Transportation Research Part E 131 (2019) 80–95

Contents lists available at ScienceDirect

Transportation Research Part E

journal homepage: www.elsevier.com/locate/tre

Using machine learning to analyze air traffic management actions: Ground delay program case study

Yulin Liu*, Yi Liu, Mark Hansen, Alexey Pozdnukhov, Danqing Zhang

Department of Civil & Environmental Engineering, University of California, Berkeley, Berkeley, CA, USA

ARTICLE INFO

Keywords:
Ground delay program
Convective weather
Support vector machine
Logistic regression
Random forest
Regularized linear models
Feature importance

ABSTRACT

We model the impact of weather and arrival demand on ground delay program (GDP) incidence. We use Support Vector Machine (SVM) to analyze how regional convective weather affects GDP incidence and find the impact depends on both distance and direction of convective activity from the airport. We then train and compare the performance of logistic regression (LR) and random forest (RF) in predicting GDP incidence using an SVM-generated regional weather variable, local weather and arrival demand. Generally, RF outperforms LR. Convective weather is the most important factor in predicting GDP incidence at Atlanta International Airport (ATL), while arrival demand has greater impact for the other airports studied. We also examined model transferability across different airports. Lastly, we build GDP duration prediction models to enable a user to assess how long a GDP is likely to continue, if it is in effect in a given hour.

1. Introduction

Ground delay programs (GDPs) are among the most common air traffic management strategies in the United States. A GDP is implemented when there are imbalances between flight demand and capacity at individual airports, or multi-airport metroplexes. When a GDP is implemented, flights bound for the affected airports are assigned controlled times of departure so that arrival demand at the destination does not exceed a specified rate. Most GDPs are the result of adverse weather that reduces airfield or terminal capacity. When initiated, GDPs severely impact flight operations, with hours of delays as well as large numbers of flight cancellations. In 2011, 1065 GDPs were issued in the US. These GDPs imposed a total amount of 26.8 million minutes of delay to 519,940 flights, with an average of 52 min per impacted flight (Liu and Hansen, 2013).

Traffic management specialists at the FAA Air Traffic Control System Command Center (ATCSCC) make decisions about ground delay programs and as well as other traffic management initiatives (TMIs) (Diao and Chen, 2018). At the beginning of each day they participate in a telecon with flight operators and FAA field facilities in which an operations plan is developed. This plan identifies likely imbalances between capacity and demand and identifies TMIs that may be needed to mitigate the imbalances. The plan may include GDPs when capacity-demand imbalances are foreseen that involve specific airports or terminal areas. The plan is then updated throughout the day, based on feedback from telecons that are take place every two hours. GDP decisions are based primarily on local airport conditions, but GDPs may also be initiated due to convective weather en route, or in the vicinity of an airport. Since weather forecasts are uncertain, GDP decisions must balance the costs of excess delay and wasted airport capacity against those of excessive airborne delay and holding. Also, as weather information and forecasts are constantly evolving, judgment is required about

* Corresponding author at: 107D McLaughlin Hall, University of California, Berkeley, Berkeley, CA 94720, USA.
E-mail addresses: liuyulin101@berkeley.edu (Y. Liu), liuyi.feier@gmail.com (Y. Liu), mhansen@ce.berkeley.edu (M. Hansen), alexeip@berkeley.edu (A. Pozdnukhov), danqing0703@berkeley.edu (D. Zhang).

https://doi.org/10.1016/j.tre.2019.09.012
Received 1 December 2018; Received in revised form 6 September 2019; Accepted 25 September 2019

when to initiate a GDP and when to defer a decision. The multi-hour durations of most flights also complicate decision making, since flights are released based on anticipated conditions when they arrive at the destination. Similar factors complicate decisions about GDP duration and scope (the set of origin airports that will be subject to departure controls), and GDP cancellations and extensions. In sum, while the basic principles that guide GDP decision making are clear and simple, in reality these decisions are complex, subjective, and accordingly difficult to predict. Moreover, interviews with traffic management specialists suggest that GDP decisions are strongly influenced by recent experiences of individual specialists, as opposed to collective experience over a longer time frame.

The purpose of this paper is to apply machine learning techniques to model the impact of convective weather, local weather and airport traffic demand on GDP incidence and duration. The motivation is three-fold. First, flight operators would benefit from having greater foreknowledge of GDPs, so that they can plan their responses further in advance, especially in the context of Collaborative Decision Making (CDM), which affords flight operators wider latitude to re-order or cancel flights during a GDP (Ball et al., 2001; Chang et al., 2001). This is also an example of a broader desire to increase the "predictability" in the national airspace system, which has been widely recognized (Liu et al., 2014; Hao and Hansen, 2014; Coppenbarger et al., 2016; Kang et al., 2017). Second, FAA specialists who are responsible for GDP decision making would value predictive models of GDP incidence and durations both to anticipate when and how long they may have to implement a GDP, and also to know when conditions are such that in the past a GDP would often have been implemented. Lastly, the analytics used in developing model, as will be discussed in Section 3, are capable of identifying where and to what extent weather matters for a particular airport, which can in turn be used to assess the similarity between weather conditions of a historical day to a given day of operation. Knowledge of what traffic management actions were taken on a similar historical day, and how well they worked, can also augment the personal experience of traffic management specialists (Gorripaty et al., 2016).

While other researchers have also linked GDP incidence with adverse weather, our analysis is unique in its attention to the *spatial pattern* of convective weather. Using Support Vector Machine (SVM), we identify the square size of the region that convective weather matters for immediate GDP incidence as $20^{\circ} \times 20^{\circ}$ latitude-longitude. Within the selected regions, we further produce $0.2^{\circ} \times 0.2^{\circ}$ latitude-longitude heatmaps capturing the impact of adverse convective weather in each square mile on GDP incidence. Additionally, the SVM results are combined with local airport weather variables such as wind and visibility to yield models that consider both local conditions at the airport and convective weather in the surrounding area. In light of our focus on convective weather, this paper limits the scope of airports in the eastern US where convective weather is believed to be an important cause of GDPs.

Our work focuses on the relation between current realized weather and GDP incidence. While for most practical uses weather forecasts would be required, our ability to accurately predict weather in future hours is continually increasing. Use of our models with weather forecasts, rather than realized weather, will reduce their reliability to some extent, but we will not address this problem in this research. Nor will we consider explicitly the fact that GDP decisions are themselves based on weather forecasts. The complex dynamics of "what did they know and when did they know it" are not readily observable; our working assumption is that the realized weather is a reasonable proxy for what was anticipated by GDP decision makers.

While we applied ML techniques to GDP prediction, the analysis can be applied to model other air traffic management strategies where the decision variable is binary. For instance, we may use SVM to understand the pattern in convective weather that would trigger the occurrence of a ground stop at an airport, an airspace flow program for an area, or a miles-in-trail (MIT) for a sector.

The remainder of this paper is organized as follows. In Section 2, we review previous work on GDPs with emphasis on works that relate GDP incidence to weather and traffic conditions. Our data and methodology are described in Section 3. In Section 4, we present our results. Section 5 offers conclusions and directions for future research.

2. Literature review

Most of the research on GDPs has focused on how to optimize GDP parameters, such as GDP start time, rates, and GDP scope, assuming a GDP is needed (Ball et al., 2003; Cook and Wood, 2009; Ball et al., 2010; Manley and Sherry, 2010; Glover and Ball, 2012; Mukherjee et al., 2012; Kuhn, 2013; Bertsimas and Gupta, 2015; Liu and Hansen, 2015; Estes and Ball, 2017; Yan et al., 2018). In general, those work emphasizes the tradeoffs that result from uncertainty about future capacity. Earlier work used capacity scenarios to characterize capacity uncertainty, and optimized planned arrival rates so as to minimize the total cost of ground delay and airborne delay (Ball et al., 2003; Cook and Wood, 2009). More recent research considers dynamic decision making in a setting where more information about future capacity becomes available at the day proceeds, and flight release decisions should therefore consider the benefit of holding a flight on the ground in order to obtain more information (Mukherjee and Hansen, 2007; Mukherjee et al., 2012). Later work also broadens the objective function to consider not just the cost differences between ground delay and airborne delay, but also delay that can be predicted when the GDP is planned and changes to the predicted delay that arise from GDP extensions and early cancelations (Liu and Hansen, 2015). The introduction of these dynamics leads to greater attention to the benefits of delaying flights from origins closer to the destination airports, which results in a shorter time lag between adjustments to GDPs based on updated information and when arrival rates at the destination airport can actually change (Ball et al., 2010). One the other hand, focusing GDP delays on shorter flights raises equity concerns, since first-scheduled-first-served has been adopted as the basic criterion for fair resource allocation in the traffic flow management community (Vossen et al., 2003). Manley and Sherry (2010) further empirically revealed the discrepancy of GDP efficiency and equity. Later Glover and Ball (2012) built upon the model of Ball et al. (2010) to incorporate both weather uncertainties and airline fairness into their objective functions. Efforts in this thread also have been devoted in Kuhn (2013), Bertsimas and Gupta (2015) and Yan et al., (2018). Another theme in recent work is to choose GDP parameters based on data about past programs under similar conditions (Estes and Ball, 2017).

Recently, the aviation community has begun applying machine learning methods to understand how and why GDP decisions were

made using historical data (Wang and Kulkarni, 2011; Mukherjee et al., 2014; Bloem and Bambos, 2015; Kuhn, 2016; Mangortey et al., 2019a; Mangortey et al., 2019b). Wang and Kulkarni (2011) used bagging decision tree to predict GDP revision events and Neural Networks (NN) to predict the GDP duration. The predictions were made at actual GDP initiation time. Their models employed the following variables: actual airport weather data from aviation system performance metrics, forecast of weather impacted traffic indexes (WITIs), and air traffic data such as scheduled arrivals. For convective weather, they considered two WITIs: en-route convective weather WITI with a scope of approximately 500 nm range and local convective weather WITI with a scope less than 100 nm.

