceilings reduces the capacity of an airport. GDPs are used hundreds of times per year at some airports [1], and each GDP can generate thousands of minutes of ground delay. TFM actions are selected by human decision makers who must rely primarily on experience and intuition because available forecasts of traffic and weather do not adequately account for uncertainties and because they have access to few "what-if" simulation capabilities or other decision-support tools [2, 3]. Furthermore, differences in style or preferences may lead decision makers to select different actions even when confronted with the same situation [4]. Given the importance of TFM actions like GDPs and the variations in their usage that result from forecast uncertainties, the lack of decision-support tools, and differences in decision maker experience, style, or preferences, researchers have mined historical data in an effort to better understand and predict these actions.

In particular, researchers have begun applying *imitation learning* techniques in which demonstration expert actions found in historical data are used to develop models that mimic expert actions [5]. Some studies have used traditional classification or clustering techniques to build models that predict or suggest TFM actions based on features describing the state of the air traffic system [6–10]. For example, Mukherjee, Grabbe, and Sridhar compared the performance of two classification models (logistic regression and decision tree) that they trained to predict the probability of GDP implementation at an airport based on features describing the current weather and traffic state at the airport [10]. This approach to imitation learning is known as *behavioral cloning* (BC), and it assumes that expert actions can be characterized and modeled as a reaction to the current state. An alternative imitation learning approach is *inverse reinforcement learning* (IRL). IRL is based upon a model of the system dynamics and how these dynamics are affected by actions; a Markov decision process (MDP) model is typical. IRL assumes that the expert is selecting actions in an attempt to maximize a total reward accumulated over time while operating within the

system model. IRL uses demonstration expert actions found in historical data to infer a reward function that is consistent with the expert actions (assuming strategic and total-reward-maximizing expert decision making). While BC can leverage powerful classification and clustering algorithms to make use of many features describing the system state, it does not consider system dynamics as explicitly as IRL [5]. Furthermore, in problems described as “long-range" and “goal-directed," IRL has been shown to produce models that generalize to new environments better than models produced by BC [11]. All TFM problems certainly involve a dynamical system and, since GDPs can last for more than ten hours and cause ground delays for flights not scheduled to arrive at a constrained airport for several hours, seem to be strategic. Although we are only aware of applications of IRL to systems considerably simpler than the selection of GDP actions, these characteristics of GDPs suggest that IRL may produce models that can predict and provide insight into expert GDP actions. Even if IRL models produce relatively poor predictions, the reward function inferred by IRL algorithms may provide a “succinct, robust, and transferable" definition of when to implement GDPs [12]. Insights gleaned from such a reward function could guide researchers as they develop TFM decision-support tools [4, 13].

Our objective is to build models that can (1) predict GDP actions and (2) provide insight into how and why the actions are selected. We are not working towards a particular application of these predictions and insights. They might be useful in the development of decision-support tools, for the evaluation of current procedures, or when simulating airspace systems. To build the models, we utilize historical data describing GDP actions and factors that might influence these actions. While this data-driven approach means that the models are necessarily rooted in the past and therefore will not account for the impact of new technologies or procedures, it also allows us to leverage powerful tools from machine learning and leads to quantified insights into historical GDP decision making. In this article, we make four main contributions. First, we deploy both BC and IRL techniques to model and predict hourly expert GDP implementation actions. As far as we know, this is the first application of IRL to a problem in air traffic management, and by comparing these two fundamentally-different approaches we gain greater insight into GDP implementation decision making. Second, by implementing an IRL technique, we infer a reward

function that is consistent with historical GDP implementation actions. While utility functions that may guide runway configuration decisions have been inferred from historical data using discrete-choice models [14], we are not aware of any other inference from historical data of a reward function that may be guiding TFM actions. Third, we analyze GDP implementation by focusing on important GDP initialization and GDP cancellation events, which occur ten or more times less frequently than non-initialization or non-cancellation events at most airports. This approach yields an operationally relevant evaluation of the predictive performance of models, and also more nuanced insight into GDP implementation decision making. Other research involving predicting GDP implementation has *not* distinguished between initialization and cancellation events (e.g., see Ref. [10]). Fourth, we use historical data to infer the degree to which future conditions are considered when GDP implementation decisions are made. Again, we are not aware of other work in which historical data was used to make this sort of inference.

The remainder of this article is structured as follows. Section II reviews the data we use in this analysis. Next, we specify the imitation learning GDP models that are developed in this research in Section III. We conduct parametric studies related to decision maker look-ahead and evaluate the quality of model predictions in Section IV. We also investigate the models to glean insights into GDP implementation decision making in that Section. Finally, we provide conclusions in Section V.

II. Data

In this Section, we describe the data that we used to quantify the system state for GDP models attempting to predict GDP actions. These data have been collected for EWR and SFO for the 151 days from 1 May 2011 through 29 September 2011, the 152 days from 1 May 2012 through 30 Septemeber 2012, and the 152 days from 1 May 2013 through 30 September 2013; there were 455 total days in the data set from these three summers. The causes of reduced airport capacity can vary between summer and winter months, so we simplified the problem by choosing to study only summer months. For example, some GDPs at EWR are reported as caused by “snow/ice," but by using only summer months we avoid the need to develop models that can account for such GDPs [1]. Night-time data were removed from consideration because, due to low traffic volumes, GDPs are

typically not used during the night. Hourly samples of each type of data were generated from 11:00 UTC (7:00 am EDT) through 06:00 UTC (2:00 am EDT) on the next day (20 hours per day) for EWR and from 12:00 UTC (5:00 am PDT) through 09:00 (2:00 am PDT) on the next day (22 hours per day) for SFO. Overall, this leads to 9100 hourly data points for EWR and $10,010$ for SFO.

We studied EWR and SFO because in recent years, GDPs have been used at these airports more than at any other airports [1]. Indeed, there were 208 and 297 GDPs utilized at EWR and SFO, respectively, during these 455 days. Assuming no more than one GDP was used per airport per day (as is typical), a GDP was utilized at EWR in more than 45% of these days and at SFO in more than 65% of these days. Furthermore, we selected these airports because we expect there to be meaningful differences in GDP decision-making practice for the two airports. This expectation is motivated in part by differences between the weather phenomena typically leading to decreased arrival capacities at these two airports (winds at EWR but low ceilings at SFO) [1]. This expectation is also motivated by differences between the characteristics of arrival traffic demand at the two airports. For example, due to its relative proximity to multiple metropolitan areas in the eastern and midwestern United States, a larger fraction of flights bound to EWR than to SFO originate within a relatively short distance of the airport.

At each hour, some predictions of future states are available. For example, arrival schedules are a simple prediction of future arrival traffic levels, and weather forecasts provide predictions of future weather conditions. The set of features for each hour includes these predictions at one-hour intervals starting an hour from the current time and extending to the hour starting ten hours from the current time. In all, 257 features describing the state were made available to the GDP models. These features will be described in the following sub-sections. Based on the results of experiments described in sub-section IV.A.1, we eventually only include features extending to the hour starting four hours from the current time. This leads to 125 features.

It may seem that by using hundreds of features to describe the state, we are providing the models with a relatively complete picture of the factors that influence GDP decision making. However, GDP decision making is so complex and can depend on so many factors that this is certainly not this case. Data quantifying some factors that may impact GDP decision making are not readily available. For

example, equipment outages or an arrival by Air Force One can reduce the arrival capacity of an airport and lead to the need for a GDP. Stakeholder (i.e., airline) preferences expressed during a planning teleconference can also impact GDP implementation decisions, but detailed minutes of these teleconferences are not available, and even if they were, it may be difficult to meaningfully quantify such expressions of preferences. To focus and simplify this study, we also chose *not* to consider certain other factors that impact GDP decision making, even though they are quantified in available data. For example, to account for how operations at nearby airports impact operations at these airports, we might have added features describing runway configurations at these other airports. The exclusion of these runway configuration data means that the models do not account for the impact of operations at nearby airports.

A. Weather Observations

Observations of weather conditions at an airport are recorded in METAR reports, which we retrieve from the FAA's Aviation System Performance Metrics (ASPM) database [15]. These data include the meteorological conditions (instrument or visual), ceiling height and its change since the last hour, visibility distance, wind speed, wind angle, and landing runway head wind and landing runway cross wind for the active runway configuration (eight features).

B. Weather Forecast

Terminal Aerodrome Forecasts (TAFs) provide predictions of weather conditions at an airport that extend 24 hours or more into the future, and they are often used by TFM decision makers. At EWR and SFO, TAFs were published at least every three hours during the time period we are studying. From TAFs, we extract predictions of conditions at the hour starting at the current time through the hour starting 10 hours from the current time. For each of these 11 hours, we include a feature for the meteorological conditions (visual, marginal visual, instrument, or low instrument), ceiling height and its change since the last hour, whether ceiling heights are "temporary," wind speed, wind direction, wind gust speed, landing runway cross wind speed for the most common runway configuration, landing runway head wind speed for the most common runway configuration, visibility distance, whether the visibility distance is greater than the specified quantity, whether the

visibility conditions are "temporary," the intensity of precipitation, the intensity of obscuration, and whether precipitation and obscuration are "temporary." There are 14 features specified for 11 hours and one (change in ceiling height) for 10 hours, which means there are 164 features derived from the TAF for each hour.