Mukherjee et al. (2014) used logistic regression and decision tree to predict GDP incidence at an hourly level for EWR and SFO. Predictor variables in the models include actual weather condition variables, such as visibility and ceiling, and variables reflecting traffic condition at the airport, such as nominal queueing delay and demand capacity ratio. Convective weather in the airport belonged air route traffic control center is considered by using WITI. They found that logistic regression model outperforms the decision tree model in terms of prediction performance on the test data set where they used the area under receiver operating characteristic curve as the performance metric. They also found that while WITI of New York Center (ZNY) is an important factor impacting GDP at EWR, the Oakland Center (ZOA) WITI does not have such a strong influence on GDPs at SFO.

Bloem and Bambos (2015) used random forest and inverse reinforcement learning (IRL) to predict GDP initiation, cancellation and GDP parameters if a GDP is predicted to be in place. They modeled the GDP decisions in two submodels: first they predicted whether or not a GDP would be implemented for a given hour; then they predict GDP parameters such as GDP scope and GDP start time if a GDP is predicted to be needed. The model is a simplification of reality in that it requires that a GDP plan either progress as planned or be cancelled (no modifications or extensions are permitted). The predictors are essentially the same for the two submodels: actual and predicted weather condition, traffic schedule, actual and predicted airport arrival rate, runway configuration, departure queue, reroute variables and ground and air buffers. The models left out convective weather variables and thus did not consider the impact of convective weather on GDP decisions. They applied their methods to EWR and SFO airports. They found that while random forest was better than IRL in predicting hourly GDP implementations at the two airports, both models struggled to predict the initiation and cancellation of GDP.

Kuhn (2016) used random forest to predict hourly GDP incidences for the three airports in the New York area: JFK, EWR and LGA. The features considered in the model included scheduled arrivals, weather forecast for crosswind speed, visibility, thunderstorm, rain and snow, and distance from New York city to three levels of precipitation: very high level, high level, and moderate level. The three distance variables, which to some degree reflected convective weather, turned out to be the most important factors.

Mangortey et al. (2019a) developed a data fusion infrastructure to support FAA facilitating the analysis of their big data systems. The infrastructure enables automatic extraction and fusion of TMI-related data such as GDP and weather-related data from the METAR system, which contains the local weather reports at the airport level. In the paper, the authors used the infrastructure to run a small-sample case in EWR, LGA and Boston (BOS) and trained a decision tree model to predict the GDP incidence, using purely the METAR data as their features. Later Mangortey et al. (2019b) compared seven popular machine learning algorithms, including SVM, boosting ensemble, and random forest, to predict GDP incidences for EWR, LGA, San Francisco (SFO), and Los Angeles (LAX). In building their models, the authors derived local weather variables from the METAR database such dew point temperature, sea level pressure, visibility, and wind speed. The comprehensive study, while still only focusing on local airport weather condition (similar to WITI), indicates that the weather-related predictors are among the most important variables.

This paper builds upon the earlier research and the main contributions can be summarized in the following aspects. First, it incorporates convective weather using an SVM rather than WITI. WITI can be viewed as unsupervised approach to summarizing high dimensional spatial weather data into a scalar metric. The WITI assumes, based on intuition but without any theoretical or empirical justification, that the sole determinant of the importance of convective weather in a given location is the amount of traffic in that location on a normal day. In contrast, in our approach we use detailed convective data accompanied by a label indicating whether a GDP is in effect to learn weights that indicate the relative importance of convective weather in different locations in predicting GDP incidence at a given airport. These weights are then used to obtain an hourly, airport-specific, convective weather score that is combined with local airport weather and other features in a second-stage model of GDP incidence. Second, we systematically compare the performance of two classification models—random forest and logistic regression—at five different airports and a comprehensive set of metrics for predictive performance. In addition to metrics based on the confusion matrix, we also consider the ability of our models to predict temporal patterns of GDP incidence. Third, we build GDP duration prediction models based on the GDP instance prediction model. For a given hourly observation, this model predicts the number of hours that the GDP will continue to be in effect, conditional on their being a GDP in the observed hour. Thus, given reliable predictions of the feature variables, our models in combination could enable a user to assess the likelihood of a GDP being in effect in a future hour, and, if a GDP is in effect, how long it is likely to continue. Lastly, we investigate the important question of model transferability. We test the performance of a model trained on one airport in predicting GDP incidence at a different airport, as well as the predictive performance of a model trained on the data from all five airports combined.

3. Methodology

We model hourly GDP incidences using three types of information: actual regional convective weather condition, actual airport local weather condition and airport demands. Three machine learning techniques are used: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). We first use SVM to understand the spatial correlation between high dimensional regional convective weather and GDP incidences and construct a regional convective weather variable using the SVM output. We then model GDP incidence using LR and RF with the SVM-generated regional convective weather variable, local weather variables, and airport

Table 1
Number of GDP and no GDP hours.

	EWR	JFK	LGA	PHL	ATL
Number of Non-GDP hour	20,305 (82.1%)	23,160 (91.9%)	21,588 (86.6%)	22,521 (90.3%)	24,355 (98.6%)
Number of GDP hour	4418 (17.9%)	2055 (8.1%)	3354 (13.4%)	2426 (9.7%)	355 (1.4%)
Average GDP duration (hour)	5.91	4.78	6.81	5.80	4.21

traffic demand variables. Lastly, we build GDP duration prediction models using both features that have been incorporated in the GDP instance prediction model and the predicted GDP probabilities. Seven models have been investigated and compared: naïve average model, ordinary least squares (OLS) model, ridge regression, LASSO regression, elastic net regression, SVM regression, and random forest regression. We apply the analysis to five airports: EWR, JFK, LGA, PHL, and ATL. Below, we will introduce our data, variable construction, and machine learning methods.

3.1. Data sources and preprocessing

We used four different datasets from three data sources, FAA's National Traffic Management Log (NTML), National Convective Weather Forecast (NCWF) product and FAA's Aviation System Performance Metrics (ASPM).

NTML provides information on hourly GDP incidence. We used data from year 2012–2014. Each GDP incidence indicator was set to 1 if there was a GDP in effect in a given hour and airport, and 0 otherwise. For GDP incidences, we consider them independent and calculate the duration of each GDP instance by counting the number of consecutive GDPs after one occurs. Basic statistics of GDP incidences at the five studied airports are listed in Table 1.

NCWF, designed and implemented by the National Center for Atmospheric Research (NCAR), provides regional convective weather information. NCWF records current convective hazards with locational information and the direction of movement and storm tops altitude, and is updated every five minutes. The dataset covers the continental US, and all the hazardous convective weather are stored as coordinates of the boundaries of polygons projected on the 2D map. Fig. 1 presents the weather polygons in yellow from 07/21/2013 17:00 Zulu to 07/21/2013 22:00 Zulu.

ASPM provides local weather information. The weather information was recorded every 15 min in the data. We aggregated them to an hourly basis by taking the average. We used five weather columns from the ASPM data: visibility (in statute miles), ceiling (in feet), instrument meteorological conditions (IMC) dummy, wind magnitude (in knots), and wind direction. We decomposed wind to crosswind and tailwind/headwind according to the main runway configuration at the airport. Moreover, when the wind was tail wind, we set headwind to zero, and vice versa. For the five airports – EWR, LGA, JFK, ATL and PHL – the main runway configurations were selected as 22 R/L, 31, 31 R/L, 27 R/L and 27 R/L respectively.

ASPM also provided information on airport traffic demand. While the database recorded both the scheduled arrivals/departures and actual arrivals/ departures, we only kept the scheduled demands due to the fact that the actual arrivals/departures are influenced by TMIs. Similar to the local weather variables, we aggregated the demands to an hourly basis by taking the summation. As an additional indicator of demand, we include the hour of the day in local time. The local hour captures recurrent operational patterns

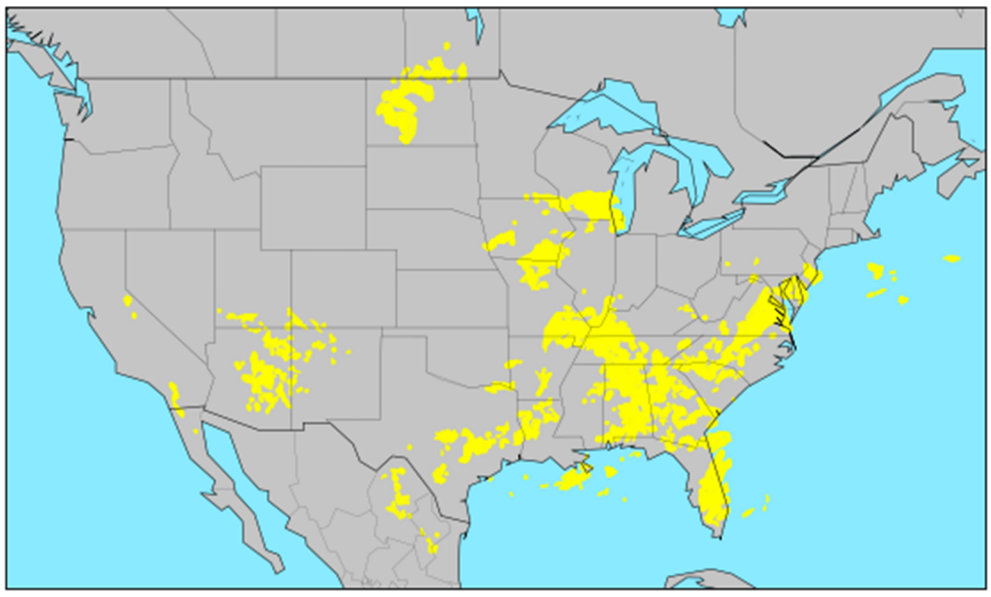

Fig. 1. NCWF convective weather map, 07/21/2013 1700Z to 07/21/2013 2200Z.