C. Number of Scheduled Arrivals

ASPM records also contain the number of scheduled arrivals at the airport during each quarter hour. We use these records to generate a feature denoting the scheduled arrivals during the hour starting at the current time and extending through the hour starting 10 hours from now (11 features).

D. Current Airport State

Other features that describe the current state of the airport are also extracted from ASPM records; two of these are the airport arrival rate (AAR) for the current hour and the runway configuration. The *deterministic departure queue* is a third such feature. It is calculated by constructing a simple deterministic queuing model based on the scheduled number of departures and the airport departure rate. Large values for this feature might be correlated with surface movement congestion and long takeoff queues.

E. Predictions of Future Airport Arrival Rates

GDPs are typically used when an airport's arrival capacity, quantified by the AAR, is expected to be too low to handle the predicted number of arrivals. The AAR selected by traffic managers depends on factors like runway configurations, weather conditions, and the type of aircraft that are scheduled to arrive at the airport. Other researchers have developed AAR prediction models [16–23], and we implemented a model similar to the bagged decision tree model proposed by Provan, Cunningham, and Cook in Refs. [21] and [22]. This type of model worked well not only for Provan, Cunningham, and Cook, but also for Wang in Refs. [18] and [20]. The AAR predictions from this model for the hour starting one hour from the current time through the hour starting ten hours from the current time are provided to the GDP models (ten features).

F. Reroutes

Reroute advisories were collected from the FAA's National Traffic Management Log (NTML) and used to construct features describing reroutes required or recommended for flights bound to the airport. GDPs are sometimes used to help address reductions in the capacity of airspace typically used by flights headed to the airport, such as airspace near arrival fixes. This sort of capacity reduction also may result in reroutes, so these features attempt to quantify reductions in airspace capacity that may cause GDP implementation. More precisely, for the hour starting at the current time through the hour starting 10 hours from the current time, there are 4 reroute-related features (44 total). Two of these report the number of departure Air Route Traffic Control Centers (ARTCCs or Centers) and departure airports for which reroutes are recommended for flights bound to the airport. The other two features report the number of departure Centers and departure airports for which reroutes are required for flights bound to the airport. These features by no means completely describe congestion in relevant airspace, however, and additional features could be added to provide describe this more fully.

G. Previous GDP Plan

GDP plan data was also retrieved from the FAA's NTML. While GDP plans are sometimes modified, selected GDP actions often are a continuation of a previously-announced plan. For each hour, we quantify the GDP action that would be pursued at the start of the hour, assuming the GDP plan as specified one hour ago is simply continued. The features describing the previous plan include whether or not there would be a GDP implemented, the GDP scope, the number of hours until the first hour of controlled GDP rates, the number of hours until the last hour of controlled GDP rates, and the GDP rate that should be in place for each relevant hour of the GDP (from the hour starting at the current time through the hour starting 10 hours from the current time). This means that there are 15 features describing the previous GDP plan for each hourly time step.

H. Ground and Air Buffers

One of the fundamental purposes of a GDP is to prescribe ground delay, which is cheaper than delay absorbed in the air, when some delay must be absorbed. Therefore, we defined a simple

deterministic queuing network model with buffers for flights delayed on the ground and in the air. It provides estimates of two quantities that are essential to GDP planning: how much delay is absorbed on the ground and in the air as a result of GDP actions. This model is similar to the deterministic queuing model utilized by Kim and Hansen in Ref. [24], the stochastic queuing model proposed by Odoni and discussed by Ball et al. in sub-section 4.3 of Ref. [25], and the model used in a stochastic ground-holding problem specified by Ball et al. in Ref. [26]. Although this model certainly fails to capture several relevant aspects of GDPs (such as differences in flight times between flights), our hope is that it is simple but not simplistic: that it captures enough of the relevant characteristics of the real world to be useful for GDP analytics, but without the burden of unnecessary complexity.

The number of flights in the ground buffer at the start of time step t is b_t^{g} and the number in the air buffer is b_t^{a}. At the start of each day, the ground and air buffers are initialized to zero. Then, the system dynamics specify that the buffer levels at the start of the next time step (b_{t+1}^{g} and b_{t+1}^{a}) depend on the scheduled arrivals ($\mathrm{scheduled}_t$) and AAR (AAR_t) during time step t, as well as b_t^{g}, b_t^{a}, and the GDP action implemented in this time step. If a GDP was implemented during t (GDP_t), then the relevant components of the GDP action are whether or not the arrival rate is controlled by the GDP during t ($\mathrm{controlled}_t$) and the controlled arrival rate during t (rate_t). The buffer levels are updated according to

$$b_{t+1}^{\mathrm{g}} = \begin{cases} [b_t^{\mathrm{g}} + \mathrm{scheduled}_t - \mathrm{rate}_t]_+ & \text{if } \mathrm{GDP}_t \text{ and } \mathrm{controlled}_t \\ 0 & \text{else} \end{cases} \tag{1}$$

and

$$b_{t+1}^{\mathrm{a}} = \begin{cases} [b_t^{\mathrm{a}} + \min(b_t^{\mathrm{g}} + \mathrm{scheduled}_t, \mathrm{rate}_t) - \mathrm{AAR}_t]_+ & \text{if } \mathrm{GDP}_t \text{ and } \mathrm{controlled}_t \\ [b_t^{\mathrm{a}} + b_t^{\mathrm{g}} + \mathrm{scheduled}_t - \mathrm{AAR}_t]_+ & \text{else.} \end{cases} \tag{2}$$

Here $[x]_+$ is equal to x if $x \geq 0$ but equal to 0 if $x < 0$. Fig. 1 depicts this simple queuing model when a GDP is implemented.

III. GDP Models

The structure we specified and used for both BC and IRL GDP models is depicted in Fig. 2. The input data that can be used by the models is described in Section II. The output of the model is

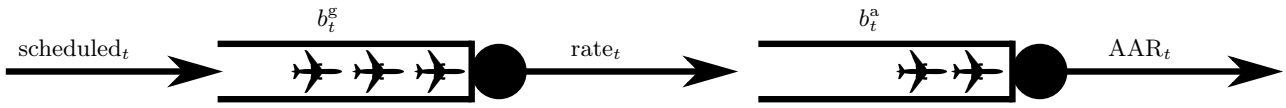


Fig. 1 Ground and air buffer system model.

either a prediction that a GDP would not be implemented in the state quantified by the input data, or a prediction that a GDP would be implemented, along with predictions of the GDP parameters that would be used in the state quantified by the input data. The GDP parameters include the GDP scope, the time when the controlled rates in the GDP begin (the GDP *start time*), the number of hours of enforced GDP rates (the GDP *duration*), and the enforced rate for each hour in the GDP duration, extending from the hour starting at the current time through the hour starting 10 hours in the future (up to 11 hourly rates). There are two sub-models: one that predicts whether or not a GDP will be implemented and another that predicts GDP parameters when a GDP is initialized. The model is a simplification of reality in that it requires that a GDP plan either progress as planned or be canceled (no modifications or extensions are permitted).

A. GDP Implemented Models

A GDP implemented model predicts whether or not a GDP will be implemented during a given hour, given a set of features describing the state (see Section II). BC and IRL GDP implemented models will be developed and compared. The specific BC and IRL algorithms will be discussed in the next two sub-sections.

1. BC: Random Forest Classifiers for Cancellation and Initialization

The structure of the BC model developed and analyzed for this research is depicted in Fig. 3. Depending on whether or not the previous GDP plan specified that a GDP would be implemented in the current hour, either a GDP cancellation or GDP initialization model is invoked. We hope that creating specific models for what seem to be different decisions (GDP cancellation and GDP initialization) will lead to better predictive performance, as well as to more refined insight into

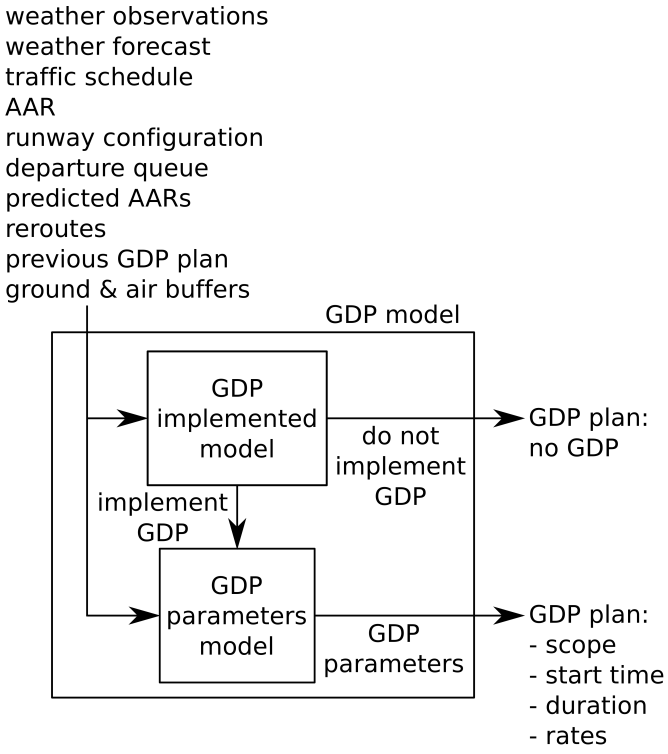


Fig. 2 Structure of GDP model.