Table 2
Description of explanatory variables.

Category	Explanatory variable notation	Variable description
Local weather activities	*IMC*	Dummy variable. = 1 if instrumental meteorological condition.
	Ceiling	Ceiling (in 1000 ft).
	Vis	Visibility (in mile). = 10 if very good visibility condition.
	TW	Tailwind speed (knot), = 0 if the wind is headwind.
	HW	Headwind speed (knot), = 0 if the wind is tailwind.
	CW	Crosswind speed (knot).
Traffic demand	*SchArr*	Number of scheduled arrival flights.
	SchDep	Number of scheduled departure flights.
	LH	The local hour of GDP modeling time.

that are shaped both by scheduled demand and deviations from the schedule.

3.2. Summary statistics

After preprocessing and fusing our data sources, we obtained a dataset which includes 124,537 observations from five airports (EWR, JFK, LGA, PHL, ATL) from calendar year 2012–2014. Among those observations, 10.12% of which has GDP incidence. The description and summary statistics of ASPM-related explanatory variables are presented respectively in Table 2 and Table 3. The table shows that ATL airport in general has more observations under bad weather periods (where $IMC = 1$), and has higher arrival and departure demand than the other airports (*SchArr* and *SchDep* are much larger). The other four airports, while are geographically close to each other, have similar weather and demand patterns.

3.3. Regional convective weather modeling with SVM

To understand the *spatial* correlation between the GDP incidences and regional convective weather information, we first discretized the polygonal representations of convective areas provided by NCWF product as in Fig. 1 into geo-referenced images spanning the areas around the selected five airports. The range of area sizes and resolutions were evaluated experimentally, resulting in a squared area of $20^{\circ} \times 20^{\circ}$ latitude-longitude centered at the airport of interest, with a resolution of 100×100 pixels and one pixel per $0.2^{\circ} \times 0.2^{\circ}$ latitude-longitude. We then transformed the NCWF data into a binary variable signifying the presence of convective weather activities for each pixel. The weather incidences were recorded every five minutes, whereas the GDP data were stored every hour. Therefore, we overlaid all convective weather geo-referenced "images" within an hour and the resulting hourly convective weather feature vectors $\mathbf{x}$ are stored as lists of binary covariates.

Second, we use airport-specific linear soft-margin SVM classifiers (Cortes and Vapnik, 1995) to relate GDPs with convective weather. The input to the model is the realized convective weather data for an hour based on the overlay described above, and the outputs are the GDP labels in the corresponding hours. Each SVM model learns a set of weights $\mathbf{w}$ and an offset b that lead to the decision function Eq. (1) that minimizes misclassification rate of the decision rule Eq. (2) where $\hat{y}$ is the predicted GDP incidence, under a maximum margin separation criterion and a given trade-off hyperparameter C. Mathematically, the weights $\mathbf{w}$ and offset $b$are obtained by solving (3), where $(\mathbf{x}_i, y_i)$ is a tuple of convective weather image pixels and GDP incidence label. To fine tune the hyperparameter C from the range specified in Table 4, we use five-fold cross validation. However, we notice that the dimensionality of the feature space (100×100) is high and close to the number of data samples (e.g., EWR has 24,723 samples), which makes the classifier prone to overfit; moreover, the dataset itself is heavily imbalanced. Therefore, using accuracy alone is inappropriate. Instead, we use the area under the receiving operating characteristic curve (AUC) to select C (Wilks, 2011). To be more specific, we partition the sample into five equal sized subsamples and each time we use four subsamples to train the model and record the AUC on the fifth fold. We repeat the model training and testing five times where each time the AUC is reported on a different subsample. The

Table 3
Summary of statistics.

Variable	EWR			JFK			LGA			PHL			ATL		
	avg	min	max	avg	min	max	avg	min	max	avg	min	max	avg	min	max
IMC	0.18	0	1	0.15	0	1	0.19	0	1	0.14	0	1	0.25	0	1
Ceiling	0.40	0.	1.0	0.42	0.	1.0	0.42	0.	1.0	0.43	0.	1.0	0.45	0.	1.0
Vis	9.1	0.	10	9.1	0.	10	9.2	0.	10	9.2	0.	10	9.1	0.	10
TW	2.2	0.	40.	2.5	0.	52	2.0	0.	55.	1.5	0.	27.	2.2	0.	33.
HW	3.6	0.	39.	4.9	0.	37.	4.8	0.	34.	4.3	0.	38.	3.3	0.	29.
CW	5.9	0.	43.	7.0	0.	45.	6.3	0.	45.	5.6	0.	38.	4.3	0.	36.
SchArr	22.	0	47	23.	0	57	21.	0	46	23.	0	69	50.	0	118
SchDep	22.	0	53	23.	0	58	21.	0	49	23.	0	64	50.	0	124

Table 4
Hyperparameters in SVM, LR and RF.

Model	Hyperparameter	Description	Parameter Search Range
Soft-margin SVM	C	Penalty term for misclassification	$C = [10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1]$
Random Forest	D	Maximal depth of the trees	$D = [5, 7, 11, 15, 17, 25]$
	L	Minimal number of samples required to split a node	$L = [1, 2, 5, 10, 50, 100]$
	N	Number of trees	$N = 300$
Logistic Regression	λ	Penalty term for the l_2 regularization	$\lambda = [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 1, 10, 100, 1000]$

5-fold cross-validation result is then the average AUC's over different subsamples, and we select the C with the highest average AUC value.

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b \tag{1}$$

$$\hat{y} = \begin{cases} 1, & f(\mathbf{x}) > 0 \\ 0, & f(\mathbf{x}) \le 0 \end{cases} \tag{2}$$

$$\begin{aligned} &\min_{\mathbf{w},b} \tfrac{1}{2} ||\mathbf{w}||_2^2 + C \cdot \textstyle\sum_{i=1}^{n} \varepsilon_i \\ &\text{s. t. } y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \ge 1 - \varepsilon_i, \forall\, i = 1, 2, \cdots, n \\ &\qquad \varepsilon_i \ge 0, \forall\, i = 1, 2, \cdots, n \end{aligned} \tag{3}$$

Due to the fact that our feature maps (i.e., discretized NCWF weather maps around the airport) are locations of convective weather activities, the learned vector $\mathbf{w}$ corresponds to the weights that the presence of a convective weather within different $0.2^{\circ} \times 0.2^{\circ}$ latitude-longitude regions carries on the incidence of a GDP. Accordingly, we define Eq. (4) as the *convective weather score*, which reflects the likelihood of GDP given regional weather conditions, and use it as a predictor in the subsequent prediction models described in Section 3.3.

$$W_x = \mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b \tag{4}$$

3.4. GDP incidence prediction models

To model GDP incidence, we compare two popular machine learning algorithms: logistic regression (LR), which maximizes the likelihood of the observed GDP incidences in the dataset and yields a predicted probability of a GDP in a given hour based upon the associated features, and random forest (RF), which learns a set of decision trees that map the feature vector to GDP incidences in the leaves of trees by splitting dataset recursively (Breiman et al., 1984). We developed both models for every airport, and for each model, three sets of predictors are included: regional convection weather score W_x calculated by Eq. (4), local weather variables (visibility, ceiling, wind, etc.), and scheduled traffic demands (see Tables 2 and 3 for details). However, as will be described in Section 3.4.1, we applied a quadratic kernel to those variables for LR model to capture nonlinearities.

To balance bias and variance, both models require us to properly tune the hyperparameters. Table 4 summarizes the descriptions of parameters and the corresponding range. Similar to SVM model, we fine-tune those hyperparameters by firstly splitting the dataset into 80% of training set and 20% of evaluation set, and then using five-fold cross validation on the training set to select the best parameter(s) based on average AUC score. We then fit models using selected parameters on the full training set and predict on the evaluation set. We report five metrics: AUC, F1 score, true positive rate (TPR), false positive rate (FPR), and accuracy, and compare RF and LR models based on the overall performance of those metrics.

In addition, to investigate the transferability of our models, we also trained a Universal Model where we stacked all five airports training sets as the training data and employed the same fine-tune process illustrated above (cross validation). Finally, we compared its performance with the airport specific models using each of their evaluation sets.

3.4.1. Logistic regression

In a binary logistic regression model, the probability of a GDP incidence y_i given predictors $\mathbf{z}_i$ can be modeled using Eq. (5), where $\boldsymbol{\beta}$ is corresponding coefficient vector. To estimate $\boldsymbol{\beta}$, we minimize the loss function L_{LR} shown in Eq. (6), which employs a L-2 penalty term $\lambda\, ||\boldsymbol{\beta}||_2^2$ to avoid overfitting, and two scale parameter R_1 and R_2 to adjust the weights of losses of misclassification. While we fine tune λ using cross validation, R_1 and R_2 are predetermined and are inversely proportional to the class weights. Therefore, we add more weights to the loss induced by misclassifying the under-represented class (i.e., with GDP records).

$$\begin{aligned} P(y_i{=}1|\mathbf{z}_i) &= \frac{1}{1 + \exp(-\mathbf{z}_i^{\mathrm{T}} \cdot \boldsymbol{\beta})} \\ P(y_i{=}0|\mathbf{z}_i) &= 1 - P(y_i{=}1|\mathbf{z}_i) \end{aligned} \tag{5}$$

$$L_{LR} = -\left\{\sum_{i=1}^{n} [R_1 \cdot y_i \log(P(y_i=1|\mathbf{z}_i)) + R_2 \cdot (1-y_i)\log(P(y_i=0|\mathbf{z}_i))] + \lambda \cdot ||\boldsymbol{\beta}||_2^2\right\} \tag{6}$$

where

$$\begin{aligned} R_1 &= \frac{\text{card}(y=0)}{\text{card}(y=0)+\text{card}(y=1)} \\ R_2 &= 1 - R_1 \end{aligned} \tag{7}$$

Since LR model is inherently a linear model, we try to increase its prediction power by introducing a quadratic kernel $\phi(\mathbf{x}_i)$ shown in Eq. (8), whose elements are the upper triangle of $\mathbf{x}_i^T\mathbf{x}_i$, to the input feature space after normalizing all variables. Therefore, the final feature vector for the logistic regression model $\mathbf{z}_i$ has dimension $55 - 2 = 53$ (we exclude the two quadratic terms for *IMC* and *BusyHour* since they are dummy variables).

$$\mathbf{z}_i = \phi(\mathbf{x}_i) = (x_{i1}, x_{i2}, \cdots, x_{im}, x_{i1}^2, x_{i2}^2, \cdots, x_{im}^2, x_{i1}x_{i2}, x_{i1}x_{i3}, \cdots, x_{im}x_{i(m-1)})^{\mathrm{T}} \tag{8}$$

After we obtain the estimation results $\hat{\boldsymbol{\beta}}$, the predicted GDP label can be obtained by Eq. (9).

$$\hat{y}_i = \begin{cases} 1 & if \hat{P}(y_i=1|\mathbf{z}_i) \geq 0.5 \\ 0 & otherwise \end{cases}$$

where

$$P(y_i=1|\mathbf{z}_i) = \frac{1}{1+\exp(-\mathbf{z}_i^{\mathrm{T}} \cdot \hat{\boldsymbol{\beta}})} \tag{9}$$

3.4.2. Random forest

The random forest model is developed by training an ensemble of decision trees. Each decision tree, which is typically shallow, only uses a subset of the features, and therefore, random forest can capture nonlinearities without augmenting the feature space and yet prevent overfitting. During each tree's training, we first obtain a training set that is drawn by sampling with replacement (bootstrap samples: $\mathbb{S}$) from the original dataset ($\mathbb{D}$), and an evaluation set that is constructed by the left-out samples (out-of-bag samples: $\mathbb{O} = \mathbb{D} \setminus \mathbb{S}$). Then we find splits at each node by minimizing the Gini criteria formulated by Eq. (10), where R_1 and R_2 are defined by Eq. (7) to penalize misclassification on under-represented class (Chen, Liaw and Breiman, 2004). Notice that in the random forest, we don't introduce the kernel function $\phi(\mathbf{x}_i)$ since it already captures nonlinearities. Lastly, we calculate feature importance by evaluating each feature's Gini importance.

$$I_G = 1 - R_1 \cdot \left(\frac{\text{card}(y=1)}{\text{card}(y=0)+\text{card}(y=1)}\right)^2 - R_2 \left(\frac{\text{card}(y=0)}{\text{card}(y=0)+\text{card}(y=1)}\right)^2 \tag{10}$$

Via the completion of the training procedure, we collect N decision trees, denoted as $T = \{T_1, T_2, \cdots, T_N\}$, and pass feature vector $\mathbf{x}_i$ to every tree in T. The predicted GDP label is then a vote by all the trees (Eq. (11)).

$$\hat{y}_i = \begin{cases} 1 & if \text{card}(T(\mathbf{x}_i)=1) \geq \text{card}(T(\mathbf{x}_i)=0) \\ 0 & otherwise \end{cases} \tag{11}$$

For random forest, we can also evaluate the feature importance by calculating the average decreases of Gini importance for each individual variable across trees (Breimen and Cutler, 2018). The feature importance (decreases in Gini importance) is then rescaled so that they sum to 1. Therefore, different categories of features are additive in terms of feature importance.

3.5. GDP duration prediction models

To further predict the GDP duration (see Section 3.1 for definition), we investigate a set of linear models: ordinary least squares (OLS), ridge regression, LASSO regression, and elastic net regression (EN); and two nonlinear models: SVM regression (SVR) and random forest regression model. Features that enter those models include variables in Table 2, and the predicted probability ($Prob_{GDP}$) from the pre-trained GDP instance model. While the two nonlinear regression models are quite similar to the classification models that have been discussed in Section 3.3 (SVM model, Eq. (3)) and Section 3.4.2 (RF model, Eq. (10)), the four linear models can be formulated in Eqs. (12)–(15), where $\mathbf{x}_i^T$ represents the feature vector, y_i is the target variable, and λ_i indicates the hyperparameter that requires fine-tuning using cross validation (see Section 3.3). Lastly, we train and evaluate all 7 models using observations where GDP actually occurred in the history, and likewise in Section 3.3, training set contains 80% of total observations and test set includes the rest 20%.

$$\hat{\beta}_{OLS} = \underset{\beta}{\text{argmin}} \sum_i (\mathbf{x}_i^T \cdot \boldsymbol{\beta} - y_i)^2 \tag{12}$$

$$\hat{\beta}_{ridge} = \underset{\beta}{\text{argmin}} \sum_i (\mathbf{x}_i^T \cdot \boldsymbol{\beta} - y_i)^2 + \lambda_1 \cdot ||\boldsymbol{\beta}||_2^2 \tag{13}$$

$$\widehat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \sum_i (\mathbf{x}_i^T \cdot \boldsymbol{\beta} - y_i)^2 + \lambda_2 \cdot \|\boldsymbol{\beta}\|_1 \tag{14}$$

$$\widehat{\beta}_{EN} = \underset{\beta}{\operatorname{argmin}} \sum_i (\mathbf{x}_i^T \cdot \boldsymbol{\beta} - y_i)^2 + \lambda_1 \cdot \|\boldsymbol{\beta}\|_2^2 + \lambda_2 \cdot \|\boldsymbol{\beta}\|_1 \tag{15}$$

4. Results

4.1. Spatial importance of convective weather to GDP incidences

We finalize the selection of hyperparameter C in SVM based on GDP prediction performance and the meaningfulness in the convective weather spatial pattern. Fig. 2 shows the GDP incidence prediction performance using only convective weather from SVM. The heatmaps are plotted with the average AUCs from 5-fold cross-validation. The vertical and horizontal axes are the logarithm of SVM hyperparameter C and the lead time of convective weather with respective to the GDP hour. We first observe that average AUC tends to decrease with lead time. Accordingly, we will use results for a lead time of 0 in subsequent analysis. Next, we observer that, regardless of the lead time, the optimal choice of C ranges from 10^{-4} to 10^{-2}. To finalize the selection of C, we analyze the spatial pattern of the learned weights for the pixels. As shown in Fig. 3, the weights of the pixels are plotted for the selected 100-by-100 square region, with warmer color (towards red[1]) indicates higher magnitude of weight. Comparing the subplots under different hyperparameter C's, the spatial importance of convective weather becomes more centered around the airports as the C decreases. We also observe that when$C$ is smaller, the heat maps are more spatially coherent, with nearby locations having similar levels of importance. Moreover, since any near-identical weather phenomenon can result in GDP or no GDP given other factors, the samples of the two classes can overlap heavily in the feature space. Therefore, SVM models used in this setting should be tolerant of misclassifications. Considering all these reasons, we use the value of $C = 10^{-4}$ for all 5 airports (Pozdnoukhov et al., 2011).

We take a closer look at the spatial variability of the components of the weight vector $\mathbf{w}$ for EWR airport with $C = 10^{-4}$ shown in Fig. 4. As in Fig. 3, higher values of $\mathbf{w}$ indicate greater impact of the presence of the convective weather at a given location on the likelihood of GDP incidence at a given airport. We first observe that the spatially contingent pattern is centered on the EWR and spreads along the main east coast corridors. We note that the pattern is not radial, which implies that convective weather impact is not strictly dependent on distance from the airport. We then compare the heatmaps in the $C = 10^{-4}$ column of Fig. 3, which are weight variables for respectively EWR, JFK, LGA, PHL, and ATL. There is a consistency among the spatial patterns for all four airports in the New York metroplex: they also reveal that weather closer to the airports (center of the figures) has greater effects on airport GDP decision, but also that the area of high impact extends further along the eastern seaboard than inland. The pattern is somewhat different for the case of ATL airport, located in the southern US, but nonetheless is consistent in showing that convective weather impact is not simply a function of distance from the airport.

4.2. GDP incidence prediction models

The performance metrics for logistic regression and random forest models on the evaluation set are summarized in Table 5. While random forest models all have slightly better AUC metrics, the others differ significantly. Overall logistic regression models seem to better predict GDP incidences when there was a GDP initiative, however, the random forest outperforms logistic regression with respect to F1 score, false positive rate and accuracy across different airports, which suggests that LR emphasizes more on the positive predictions, and RF is more robust to overfitting with a better accuracy and F1 score performance on the evaluation set. Therefore, in the subsequent analysis, we will use random forest to further explain how important different covariates are with respect to predicting GDPs.