GDP decision making. An additional motivation for building separate models for GDP cancellation and GDP initialization is that doing so facilitates the use of over- and under-sampling to generate custom training data sets for each model, which helps each with the difficult imbalanced classification problem it faces.

The GDP initialization model is provided with features describing the current state at the airport, but no features describing the previous GDP plan because the previous plan was to not use a GDP. It then predicts either that a GDP will be initialized or that no GDP will be initialized. The GDP cancellation model is provided with features describing not only the current state at the airport, but also features describing the previous GDP plan, such as the scope, planned end time, and rates. It then predicts either that the GDP will continue as planned or that it will be canceled. In hours that are immediately after the planned end of a GDP, the GDP cancellation model is *not* used because not implementing a GDP in that hour is *not* a cancellation. The question in those hours is whether or not a new GDP will be initialized, so the GDP initialization model is invoked.

The GDP cancellation and GDP initialization models are both random forest classification

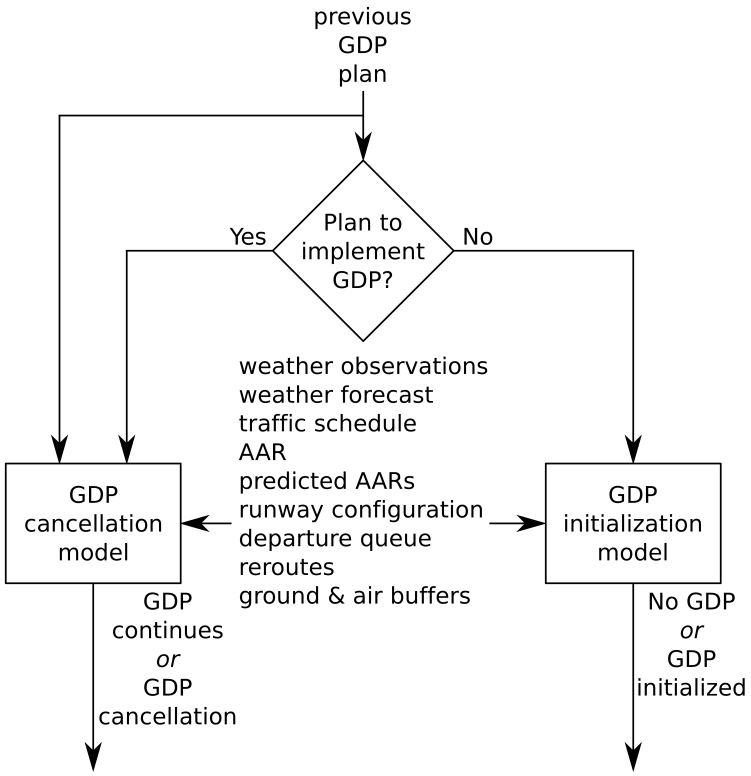


Fig. 3 Structure of the BC GDP implemented model.

models, implemented with the `RandomForestClassifier` class available in the scikit-learn package for the Python programming language [27]. Random forest models were selected because they typically perform well with minimal tuning of the algorithms that train the models, they generally do not over-fit to the training data even when provided with many features, and other researchers have found that related models predict AARs better than alternative models [18, 20, 21, 28].

Among all of the hours where a GDP is planned to be implemented, only in about one in ten hours is a GDP canceled in the data we are analyzing. Similarly, among all the hours where a GDP is not planned to be implemented, only in about one in thirty hours is one initialized. When facing imbalanced data sets such as these, predictive performance can sometimes be enhanced by using the Synthetic Minority Over-Sampling Technique (SMOTE) [29]. We used an implementation of the SMOTE algorithm to generate synthetic minority class (initialization and cancellation) data points [30], and we also under-sampled the majority class data points. In particular, the number of minority samples was doubled by using SMOTE to generate synthetic samples, and the majority

samples were under-sampled so that there were four times as many (for EWR) or twice as many (for SFO) majority-class samples as (real and synthetic) minority-class samples.

2. IRL: Cascaded Supervised Inverse Reinforcement Learning

The IRL algorithm selected for evaluation is the Cascaded Supervised Inverse Reinforcement Learning (CSI) algorithm proposed by Klein, Piot, Geist, and Pietquin in Ref. [31]. This approach was selected for several reasons: it is relatively easy to implement, it does not require multiple computations of an optimal policy for various possible reward functions, and it leverages existing classification and regression algorithms. Furthermore, it does not involve exploring the entire state space—just the states visited in the training data and possibly in some additional simulations. Finally, the expert policy that produced the demonstrations in the data is near-optimal for the reward function estimated by the CSI algorithm (see Ref. [31], Theorem 1). One downside of the algorithm is that it assumes a deterministic expert policy that is optimal for a certain reward function, which may not be the case for GDP decision making. This drawback did not prevent us from selecting the CSI algorithm, however, because for this initial investigation we are working with deterministic GDP implemented models.

There are six main elements involved in CSI. Each of these elements will be described in the subsequent paragraphs.

System Model: The CSI algorithm, like any other IRL algorithm, requires a system model. The system state at a time step t is s_t, which is a member of a set of all possible states $\mathcal{S}$. It involves an *exogenous* state and a *controlled* state. The exogenous state is a vector of features describing the weather observations, the weather forecast, the traffic schedule, the current AAR, the runway configuration, predicted AARs, and reroutes (see Section II). GDP actions have no impact on the exogenous state dynamics in this model, and the CSI algorithm allows us to proceed even though we assume that the state transition probabilities for this part of the state are not specified. The controlled state contains the GDP action for this time step prescribed by the previous GDP plan, b_t^{a}, and b_t^{g}. The controlled state dynamics are impacted by GDP actions as prescribed by Eqs. (1) and (2). The action a_t taken at t is binary: it specifies either that no GDP is implemented or that

a GDP is implemented in this time step ($a_t \in \mathcal{A} = \{\text{GDP}, \text{no GDP}\}$). This is a Markov model because the conditional distribution of future states depends only on the current state and action, not on the whole history of states and actions.

Decision Process Model: We assume that traffic managers are attempting to maximize the expected value of a discounted infinite sum of future rewards. More precisely, they seek a deterministic policy mapping states to actions that maximizes $\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$, where $\gamma \in (0, 1)$ is the discount factor and $R : \mathcal{S} \times \mathcal{A} \to \mathbf{R}$ is the reward for each time step.

Classifier Policy and Estimation of Optimal State-Action Value: Once the system and decision process model have been specified, the next step in CSI is to derive a deterministic classifier policy $\pi_{\text{C}} : \mathcal{S} \to \mathcal{A}$. This classifier policy can be any *score function-based multi-class classifier* (SFMC^2). An SFMC^2 predicts an action that achieves the largest score according to some score function $q(s, a)$: $\pi_{\text{C}}(s) \in \text{argmax}_{a \in \mathcal{A}}\, q(s, a)$. By interpreting the score function $q(s, a)$ as an optimal *state-action value* function for π_{C} and making use of the Bellman equation, the CSI algorithm can use the score function and the policy to generate a reward sample data point corresponding to each state-action pair. We will utilize a model similar to the random forest BC GDP implemented model described in sub-section III.A.1 for the classifier policy π_{C}. The classifier policy differs in that it uses a single model to predict GDP implementation, not separate models for initialization and cancellation. Furthermore, we train it after using SMOTE to double the number of time steps in which a GDP is implemented, but we do not under-sample the majority class (non-GDP time steps). Finally, this model does *not* utilize features describing the previous GDP plan. The score function for a given state and action is the average over all the trees in the random forest of the fraction of members in the leaf nodes for which the action was selected.

Reward Estimation with Regression: The next step in CSI is to use a regression algorithm to estimate a reward function R_{C} that is consistent with the reward samples produced by using the Bellman equation and the score function q when it is interpreted as an optimal state-action value function for π_{C} [31]. Any type of regressor can be used, but we use a linear regressor model (as implemented in the `OLS` class of the statmodels Python package). Linear models are relatively easy to interpret, so a trained linear regressor model should be relatively rich in insight. Furthermore,

some proposed TFM and GDP optimization approaches assume or even require a reward function that is linear in the optimization decision variables (Refs. [2] and [25] describe examples of such approaches).

The form of a linear regressor model is $\hat{R}_{\mathrm{C}}(s,a) = \hat{\theta}^{\top} f(s,a)$, where $f(s,a) \in \mathbf{R}^{M}$ is a vector of reward features for the state-action pair (s,a) and $\hat{\theta} \in \mathbf{R}^{M}$ is a vector of parameters estimated by the regression algorithm. The reward features we utilize were inspired by the objective function used in the Ground Delay Program Parameter Selection Model [13] and also by performance measures suggested by Ball et al. [25] and Liu and Hansen [4]. The eight reward features are ground and air buffer levels at the end of the current time step, the change in the air and ground buffers during this time step, an indicator that the air buffer will be greater than or equal to five at the end of the current time step, the number of arrivals during the time step, the number of unused arrival slots while the arrival rate is controlled by a GDP (i.e. while controlled$_t$ is true), and the canceled duration of a canceled GDP.

Derivation of Approximately-Optimal Policy: We derive a policy $\hat{\pi}^{\star}_{\hat{R}_{\mathrm{C}}}$ that attempts to maximize the infinite discounted total reward objective, defined based on $\hat{R}_{\mathrm{C}}$, using an approximate dynamic programming approach known as *rollouts* (see Ref. [32], sub-section 6.4). Ultimately, the prediction produced by the CSI model of whether or not a GDP will be implemented when in state s will be the action returned by $\hat{\pi}^{\star}_{\hat{R}_{\mathrm{C}}}(s)$. Other techniques from reinforcement learning and approximate dynamic programming could be used to derive this policy. We selected rollouts because we could use weather forecast and traffic schedule data already in the system state to perform the simulations required when estimating optimal state-action value functions, which made the problem tractable in spite of the unknown state transition probabilities.

B. GDP Parameters Model

As depicted in Fig. 2, a GDP parameters model was developed to predict the GDP scope, GDP start time, GDP duration, and GDP rates. This model is only used in the final policy estimation step of the CSI algorithm, described in sub-section III.A.2. Random forest BC models were used for each sub-model that predicts one of the GDP parameters, and the predictions were rounded

to the nearest typical value for the parameter in question. Random forest models were selected because they perform well with relatively little tuning of the algorithm that trains the models, they generally do not over-fit to the training data even when provided with many features, and other researchers have had some success using related models to predict AARs [18, 20, 21, 28]. Random forest regression models for these parameters were trained with the `RandomForestRegressor` class in the scikit-learn Python package [27] using settings and parameters similar to those suggested in Ref. [21]. Although these models are not the focus of this research, we are not aware of previous attempts at building models that predict GDP parameters.

IV. Experiments

Ten-fold cross validation is used in these experiments (see Ref. [28], sub-section 7.10.1). The folds are defined based on days in the data set rather than individual time steps (hours). With 455 days in the data set and ten folds, each fold consisted of all of the time steps in around 46 days. The experiments focus on the GDP implemented models.

A. Parametric Studies Related to Look-ahead

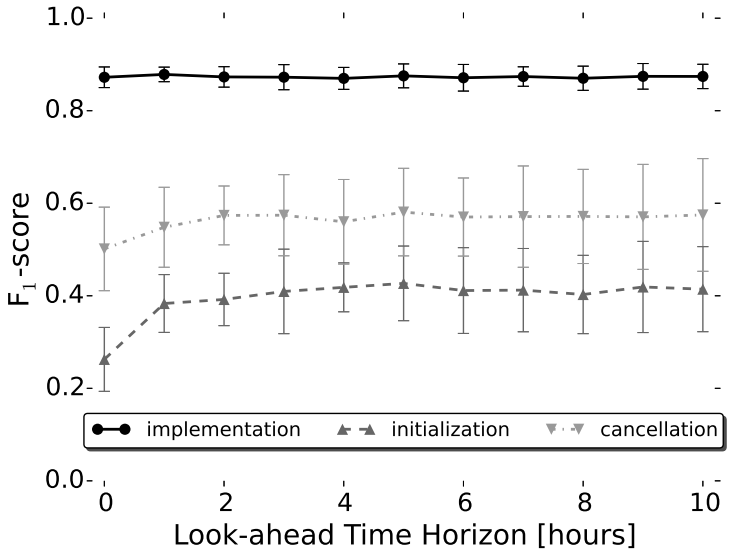
Characteristics of both BC and IRL models provide insight into the degree to which decision makers are looking ahead into the future when selecting actions. Both models are provided with features that quantify predictions of future weather conditions, future scheduled traffic, and predictions of future airport arrival rates. The data set includes predictions extending ten hours into the future, but if decision makers only consider predictions extending say five hours into the future, then providing the models with the remaining predictions will not improve and may even harm the quality of model predictions. Therefore, we conducted a parametric study to determine the impact on GDP implementation prediction quality of changing how far into the future these predictions extend (a quantity we refer to as the *look-ahead time horizon*). Furthermore, for the IRL model, the discount factor γ quantifies the relative importance of rewards accrued now and rewards accrued in the future. Therefore, we also conducted a parametric study to determine the impact of changing γ on reward regressor prediction quality.

1. BC: Look-ahead Time Horizon Parametric Analysis

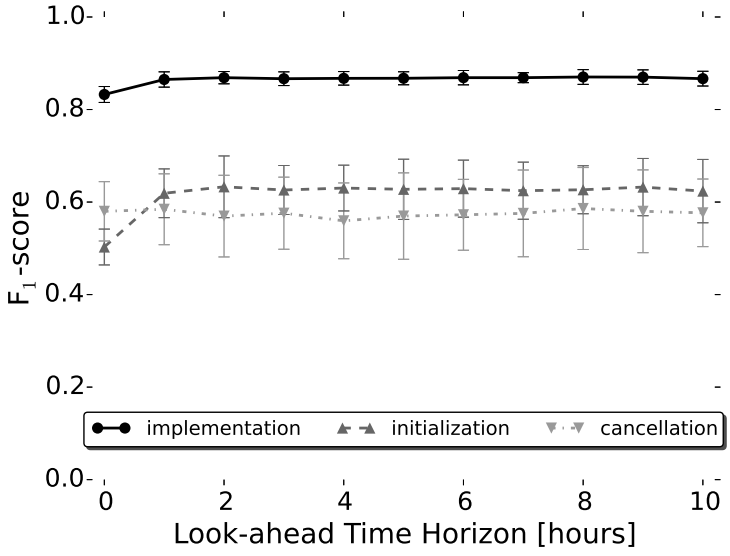
A parametric analysis was performed to determine the impact on the predictive performance of the BC GDP implemented model of changing the look-ahead time horizon. That is, we changed the number of features provided to the model by removing features describing predictions beyond a certain time horizon in the future, and we recorded the performance of the model while varying this time horizon between zero and ten hours. To evaluate predictive performance, we used the F_1-score because it is an appropriate metric for imbalanced data sets. It ranges between zero and one, with larger values indicating better performance.

Figure 4 shows the means (dots) and standard deviations (bars) of the F_1-scores achieved on the ten test data folds for various look-ahead time horizons. For each airport, the GDP implementation curve is flat, indicating that the quality of these overall predictions do not change much with look-ahead time horizon. Predictions of GDP initialization and cancellation at EWR improve as the horizon increases to four and two hours, respectively, which supports the claim that decision makers use predictions extending at least that far into the future when determining whether to initialize or cancel a GDP. For SFO, on the other hand, the characteristics of the quality of predictions as the look-ahead time horizon increases are consistent with the claim that decision makers only use predictions extending to the next hour when deciding whether or not to initialize a GDP, and that they use no predictions at all when deciding whether to cancel a GDP. Given that initial GDP plans can specify GDPs extending ten or more hours in the future, and that GDPs can cause ground delays now for flights that will not arrive for four or more hours, it is somewhat surprising that these results are consistent with relatively tactical GDP implementation decision making. However, there are several possible explanations for such tactical decision making. The dynamics of traffic demand and weather conditions are stochastic, and so predictions at longer look-ahead time horizons may be subject to too much uncertainty to be helpful for decision makers. Furthermore, GDP implementation decisions can always be changed later, so decision makers may not always need to look very far into the future. A related explanation is that GDP parameters can be revised. For example, GDP rates can be increased or decreased, and the end time can be extended. This gives decision makers the ability to adjust a GDP to better fit unforeseen conditions, and may therefore

reduce the need to look far into the future when determining whether to implement a GDP. Finally, when there is sufficient demand originating from nearby airports, GDPs may be able to effectively eliminate airborne delay that is expected just a couple of hours in the future.



(a) EWR.



(b) SFO.

Fig. 4 Prediction quality of BC GDP implemented model as look-ahead time horizon changes.