Feature importance is a metric that helps us select relevant variables, and further understand the contributions of different predictors. Table 6 reports the feature importance results. The SVM weather metric is the single most important weather variable at all airports except LGA and PHL, where ceiling is more important. However, if we sum the importance of the local weather variables – IMC, ceiling height, visibility, tailwind, headwind and crosswind speed – the result exceeds that of convective weather at all airports except ATL, where regional convective weather is uniquely dominant as a driver of GDP activity. The demand variables local hour, scheduled arrivals, and scheduled departures exhibit different degrees of importance among the airports, although their sum, which can be interpreted as a summary measure of the importance of demand, is fairly similar across all airports except Atlanta. Among the demand variables, local hour in generally is more important than scheduled demand as a predictor of GDP occurrence. Scheduled demand at an airport follows recurrent daily patterns; therefore, the hour of the day provides considerable information about scheduled demand. Moreover, busy airports follow daily operational patterns that derive both from scheduled demand and deviations from the schedule. For this reason, it is not surprising that the hour of the day is a more important predictor of GDP incidence than scheduled demand.

We further generate scatter plots to visualize the relation of the convective weather score and demand variables to predicted GDP incidence in Figs. 5 and 6, in which both horizontal axes show convective weather score, and vertical axis indicates respectively arrvial demand and local hour, and data points are colored to indicate whether a GDP is predicted or not predicted for a given hour.

[1] For interpretation of color in Figs. 3 and 7, the reader is referred to the web version of this article.

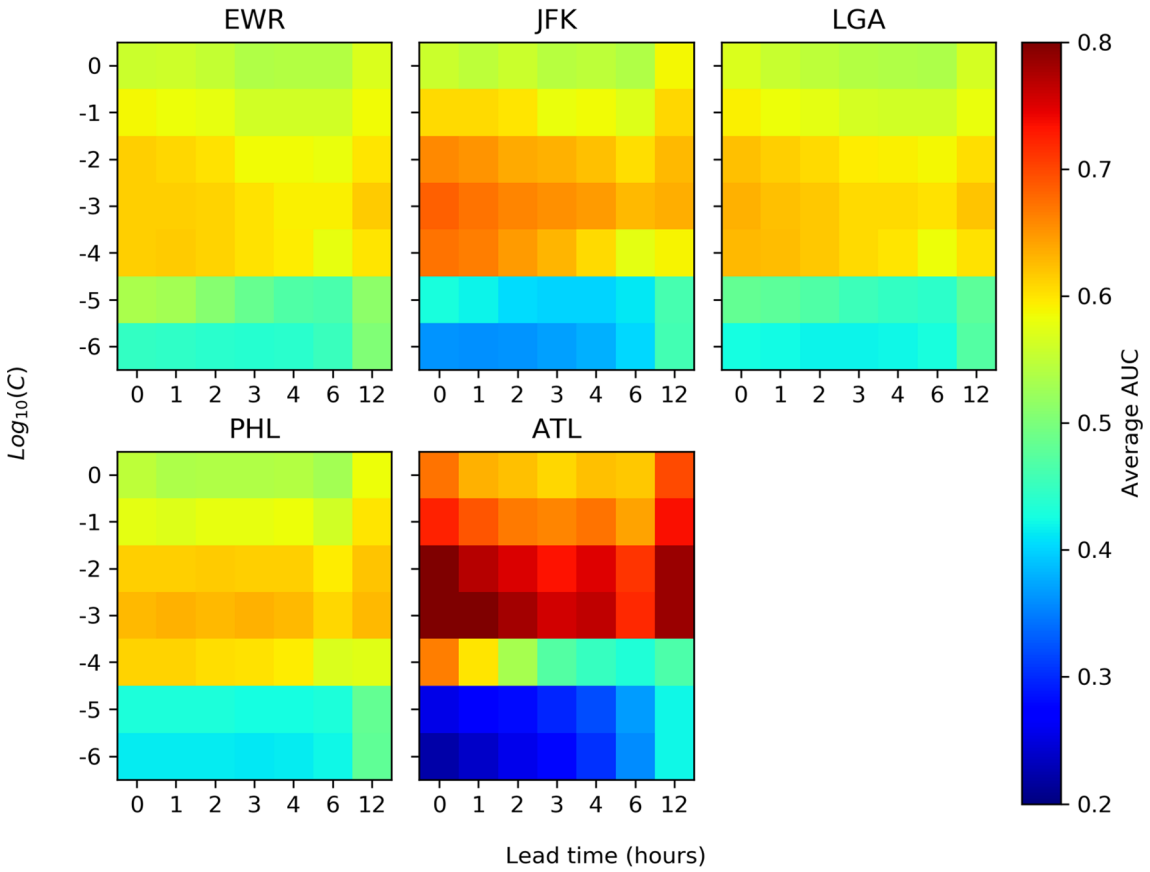

Fig. 2. Average Area under ROC Curve over 5-fold Cross-validation for five airports, as a function of hyper-parameter value C (in log scale) and GDP lead time (h).

From Fig. 5, we first notice that GDPs are more likely to be predicted when demand and convective weather score is high. However, we observe a crisper vertical boundary between predictions of GDP and no-GDP for ATL, indicating that the convective weather score has a larger influence on GDP predictions for that airport, as Table 6 also indicates. Conversely, the New York airports exhibit a sharper delineation between GDP and no-GDP predictions based on arrival demand level; this is again consistent with Table 6. In Fig. 6, since local hour and scheduled arrival demand are highly correlated, we notice that GDP incidence occurs mostly from 10 AM to 8 PM, when the airports are much busier than other time periods.

We further investigate performance of our models by their predicted temporal patterns. To achieve this goal, we first applied our trained model to the full datasets within the time period of investigation and obtained the predicted GDP incidence for every hour of the dataset. Second, we applied a moving average filter with size $2 \times 24 = 48$ h (since GDP impact rarely extend over 2 days) to both the predicted and ground truth GDP labels to create less "spiky" representations of the GDP temporal density maps. Lastly, we compared the predicted and ground truth density maps by calculating their Pearson correlation coefficient (Benesty et al, 2009). The results for five airports are summarized in Fig. 7. In each subplot, the ground truth and predicted temporal density maps are respectively shown in the top (blue) and bottom (gray) figures; the x and y axes show respectively the time stamp and density, and the Pearson correlation coefficient and its p-value are shown in the title. For all airports, the predicted temporal patterns exhibit high correlation, which ranges from 0.693 to 0.861, with the ground truth, which indicates good temporal prediction power. We also notice that our models tend to predict more GDP incidences than the ground truth, especially for the ATL airport (e.g., 2012-05-06 – 2012-07-14), which agrees with the true positive rate reported in Table 5.

4.3. Model transferability

In this section, we investigate the model transferability, which reflects how well the model trained on a dataset can be generalized to another related but different dataset. In particular, we gathered all six trained random forest models – EWR, JFK, LGA, PHL, ATL, and Universal (termed as "All" hereafter) – and the evaluation sets for the five airports. Then we applied each of the six models to every evaluation set and obtained the GDP predictions. Lastly, we compared those predictions with the ground truth in terms of F1 score, area under the ROC curve (AUC ROC score), true positive rate (TPR), false positive rate (FPR), and accuracy. Therefore, each model ended up with five transferability metrics, each of which contains a list of five measurements for the five airports. We present

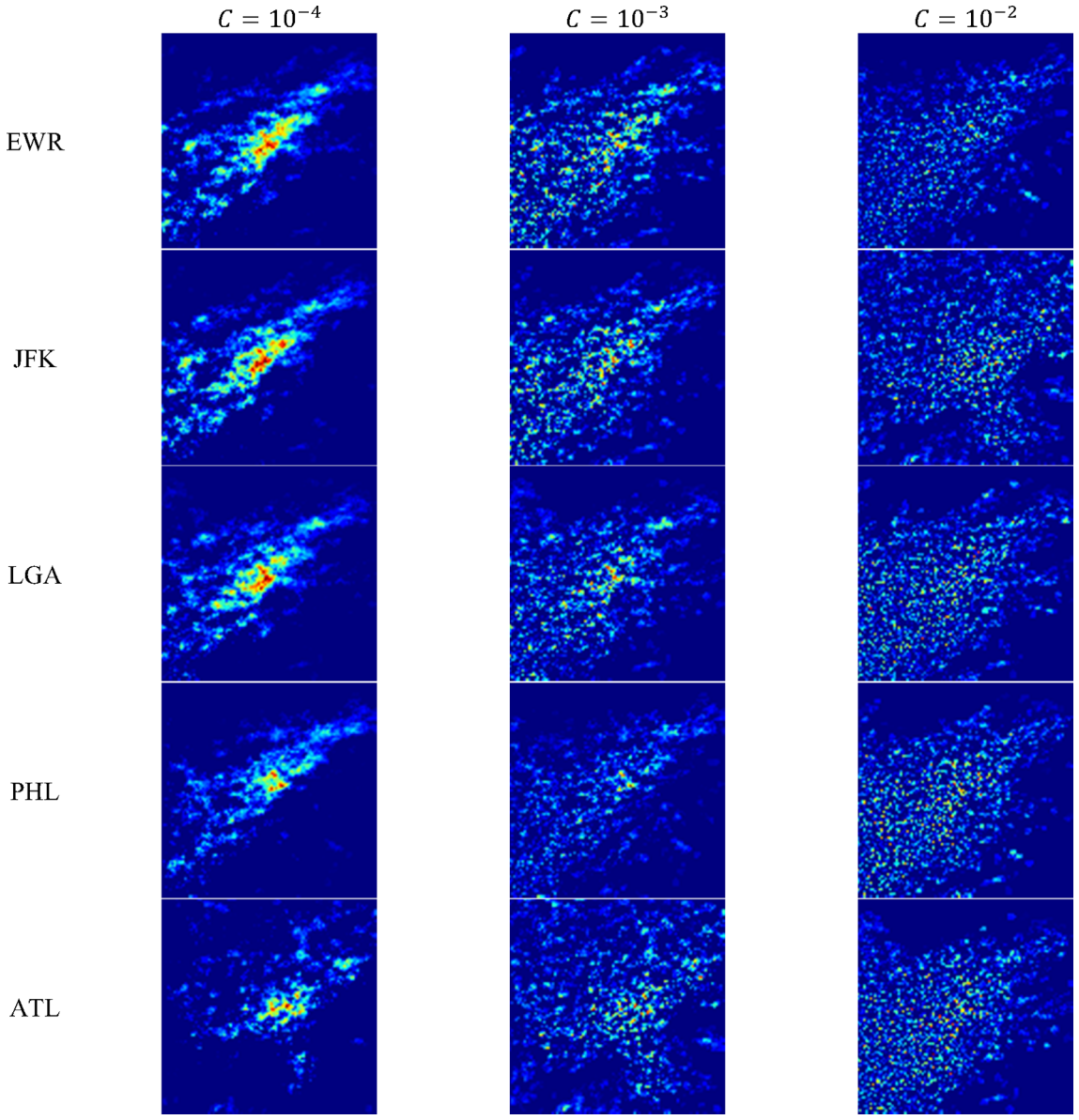

Fig. 3. Weight vectors for airports under different hyperparameters. From top to bottom shows the airports EWR, JFK, LGA, PHL, and ATL. From left to right shows the hyperparameter $C = 10^{-4}$, $C = 10^{-3}$, and $C = 10^{-2}$.