For the results in the remainder of this article, we use a look-ahead time horizon of four hours. Prediction quality does not meaningfully change for any of the models with longer look-ahead time

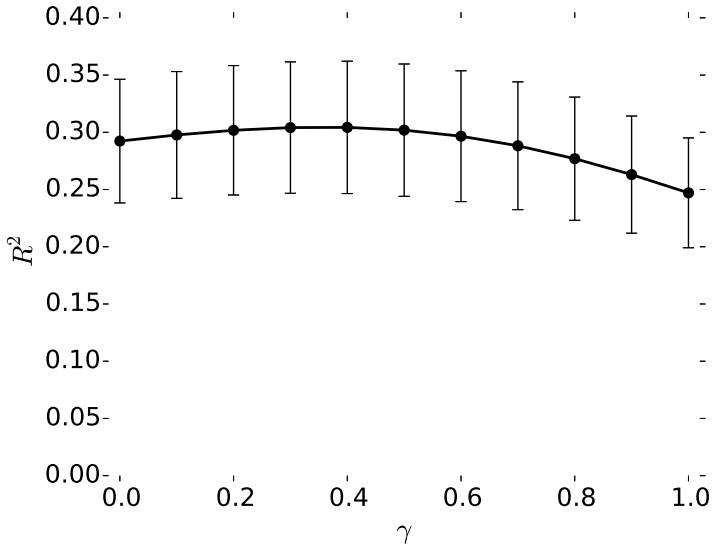
horizons, so using this horizon should lead to more succinct models without sacrificing prediction quality. With a four-hour look-ahead time horizon, the models are given 125 features describing the state.

2. IRL: Parametric Investigation of Discount Factor

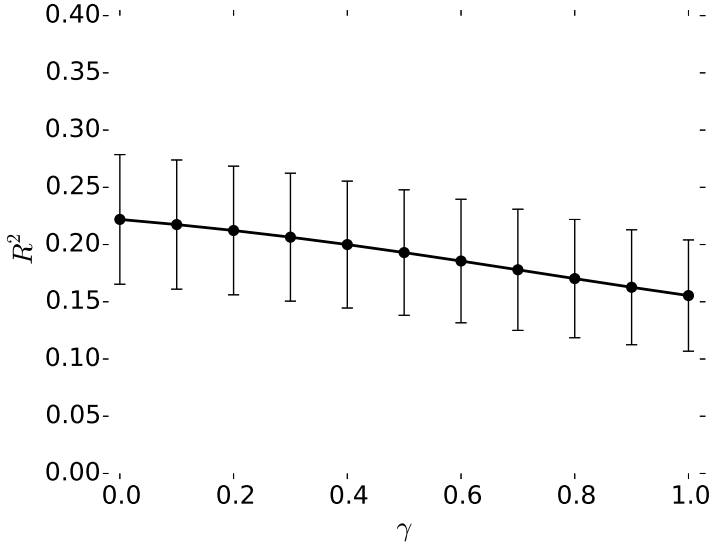
We also performed a parametric analysis to determine the impact of changing γ on the predictive performance of the reward regressor $\hat{R}_C$. More precisely, we investigated 11 possible values of γ ranging from 0.0 to 1.0. A γ of 0.0 means that the objective does not consider rewards in future time steps at all, while a γ of 1.0 means that the objective places equal weight on rewards earned in any time step—even time steps far in the future. For each possible γ, we constructed a corresponding set of reward samples and then trained a corresponding reward regressor. Ten-fold cross validation was used to do this repeatedly, and the means (dots) and standard deviations (bars) of the R^2 values achieved on the test data sets by each regressor are plotted in Fig. 5. We used R^2 to quantify predictive performance because changing γ changes the magnitude and distribution of the truth data. The R^2 metric involves normalization in a way that permits comparisons of predictive performance as changes in γ change the truth data distribution.

The R^2 values in Fig. 5 range between around 0.15 and 0.30, which indicates that the regressor achieves a poor fit of the data. This may be a cause of the relatively poor predictive performance of the IRL model (see sub-section IV.B). For EWR, the R^2 increases as γ increases from 0.0, peaks at 0.4, and then decreases as γ increases further. For SFO, the R^2 decreases monotonically as γ increases. These results also suggest that GDP implementation decision making is more tactical than strategic: it is not concerned much with rewards earned after the current time step. As was the case when we investigated the impact of look-ahead time horizon in sub-section IV.A.1, the results suggest that decision makers are more concerned with future conditions at EWR than at SFO.

For the remainder of this article, we will use $\gamma = 0.4$ for the EWR IRL model because that value leads to the largest average R^2 value on the test data sets. For the SFO IRL model, we use $\gamma = 0.2$ even though a better fit is achieved at $\gamma = 0.0$. We select this γ for SFO because using $\gamma = 0.0$ would make the IRL model simply a complicated BC model, and we want to compare an IRL model



(a) EWR.



(b) SFO.

Fig. 5 Fit of reward regressor as γ changes.

to a BC model. Furthermore, the average R^2 at $\gamma = 0.2$ is close to that at $\gamma = 0.0$, and well within the one-standard-deviation range, so we are not sacrificing much predictive performance.

B. Prediction Quality Results

The prediction quality metrics for GDP implemented models are computed based on three confusion matrices that can be constructed with the predictions for the testing data. The three matrices investigated here are constructed with the data in the ten test data folds. Each hour-long time step in the full data set is in a testing data fold exactly once, so the testing data contains each sample from the full data set exactly once. The first confusion matrix describes predictions of GDP implementation, the second describes predictions of GDP initialization, and the third describes predictions of GDP cancellation. The first matrix involves all the testing data points, the second and third matrices are based on only some of the testing data points, and each data point is involved in either the second or the third matrix but not both. For each confusion matrix, the accuracy, precision, recall, and F_1-score will be reported. Each of these metrics will be in the range $[0, 1]$, with larger values indicating better predictive performance. Since precision and recall are particularly relevant for imbalanced data sets, such as those faced by the initialization and cancellation models, and since the F_1-score is the harmonic mean of these two metrics, we view the F_1-score as the most important single metric that quantifies the predictive performance for each confusion matrix.

1. Baseline: Quality of GDP Plan Model Predictions

We will summarize the quality of predictions produced by a baseline model that simply predicts that the previous GDP plan will be executed. If no GDP is planned for the hour in question, the model predicts that no GDP will be initialized, and if a GDP is planned, then the model predicts that it will continue and not be canceled. The F_1-scores achieved by this model are 0.90 and 0.87 for GDP implementation at EWR and SFO, respectively, which is suggestive of high-quality predictions. However, this model achieves a recall of 0.00 and undefined precision and F_1-score metrics for predictions of initializations and cancellations at both airports. Initialization and cancellation events are important operationally, so this model's failure to predict these events reveals its limited operational value. Furthermore, this model's performance illustrates the importance of evaluating models based on their predictions of initializations and cancellations, not just their predictions of GDP implementation.

2. Quality of BC GDP Implemented Model Predictions

Tables 1 and 2 show the three confusion matrices and related metrics achieved by the BC GDP implemented models for EWR and SFO, respectively, when they are presented with the ten test data folds. For both EWR and SFO, the accuracy, precision, recall, and F_1-score of the overall GDP implemented models are relatively close to one. However, this strong overall performance masks how much the GDP initialization and GDP cancellation models struggle to predict the relatively infrequent initialization and cancellation events. Although the F_1-scores for predictions of GDP initialization and cancellation are slightly higher for SFO than for EWR, they all range between 0.41 and 0.64, which is low. For both EWR and to a lesser extent SFO, the low precision scores achieved by the GDP initialization models indicate a high false alarm rate—that they often predict that GDPs will be initialized when they are not. The GDP initialization data set is particularly imbalanced, with 34 and 26 times more non-initialization events than initialization events for EWR and SFO, respectively. This makes predicting initializations difficult. The GDP cancellation models also suffer from low precision scores, and again this is partially a result of the imbalanced nature of the data set. There are 11 and 7 times more non-cancellation events than cancellation events for EWR and SFO, respectively.

3. Quality of IRL GDP Implemented Model Predictions

Tables 3 and 4 show the three confusion matrices and related metrics achieved by the IRL GDP implemented models for EWR and SFO, respectively, when they are presented with the ten test data folds. For both EWR and SFO, the IRL GDP implemented models demonstrate substantially lower predictive performance than the BC GDP implemented models. This is most evident for predictions of GDP initialization: the EWR IRL model predicts initialization much too frequently while the SFO IRL model predicts initialization too infrequently. The SFO IRL model also predicts GDP cancellation too infrequently.