the transferability results by heatmaps shown in Fig. 8, where each heatmap indicates one metric, and rows and columns of each heatmap represent respectively the model and the evaluation set. For example, in the first heatmap (F1 score), the first row suggests that the EWR random forest model performs the best on the EWR evaluation set, followed by LGA, PHL, JFK, and ATL evaluation sets. Furthermore, the first column suggests that on the EWR evaluation set, the EWR model has the best performance, followed by LGA, "All", JFK, PHL, and ATL models. For all heatmaps, the diagonal dominates the rest of elements both row-wise and column-wise except for the TPR heatmap, where "All" model dominates other five airport-specific models. We also notice that the three New York models and PHL model have similar (good) performance on each other's evaluation sets, which suggests good transferability across those geographically adjacent airport models.

4.4. GDP duration prediction models

We report both the mean-absolute-error (MSE) and root-mean-squared-error (RMSE) across seven GDP duration regression models and five different airports. Recall that these models predict for a given hour, how long a GDP is likely to continue, assuming that the observed hour includes a GDP. To further assess the performance of our predictive models, we also evaluate these two

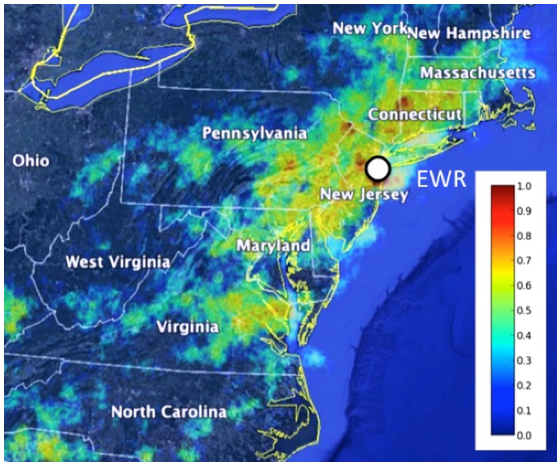

Fig. 4. Map overlay of the convective weather weight vector w for EWR model at $C = 10^{-4}$.

Table 5
Model performance.

Airport	Model	Hyper-parameter(s)	AUC	F1 score	TPR	FPR	Accuracy
EWR	LR	$\lambda = 0.1$	0.90	0.59	85.68%	23.24%	78.36%
	RF	$D = 15$; $L = 10$	0.91	0.62	81.62%	17.94%	81.98%
JFK	LR	$\lambda = 1$	0.93	0.45	87.97%	17.33%	83.09%
	RF	$D = 15$; $L = 10$	0.94	0.56	74.19%	7.97%	90.62%
LGA	LR	$\lambda = 100$	0.92	0.55	84.31%	18.25%	82.08%
	RF	$D = 15$; $L = 10$	0.93	0.64	73.69%	8.73%	88.98%
PHL	LR	$\lambda = 1000$	0.92	0.51	83.47%	15.87%	84.07%
	RF	$D = 17$; $L = 10$	0.94	0.64	74.80%	6.56%	91.58%
ATL	LR	$\lambda = 0.001$	0.92	0.13	85.90%	18.89%	81.18%
	RF	$D = 11$; $L = 50$	0.93	0.35	64.10%	3.25%	96.24%
All	LR	$\lambda = 0.001$	0.89	0.44	83.55%	22.16%	78.41%
	RF	$D = 11$; $L = 50$	0.93	0.52	87.33%	16.87%	83.56%

Table 6
Feature importance.

Variable	EWR	JFK	LGA	PHL	ATL
SVM generated *Wx*	0.087	0.122	0.126	0.128	0.572
Local Weather Variables (sum)[a]	0.235	0.261	0.354	0.460	0.185
IMC	*0.031*	*0.000*	*0.004*	*0.024*	*0.001*
Ceiling	*0.051*	*0.113*	*0.195*	*0.257*	*0.058*
Vis	*0.023*	*0.012*	*0.014*	*0.035*	*0.006*
TW	*0.016*	*0.053*	*0.028*	*0.019*	*0.017*
HW	*0.036*	*0.035*	*0.055*	*0.051*	*0.058*
CW	*0.077*	*0.047*	*0.060*	*0.074*	*0.045*
Demands (sum)[b]	0.678	0.616	0.520	0.412	0.242
LH	*0.568*	*0.527*	*0.109*	*0.320*	*0.138*
SchArr	*0.048*	*0.044*	*0.375*	*0.043*	*0.051*
SchDep	*0.062*	*0.045*	*0.036*	*0.049*	*0.053*

[a] Sum of feature importance of varibles *IMC, Ceiling, Vis, TW, HW, CW.*
[b] Sum of feature importance of variables *LH, SchArr, and SchDep.*

metrics using a naïve historical average as the predicted GDP duration. The performance metrics on the evaluation set are summarized in Table 7, where the best MAE across all models and airports ranges from 0.8 to 1.9 (h), and the best RMSE ranges from 1.1 to 2.6 (h). Overall linear models, especially the elastic net regression model, significantly outperform both nonlinear models and the naïve model (except the PHL airport, where random forest model has the best MAE). However, among all linear models, the

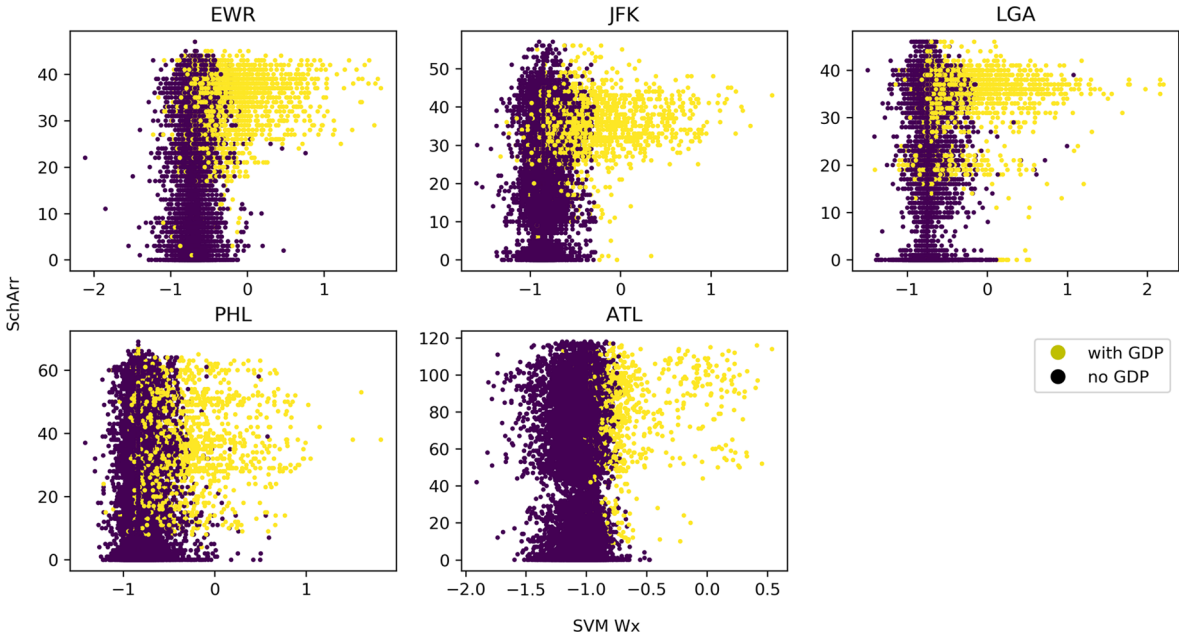

Fig. 5. GDP predictions with respect to SVM weather variable, scheduled arrivals.

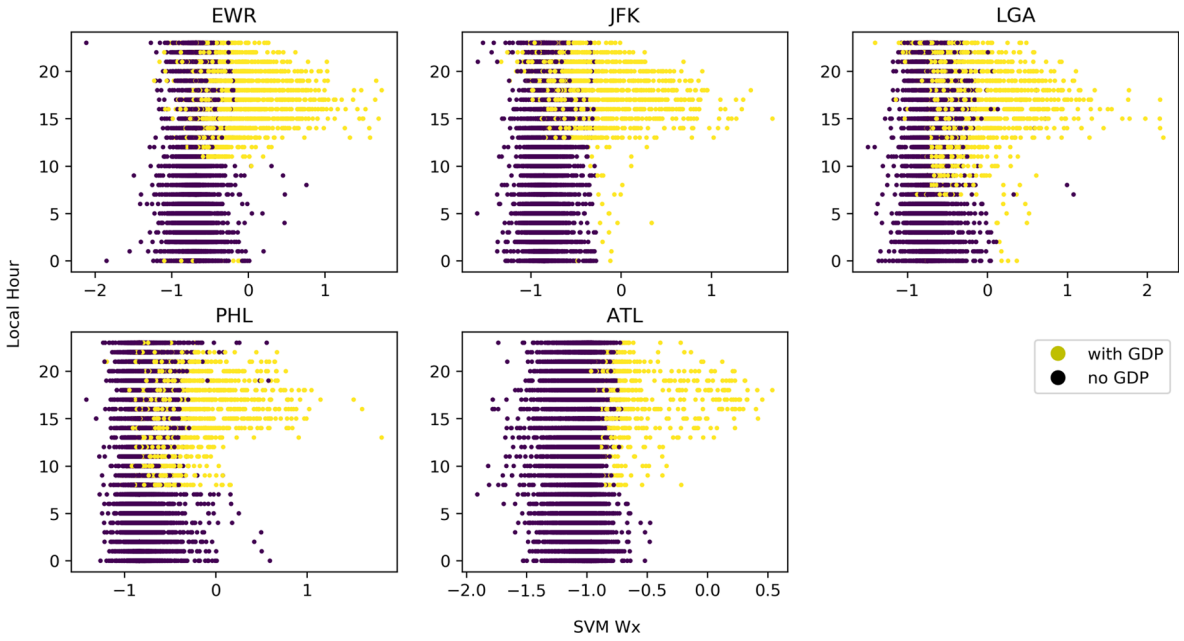

Fig. 6. GDP predictions with respect to SVM weather variable and local hour.

performances are similar. Therefore, to further uncover how different (weather) variables contribute to the GDP duration prediction models, we summarize the estimation results of the OLS model in Table 8, with emphasis of weather-related variables. We first observe that the SVM-generated convective weather score variable (W_x) has large positive and significant estimates across all five airports, especially for ATL airport, which suggests a worse weather condition is more likely to cause a longer GDP incidence. The predicted GDP probability, while not significant for ATL airport, is also highly significant for the four east coast airports. This implies that, for an hour in which a GDP is in effect, that GDP is likely to last longer if it also has a strong proclivity toward GDPs based on the GDP incidence model. Other local weather variables, while have mixed effects on GDP duration across different airports, are largely consistent for the three New York airports.

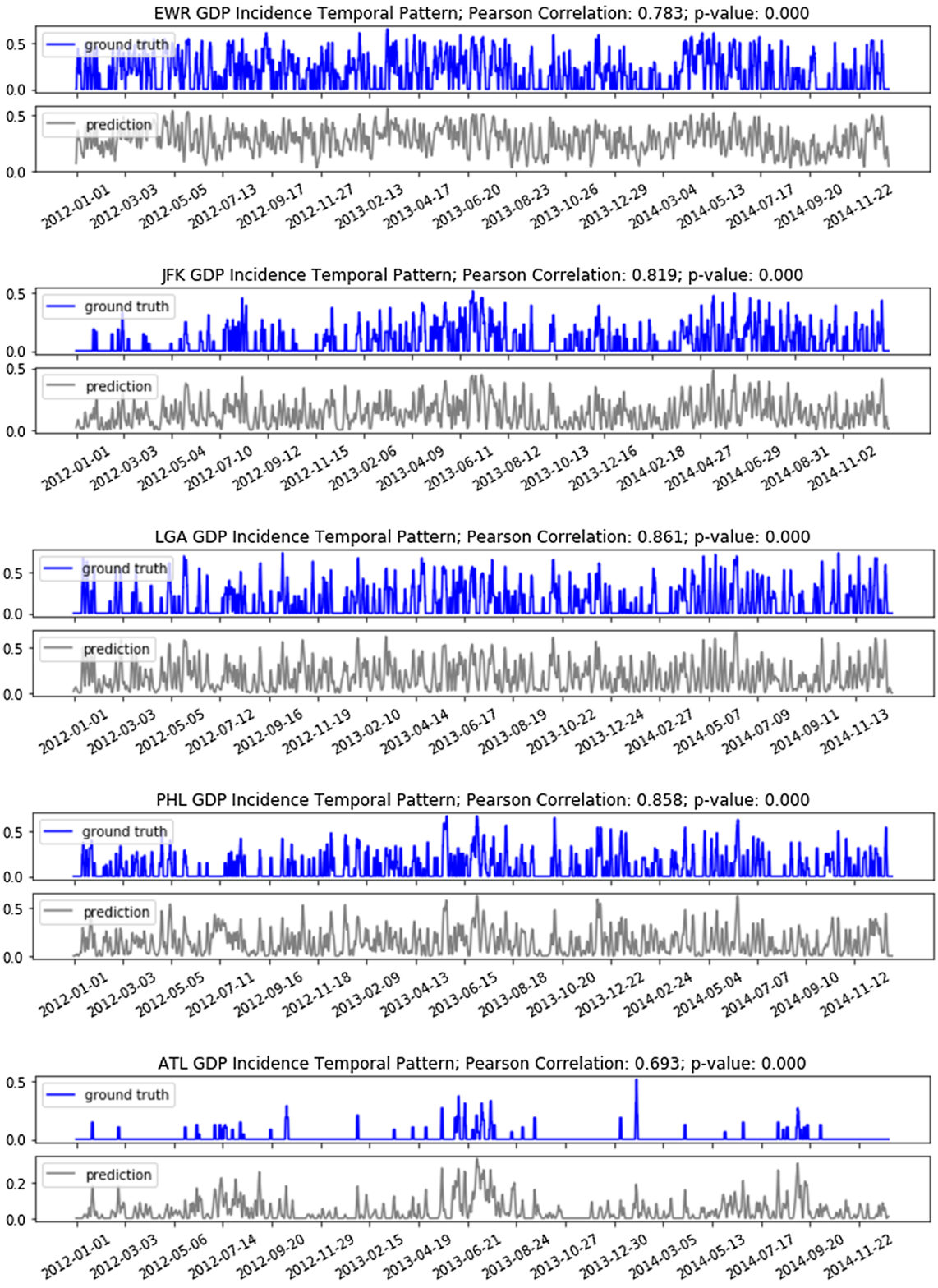

Fig. 7. Predicted GDP temporal patterns.

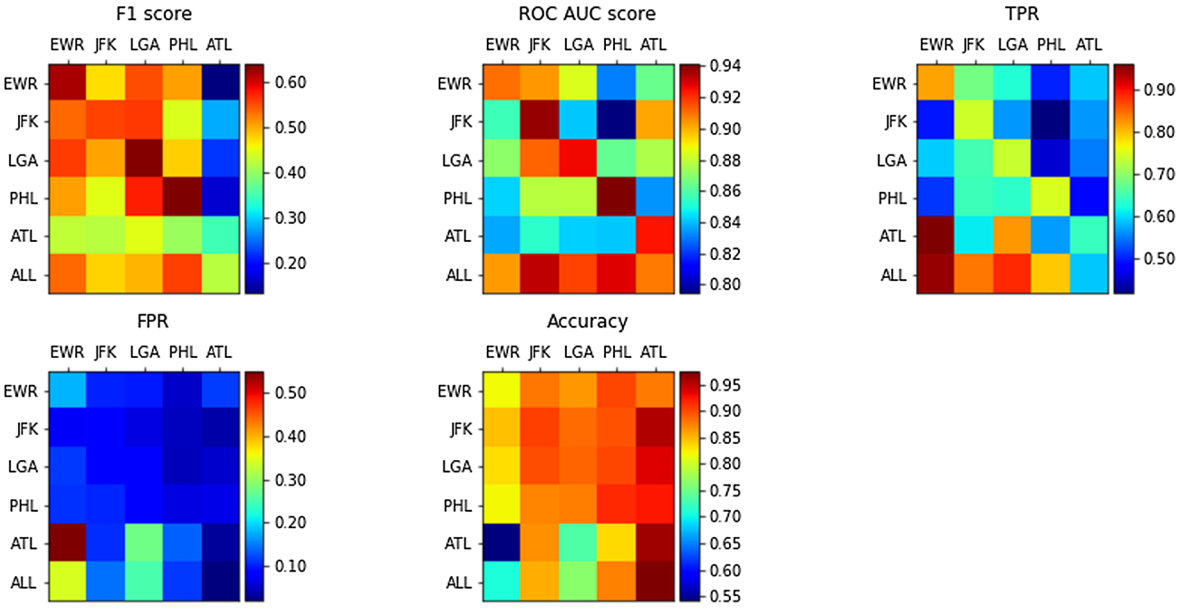

Fig. 8. Model transferability measures.

Table 7
GDP duration model performance.

Airport	Error Metric	Naïve	OLS	Ridge	LASSO	Elastic Net	SVR	RF
EWR	MAE	2.843	1.032	1.034	1.034	**1.021**	1.118	1.029
	RMSE	3.403	1.403	1.405	1.407	**1.393**	1.571	1.407
JFK	MAE	2.388	**0.812**	**0.812**	0.817	0.830	0.935	0.857
	RMSE	2.861	**1.139**	**1.139**	1.144	1.145	1.329	1.180
LGA	MAE	3.393	0.829	0.832	0.835	**0.826**	1.871	0.906
	RMSE	4.057	1.198	1.198	1.201	**1.197**	2.160	1.398
PHL	MAE	3.076	2.016	2.016	2.018	2.008	2.046	**1.911**
	RMSE	3.716	2.651	2.650	2.650	**2.649**	2.732	2.682
ATL	MAE	2.294	**1.205**	1.256	**1.205**	1.231	1.423	1.462
	RMSE	2.809	1.790	1.822	**1.783**	1.783	2.262	1.895

Note: the best model for each airport-metric tuple has been highlighted with bold texts.