Although it is difficult to determine exactly what causes the relatively poor predictive performance of the IRL GDP implemented models, the results of diagnostic tests and analysis of the estimated regressor parameters (see sub-section IV.C.2) suggest that the poor predictive performance of

Table 1 Confusion matrices for EWR BC GDP implemented model

(a) *Implementation:* accuracy=0.94, precision=0.84, recall=0.92, F_1-score=0.88

Actual	**Predicted No GDP**	**Predicted GDP**	**Total**
No GDP	6810	348	7158
GDP	154	1788	1942
Total	6964	2136	9100

(b) *Initialization:* accuracy=0.95, precision=0.31, recall=0.60, F_1-score=0.41

Actual	**Predicted No Initialization**	**Predicted Initialization**	**Total**
No Initialization	6720	273	6993
Initialization	84	124	208
Total	6804	397	7201

(c) *Cancellation:* accuracy=0.92, precision=0.56, recall=0.55, F_1-score=0.55

Actual	**Predicted Continuation**	**Predicted Cancellation**	**Total**
Continuation	1664	70	1734
Cancellation	75	90	165
Total	1739	160	1899

the reward regressor is likely an important cause [33]. The reward regressor's poor performance may in turn be the result of selecting a reward regressor model form that is not suited to the regression problem or of using an inaccurate or incomplete set of reward features. Several reward features are derived from the state of the ground and air buffer system model, and our intuition is that the fidelity of this system model may need to be improved. Finally, the IRL and BC models are based on different assumptions about decision making, and this could help explain the difference in predictive performance. In particular, GDP implementation decision making may be more of a reaction to the current state that does not explicitly consider the impact of actions on future states, as assumed by BC approaches, than an attempt to deterministically and strategically achieve rewards accrued over time, as assumed by the CSI IRL algorithm. The results of our analysis do not provide any insight into the reasons that GDP implementation decision making might exhibit this characteristic, but it could be related to training, procedures, or system uncertainties.

Table 2 Confusion matrices for SFO BC GDP implemented model

(a) *Implementation:* accuracy=0.95, precision=0.85, recall=0.90, F_1-score=0.87

Actual	Predicted No GDP	Predicted GDP	Total
No GDP	7625	337	7962
GDP	206	1842	2048
Total	7831	2179	10010

(b) *Initialization:* accuracy=0.96, precision=0.50, recall=0.87, F_1-score=0.64

Actual	Predicted No Initialization	Predicted Initialization	Total
No Initialization	7456	257	7713
Initialization	39	258	297
Total	7495	515	8010

(c) *Cancellation:* accuracy=0.88, precision=0.50, recall=0.68, F_1-score=0.58

Actual	Predicted Continuation	Predicted Cancellation	Total
Continuation	1584	167	1751
Cancellation	80	169	249
Total	1664	336	2000

C. Insight Results

1. Insight from BC GDP Implemented Model

The main form of insight available from the random forest models used in the BC GDP implemented model is feature *importance scores*. The importance score for a feature is the total decrease in node "impurity" (as measured by the Gini splitting criterion) resulting from splits defined based on the feature and weighted by the proportion of samples reaching the corresponding nodes, averaged over all the trees in the ensemble. According to this definition, important features define frequently-used splits and/or generate large improvements in the split criterion when they are used to define splits. Larger scores imply greater importance. We recorded the importance scores of the input features used by the models for each of the ten times the models are trained.

Figures 6 and 7 show the scores for the features with the ten highest importance scores for initialization and cancellation models that make up the BC GDP implemented models for EWR

Table 3 Confusion matrices for EWR IRL GDP implemented model

(a) *Implementation:* accuracy=0.81, precision=0.53, recall=0.92, F_1-score=0.67

Actual	**Predicted No GDP**	**Predicted GDP**	**Total**
No GDP	5554	1604	7158
GDP	159	1783	1942
Total	5713	3387	9100

(b) *Initialization:* accuracy=0.78, precision=0.085, recall=0.67, F_1-score=0.15

Actual	**Predicted No Initialization**	**Predicted Initialization**	**Total**
No Initialization	5491	1502	6993
Initialization	69	139	208
Total	5560	1641	7201

(c) *Cancellation:* accuracy=0.90, precision=0.41, recall=0.38, F_1-score=0.40

Actual	**Predicted Continuation**	**Predicted Cancellation**	**Total**
Continuation	1644	90	1734
Cancellation	102	63	165
Total	1746	153	1899

and SFO, respectively. The height of each bar is the mean of the importance scores for each feature according to the ten models constructed for the ten training data sets used in cross validation, and the error bars show the standard deviation of these ten importance scores. Features names ending with a "_k" for some integer k are predictions of the feature for the hour starting k hours from the time of the prediction.

For the EWR GDP initialization model, features related to the AAR ("AAR") or a prediction of the AAR ("Pred_AAR") make up five of these ten features. This suggests that future AAR levels have a relatively strong influence on GDP initialization, which makes sense given that GDPs are used when expected arrival demand exceeds expected arrival capacity and that arrival capacity is quantified by AAR. Features related to scheduled arrivals ("SCHARR") account for another four of the ten features, which is also not surprising because GDPs are used to reduce arrivals when predictions suggest that there might be an excessive number of arrivals. The tenth feature in this

Table 4 Confusion matrices for SFO IRL GDP implemented model

(a) *Implementation:* accuracy=0.94, precision=0.85, recall=0.84, F_1-score=0.84

Actual	Predicted No GDP	Predicted GDP	Total
No GDP	7659	303	7962
GDP	337	1711	2048
Total	7996	2014	10010

(b) *Initialization:* accuracy=0.95, precision=0.17, recall=0.081, F_1-score=0.11

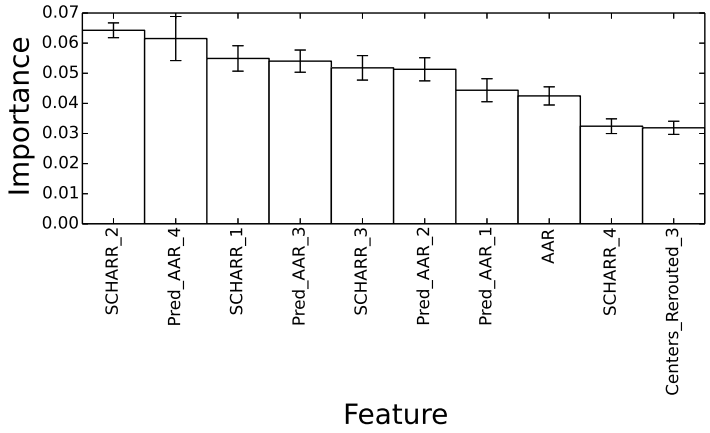
Actual	Predicted No Initialization	Predicted Initialization	Total
No Initialization	7595	118	7713
Initialization	273	24	297
Total	7868	142	8010

(c) *Cancellation:* accuracy=0.88, precision=0.50, recall=0.26, F_1-score=0.34

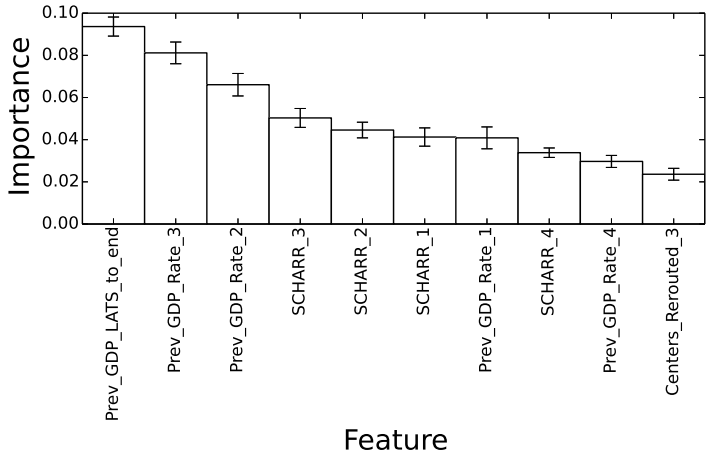
Actual	Predicted Continuation	Predicted Cancellation	Total
Continuation	1687	64	1751
Cancellation	185	64	249
Total	1872	128	2000

set is a prediction of the number of departure Centers for which reroutes are required three hours in the future ("Centers_Rerouted_3"), suggesting that GDPs are initialized at EWR to help reduce demand for constrained airspace. This makes sense because flights bound for EWR typically traverse highly congested airspace in the northeastern United States.

For the EWR GDP cancellation model, five features related to parameters of the previous GDP plan, such as the planned time until the end of the GDP ("Prev_GDP_LATS_to_end") and planned GDP rates ("Prev_GDP_Rate"), achieve high average importance scores. This suggests that GDPs are unlikely to be canceled when the previous plan indicates that they are far from finishing, which makes sense because stakeholders may value predictable GDP plans and so presumably experts attempt to select appropriate GDP end times [4]. The planned rates may be important because they are another way to learn the planned remaining duration, or perhaps because they may indicate the degree to which capacity is diminished (GDPs might be less likely to



(a) GDP initialization model.

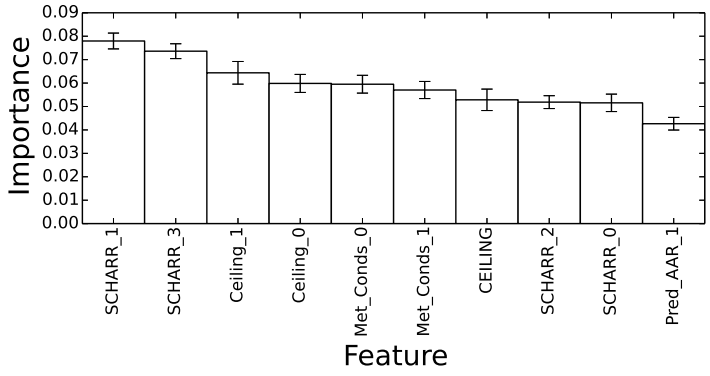


(b) GDP cancellation model.

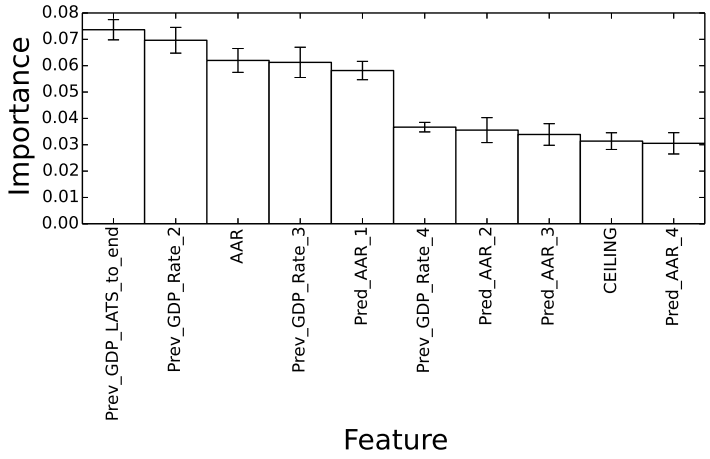
Fig. 6 Features with highest importance scores for the EWR BC GDP implemented model.

be canceled when capacity is lower). Features related to scheduled arrivals ("SCHARR") account for four of the remaining features achieving the ten highest importance scores. This makes sense because if scheduled arrivals are low, a GDP may no longer be needed. The final feature in this set is a reroute-related feature ("Centers_Rerouted_2"), again suggesting that EWR GDPs might be partially caused by congested airspace that also requires new routes for flights bound for EWR.

Of the ten features with the largest average importance scores for the SFO GDP initialization model, five are related to observations or forecasts valid at some hour in the upcoming two hours of the ceiling or meteorological conditions ("CEILING", "Ceiling_0", "Ceiling_1", "Met_Conds_0", and "Met_Conds_1"). "CEILING" is the observation of the ceiling at the current hour recorded



(a) GDP initialization model.



(b) GDP cancellation model.

Fig. 7 Features with highest importance scores for the SFO BC GDP implemented model.

in a METAR report. “Ceiling_0" is the prediction of the ceiling from a TAF forecast that is valid for the current hour. The TAF forecast may have been published very recently, or it may have been published up to three hours ago. This dependence on ceilings or meteorological conditions is not surprising given that SFO GDPs are largely caused by low ceilings [1]. Only one feature in this set is an AAR (“Pred_AAR_1"), which is fewer than the five such features for the EWR GDP initialization model. This might be because while the weather conditions that lead to lower capacity at SFO are relatively straightforward to describe and quantify, enabling the SFO model to identify and depend upon them directly, the capacity at EWR is more a more complicated function of a variety of weather and other conditions, leading the EWR model to depend on AAR predictions

(rather than directly on weather conditions). Features related to scheduled arrivals make up the remaining four features, which is not surprising for reasons described earlier.

Finally, for the SFO GDP cancellation model, four of the ten features achieving the largest average importance scores come from the previous GDP plan (such as "Prev_GDP_LATS_to_end" and planned GDP rates). This makes sense for the reasons described in our discussion of the EWR cancellation model importance scores. Predicted or current AARs make up five other of these features, and the current ceiling is the final feature in the set.

2. Insight from IRL GDP Implemented Model

Although our evaluation suggests that the predictive power of the IRL GDP implemented model is low, the estimated parameters $\hat{\theta}$ for the reward function regressor $\hat{R}_{\mathrm{C}}(s,a) = \hat{\theta}^{\top} f(s,a)$ still may provide some useful insights. These insights should be viewed with suspicion, however, because $\hat{R}_{\mathrm{C}}$ struggled to fit the reward sample training data. In the ten testing data sets used in cross validation of the reward regressor, the average R^2 value achieved by $\hat{R}_{\mathrm{C}}$ was only 0.30 for EWR and 0.21 for SFO (see sub-section IV.A.2). These low values suggest that $\hat{R}_{\mathrm{C}}$ is not explaining much of the variation of the reward samples in the training data set. We suspect that this poor performance is an important cause of the poor predictive power of the IRL GDP implemented model.

Table 5 shows average reward parameter estimates and corresponding average p-values for the ten regressors trained with the ten training data folds used in cross validation. Before training $\hat{R}_{\mathrm{C}}$, scalar-valued reward features were standardized over the time steps for which they are defined, while the lone indicator feature (which takes a value of one when the air buffer at the end of the time step is greater than or equal to five and zero otherwise) was not standardized. This facilitates interpretation of the parameter estimates by making comparisons of their relative magnitudes more meaningful. When constructing reward samples to train $\hat{R}_{\mathrm{C}}$, we used $\gamma = 0.4$ for EWR and $\gamma = 0.2$ for SFO. Therefore, the ranges of possible and typical reward sample values differ between the two airports, and so we cannot directly compare the values of the estimated parameters between airports.

The reward regressors for the two airports are remarkably similar. This similarity is consistent

Table 5 Properties of parameters of $\hat{R}_C$ for EWR and SFO

	EWR		SFO	
Reward Feature ($f_m(s,a)$)	$\hat{\theta}_m$	p-**value**	$\hat{\theta}_m$	p-**value**
constant	0.53	< 0.001	0.70	< 0.001
ground buffer at end of time step	0.0040	0.14	-0.022	< 0.001
change in ground buffer	0.0078	0.14	0.0070	0.19
air buffer at end of time step	0.015	< 0.001	-0.0035	0.40
change in air buffer	-0.035	< 0.001	-0.027	< 0.001
air buffer at end of time step ≥ 5	-0.24	< 0.001	-0.19	< 0.001
arrivals	-0.029	< 0.001	0.0056	0.076
unused slots during rate control	-0.15	< 0.001	-0.21	< 0.001
duration of GDP canceled	-0.041	0.040	-0.050	0.039

with the conjecture that while different phenomena cause congestion issues leading to GDPs at these two airports (evidenced by the different important features in the two BC GDP initialization models described in sub-section IV.C.1), GDPs are implemented to achieve roughly the same objectives at both airports. Furthermore, this similarity illustrates the potential of IRL algorithms to identify reward functions and corresponding policies that generalize—that work in a variety of contexts, including those not represented in training data. More specifically, if non-constant reward features that achieve p-values less than 0.05 are sorted from largest to smallest magnitude of the average corresponding parameter estimate, the order of the top four features for EWR are the indicator that the air buffer at the end of the time step is greater than or equal to five, the number of unused slots during rate control, the duration of GDP canceled, and the change in the air buffer. The order for SFO is nearly identical (only the top two features are swapped). These parameter estimates are all negative, as would be expected. The estimates quantify the balance achieved by traffic managers as they face a fundamental trade-off in GDP implementation: airborne delay is expensive, and implementing GDPs can help reduce it, but excessive GDP implementation can lead to undesired under-utilization of available capacity. The objective functions used in various algorithms [2, 25], including one in an operational GDP decision-support tool [13], all specify some balance for this

trade-off. However, as far as we know, this is the first time that the balance achieved in current operations has been inferred directly from historical traffic flow management initiatives and related data. Furthermore, the estimate for the parameter corresponding to the duration of GDP canceled feature provides a quantification the relative importance of continuing an existing GDP plan, which is related to the predictability metric identified by Liu and Hansen in Ref. [4].

If we investigate the average parameter estimates for less important features (those that achieve relatively high p-values and/or average parameter estimates with small magnitudes), then we find some differences between the regressors for the two airports. For example, while the average parameter estimates for buffer levels are negative for the regressors for SFO, as would be expected, they are positive for EWR. These counter-intuitive parameter estimates may help explain why the EWR IRL GDP implemented model over-predicted GDP initialization.

V. Conclusions

GDPs seem to be a tool for strategically managing traffic in an effort to achieve desired values for certain metrics that are accrued over time, suggesting that IRL may be a promising technique for GDP analytics. Therefore, we compared IRL models of GDP implementation to BC models. More precisely, we developed BC models of GDP implementation that are based on random forest models of GDP initialization and GDP cancellation. We used the CSI IRL algorithm to infer reward functions consistent with historical state and action data and then used rollouts to find policies attempting to optimize expected total discounted reward objectives based on the inferred reward functions. Furthermore, we implemented BC models for GDP parameters that are used by the rollouts policies. The models were developed for EWR and SFO and evaluated using cross validation on a data set consisting of 455 days from the summers of 2011–2013.

Features related to predictions of conditions more than four hours in the future do not improve the predictive power of the BC GDP implemented models. Similarly, we selected low discount factors of 0.4 for EWR and 0.2 for SFO for use in the IRL algorithm because these values led to reward training data that the reward regressor was better able to fit. These characteristics of the BC and IRL models are inferred from historical data and suggest that GDP implementation decisions

are more tactical than strategic: they are made primarily based on conditions now or conditions anticipated in only the next couple of hours.

When predicting GDP implementation on testing data, the BC GDP implemented models we developed for EWR and SFO demonstrate substantially stronger predictive performance than the IRL GDP implemented models we developed. The relatively poor performance of the IRL models may be caused by inaccuracies in the simple ground and air buffer model, an incomplete set of reward features, an ill-suited reward regressor form, or the invalidity of some of the underlying assumptions made by the CSI IRL algorithm, such as deterministic decision making that seeks to strategically achieve rewards accrued over time. Our experiments also suggest that neither the BC nor the IRL models predict the relatively infrequent GDP initialization or cancellation events well.

We also investigated the structure of the models in order to gain insights into GDP implementation decision making. Feature importance scores derived from the structure of the random forest BC GDP initialization and GDP cancellation models suggest that the set of most important features varies between airports; features related to scheduled arrivals, predicted airport arrival capacity levels, the previous GDP plan, certain weather conditions, and reroutes are most important for one or both airports. The reward functions inferred by the IRL algorithm are not able to achieve a good fit of the training data, but their structures suggest that decision makers at both airports are primarily concerned with avoiding relatively large numbers of flights that must incur delay in the air, avoiding unused arrival slots while delaying flights on the ground to achieve a certain rate of arrivals at the airport, and avoiding canceling a GDP long before its planned end time.

Acknowledgments

We are grateful to Shon Grabbe, Banavar Sridhar, Avijit Mukherjee, Deepak Kulkarni, and Heather Arneson for providing feedback on this research as it progressed. We are also thankful for feedback on the initial plan for this research from Tony Evans, Roberto Bunge, Ryder Winck, Paul Varkey Parayil, Nikunj Oza, Karl Bilimoria, and Tatsuya Kotegawa. David Hattaway put us in touch with Cindy Hood, Supervisory Traffic Management Coordinator at New York TRACON, who we thank for valuable insights into current GDP decision making.

References

[1] Rios, J., "Aggregate Statistics of National Traffic Management Initiatives," *AIAA Aviation Technology, Integration, and Operations Conference*, Fort Worth, TX, October 2010.

[2] Sridhar, B., Grabbe, S. R., and Mukherjee, A., "Modeling and Optimization in Traffic Flow Management," *Proceedings of the IEEE*, Vol. 96, No. 12, December 2008.

[3] Grabbe, S., Sridhar, B., and Mukherjee, A., "Similar Days in the NAS: an Airport Perspective," *AIAA Aviation Technology, Integration, and Operations Conference*, September 2013.

[4] Liu, Y. and Hansen, M., "Ground Delay Program Decision-making using Multiple Criteria: A Single Airport Case," *USA/Europe Air Traffic Management Research & Development Seminar*, Chicago, IL, June 2013.

[5] Ratliff, N., Ziebart, B., Peterson, K., Bagnell, J. A., and Hebert, M., "Inverse Optimal Heuristic Control for Imitation Learning," Paper 48, Carnegie Mellon University Robotics Institute, 2009.

[6] Wolfe, S. R. and Rios, J. L., "A Method for Using Historical Ground Delay Programs to Inform Day-of-Operations Programs," *AIAA Guidance, Navigation, and Control Conference*, Portland, OR, August 2011.

[7] Wang, Y. and Kulkarni, D., "Modeling Weather Impact on Ground Delay Programs," *SAE Journal of Aerospace*, Vol. 4, No. 2, November 2011, pp. 1207–1215.

[8] Bloem, M., Hattaway, D., and Bambos, N., "Evaluation of Algorithms for a Miles-in-Trail Decision Support Tool," *International Conference on Research in Air Transportation*, Berkeley, CA, May 2012.

[9] Kulkarni, D., Wang, Y., and Sridhar, B., "Data Mining for Understanding and Improving Decision-Making Affecting Ground Delay Programs," *Proc. of AIAA/IEEE Digital Avionics Systems Conference*, Syracuse, NY, October 2013.

[10] Mukherjee, A., Grabbe, S., and Sridhar, B., "Predicting Ground Delay Program At An Airport Based on Meteorological Conditions," *AIAA Aviation Technology, Integration, and Operations Conference*, Atlanta, GA, June 2014.

[11] Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A., "Maximum Margin Planning," *Proc. of International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[12] Abbeel, P. and Ng, A. Y., "Apprenticeship Learning via Inverse Reinforcement Learning," *Proc. of International Conference on Machine Learning*, Banff, Canada, 2004.

[13] Cook, L. S. and Wood, B., "A Model for Determining Ground Delay Program Parameters Using a Probabilistic Forecast of Stratus Clearing," *USA/Europe Air Traffic Management Research & Development Seminar*, Napa, CA, June 2009.

[14] Ramanujam, V. and Balakrishnan, H., "Estimation of Maximum-Likelihood Discrete-Choice Models of the Runway Configuration Selection Process," *Proc. of American Control Conference*, June 2011.

[15] Federal Aviation Administration, "FAA Operations & Performance Data," http://aspm.faa.gov/.

[16] Liu, P.-C., *Managing Uncertainty in the Single Airport Ground Holding Problem Using Scenario-based and Scenario-free Approaches*, PhD dissertation, University of California, Berkeley, CA, 2007.

[17] Smith, D. A. and Sherry, L., "Decision Support Tool for Predicting Aircraft Arrival Rates, Ground Delay Programs, and Airport Delays from Weather Forecasts," *International Conference on Research in Air Transportation*, Fairfax, VA, February 2008.

[18] Wang, Y., "Prediction of Weather Impacted Airport Capacity using Ensemble Learning," *AIAA/IEEE Digital Avionics Systems Conference*, Seattle, WA, October 2011.

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

[20] Wang, Y., "Prediction of Weather Impacted Airport Capacity using RUC-2 Forecast," *AIAA/IEEE Digital Avionics Systems Conference*, Williamsburg, VA, October 2012.

[21] Provan, C. A., Cook, L., and Cunningham, J., "A Probabilistic Airport Capacity Model for Improved Ground Delay Program Planning," *AIAA/IEEE Digital Avionics Systems Conference*, Seattle, WA, October 2011.

[22] Cunningham, J., Cook, L., and Provan, C., "The Utilization of Current Forecast Products in a Probabilistic Airport Capacity Model," *AMS Annual Meeting*, New Orleans, LA, January 2012.

[23] Dhal, R., Roy, S., Taylor, C., and Wanke, C., "Forecasting Weather-Impacted Airport Capacities for Flow Contingency Management: Advanced Methods and Integration," *AIAA Aviation Technology, Integration, and Operations Conference*, Los Angeles, CA, August 2013.

[24] Kim, A. and Hansen, M., "Deconstructing delay: A non-parametric approach to analyzing delay changes in single server queuing systems," *Transportation Research Part B*, Vol. 58, December 2013, pp. 119–133.

[25] Ball, M., Barnhart, C., Nemhauser, G., and Odoni, A., *Air Transportation: Irregular Operations and Control*, Vol. 14, chap. 1, Elsevier, 2007, pp. 1–61.

[26] Ball, M. O., Hoffman, R., Odoni, A. R., and Rifkin, R., "A Stochastic Integer Program with Dual Network Structure and Its Application to the Ground-Holding Problem," *Operations Research*, Vol. 51, No. 1, January-February 2003, pp. 167–171.

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-

hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825–2830.

[28] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York, 2001.

[29] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: Synthentic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321–357.

[30] Jeschkies, K., "SMOTE implementation for over-sampling," http://comments.gmane.org/gmane.comp.python.scikit-learn/5278, November 2012.

[31] Klein, E., Piot, B., Geist, M., and Pietquin, O., "A Cascaded Supervised Learning Approach to Inverse Reinforcement Learning," *Machine Learning and Knowledge Discovery in Databases*, edited by H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Vol. 8188 of *Lecture Notes in Computer Science*, Springer, September 2013, pp. 1–16.

[32] Bertsekas, D. P., *Dynamic Programming and Optimal Control*, Vol. 1, Athena Scientific, Nashua, NH, 2005.

[33] Ng, A., "Advice for applying Machine Learning," http://cs229.stanford.edu/materials/ML-advice.pdf, 2011.