Table 8
OLS model estimation results (weather related variables).

Weather-related variable name	EWR Est./Std.	JFK Est./Std.	LGA Est./Std.	PHL Est./Std.	ATL Est./Std.
IMC	0.312*** (0.085)	−0.008 (0.092)	0.059 (0.072)	0.170 (0.191)	−0.119 (0.315)
Ceiling	−0.396*** (0.067)	−0.293*** (0.090)	−0.165** (0.083)	−0.364 (0.248)	0.841* (0.450)
Vis	−0.027*** (0.011)	−0.013 (0.013)	0.005 (0.009)	−0.1069*** (0.024)	−0.011 (0.05)
TW	0.040*** (0.006)	0.042*** (0.006)	0.021*** (0.006)	0.113*** (0.017)	−0.005 (0.03)
HW	0.016*** (0.005)	0.010*** (0.005)	0.012*** (0.004)	0.017 (0.012)	−0.071** (0.032)
CW	0.023*** (0.005)	0.009*** (0.005)	0.010*** (0.005)	0.081*** (0.011)	0.020 (0.031)
W_x	0.650*** (0.080)	0.398*** (0.086)	0.131* (0.072)	0.544*** (0.072)	1.088*** (0.456)
$Prob_{GDP}$	1.147*** (0.231)	0.889*** (0.277)	0.818*** (0.199)	2.708*** (0.575)	−0.452 (1.081)
Adjusted R squared	0.848	0.843	0.919	0.553	0.570

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

5. Conclusions

This paper has attempted to predict the incidence of GDPs using spatially detailed convective weather information in the region surrounding the airport, airport local weather data, and scheduled demand. A two-stage framework is proposed. First, we use a support vector machine (SVM) model to correlate GDP incidences with regional convective activity and compute a weather score based on the location of convective weather in the airport region. Most importantly, the outcome from the SVM exhibits the spatial

correlation between the regional convective weather and GDP incidence, and therefore can be used to further understand how different locations of convection affect the GDP decision significantly. Second, we train and compare two models – logistic regression and random forest – which incorporate the weather score, local weather variables, and demand variables. We apply our method to five airports: EWR, JFK, LGA, PHL and ATL. Lastly, we use the trained model to investigate the model transferability and temporal predictability.

We find that the SVM model, when properly tuned, provides a reasonable spatially coherent picture of where convective weather is important. While, as expected, the area near the airport has the highest weights, the pattern is not purely a radial one – spatial importance distributes along the east coast corridor for airports in New York metroplex and south coast for ATL. Thus, the importance of convective weather depends on both its distance and direction from the airport.

We compare two supervised models with hyperparameters fine-tuned by their prediction performance on the test set (20% of full dataset). While both models have similar AUCs, random forest outperforms logistic regression with respect to F1 score, false positive rate and accuracy for all five airports. However, even though we increase the weight of penalizing misclassifying lower-representative class (with GDP initiatives), the true positive rate of RF is relatively low. Therefore, if the model predicts a GDP, it is likely to happen, but it does not predict the majority of GDPs that actually occur. In practice, a user of the model could choose some lower threshold than 50% in order to increase the true positive rate, albeit at the cost of also increasing the false negative one.

We then use random forest to estimate the importance of different groups of features – regional convective weather, local weather, and scheduled demands. The results suggest that for New York airports, demands have the greatest importance, followed by local weather and regional weather, while for ATL, regional weather dominates the feature importance. Then we visualize the prediction power of random forest by plotting the predicted GDPs against regional weather and scheduled arrivals variables. The figures bear out both the correlations and relative feature importance found for the different airports. Second, we apply our trained model to predict the GDP incidence for the three-year time period, and use the Pearson correlation coefficient to compare the temporal GDP density with the ground truth GDP incidences. The resulting Pearson correlation coefficients range from 0.693 to 0.861, which suggests good temporal model predictability. We further investigate the model transferability by applying trained models on different airports' evaluation sets and comparing the resulting F1 scores, area under ROC curve, true positive rate, false positive rate, and accuracy. While the performance of each airport model on its own evaluation set almost always dominates others, the three New York airport models and PHL model exhibit substantially similar performance on each other's evaluation sets, which indicates good transferability across those geographically adjacent airport models. Lastly, using both the predicted GDP probability and features that enter the GDP instance prediction model, we investigate a set of linear and nonlinear regression models to predict the GDP duration, given that a GDP has been initiated. Overall linear models, especially the elastic net model, outperform the SVR and random forest regression models, with the best MAE ranging from 0.8 to 1.3 h. The detailed estimation results of the OLS model suggest that the regional convective weather has a positive and significant effect to a longer GDP duration.

In combination, the models present a predictive capability that would be of great potential value to flight operators and other stakeholders. Assuming reliable forecasts of feature variables, the models predict both the probability of a GDP in a given hour and, should there be a given GDP, how long it is likely to last. Since this set of predictions is available for any given hour, one can imagine synthesizing them to give a more complete picture of the temporal pattern of GDP predictions over the course of a day.

Future research should be geared to improving the model and testing it in a more real-world setting. On the modeling techniques, improvement can be made to couple the SVM tuning and subsequent supervised model estimation in order to obtain the optimal tuning parameters for the application. More attention should also be given to leveraging the transferability property to develop models that use historical data from multiple airports. This is particularly desirable for the vast majority of airports where GDPs are quite rare. Another modeling improvement is to consider GDP incidence as a sequential decision process using time series models such as the recurrent neural networks. Moving the model towards real-world use, one step is to determine how to use forecast rather than realized weather in the analysis, and find how use of forecasts, which will be unavoidable in real world application, affects model performance. Lastly, consideration should be given to correlations among GDP decisions at different airports, which are implicitly considered as independent in the work presented here. With these enhancements, the model presented here has great potential to enable flight operators when to expect GDPs, and FAA specialists when to consider implementing them.

Acknowledgment

This research is funded by NASA under Award No. NNX14AJ79A.

References

Ball, M.O., Chen, C.Y., Hoffman, R., Vossen, T., 2001. Collaborative decision making in air traffic management: Current and future research directions. In: New Concepts and Methods in Air Traffic Management. Springer, Berlin, Heidelberg, pp. 17–30.

Ball, M.O., Hoffman, R., Odoni, A., Rifkin, R., 2003. A stochastic integer program with dual network structure and its application to the ground-holding problem. Oper. Res. 51 (1), 167–171.

Ball, M.O., Hoffman, R., Mukherjee, A., 2010. Ground delay program planning under uncertainty based on the ration-by-distance principle. Transp. Sci. 44 (1), 1–14.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Springer, Berlin, Heidelberg, pp. 1–4.

Bertsimas, D., Gupta, S., 2015. Fairness and collaboration in network air traffic flow management: an optimization approach. Transp. Sci. 50 (1), 57–76.

Bloem, M., Bambos, N., 2015. Ground delay program analytics with behavioral cloning and inverse reinforcement learning. J. Aerospace Inf. Syst. 12 (3), 299–313.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth.

Breimen, L., Culter, A., http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. (Last Accessed: 09/23/2018).

Chang, K., Howard, K., Oiesen, R., Shisler, L., Tanino, M., Wambsganss, M.C., 2001. Enhancements to the FAA ground-delay program under collaborative decision

making. Interfaces 31 (1), 57–76.
Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. University of California, Berkeley 110, 1–12.
Cook, L.S., Wood, B., 2009. A model for determining ground delay program parameters using a probabilistic forecast of stratus clearing. In: 8th USA/Europe Air Traffic Management R&D Seminar, 2009.
Coppenbarger, R., Jung, Y., Kozon, T., Farrahi, A., Malik, W., Lee, H., Chevalley, E., Kistler, M., 2016. Benefit opportunities for integrated surface and airspace departure scheduling. In: 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, 25–29 Sep. 2016.
Cortes, C., Vapnik, V., 1995. Support vector networks. Machine Learning 20, 273–297.
Diao, X., Chen, C.H., 2018. A sequence model for air traffic flow management rerouting problem. Transp. Res. Part E: Logistics Transp. Rev. 110, 15–30.
Estes, A., Ball, M., 2017. Data-Driven Planning for Ground Delay Programs. Transp. Res. Rec. 2603 (1), 13–20.
Glover, C.N., Ball, M.O., 2012. Stochastic optimization models for ground delay program planning with equity–efficiency tradeoffs. Transp. Res. Part C: Emerg. Technol. 33, 196–202.
Gorripaty, S., Liu, Y., Pozdnukhov, A, Hansen, M., 2016. Identifying similar days for air traffic management. In: World Conference on Transport Research, Shanghai, China, 10–15 July 2016.
Hao, L., Hansen, M., 2014. Flight predictability: concepts, metrics and impacts. NEXTOR Final Report.
Kang, L., Liu, Y., Hoffman, R., Hansen, M., 2017. Ground delay program decision-making based on utility maximization (in preparation).
Kuhn, K.D., 2013. Ground delay program planning: delay, equity, and computational complexity. Transp. Res. Part C: Emerg. Technol. 2013 (35), 193–203.
Kuhn, K.D., 2016. A methodology for identifying similar days in air traffic flow management initiative planning. Transp. Res. Part C 69, 1–15.
Liu, Y., Hansen, M., 2013. Evaluation of performance of ground delay program. Transp. Res. Rec. 2400, 54–64.
Liu, Y., Hansen, M., 2015. Incorporating predictability into cost optimization for ground delay programs. Transp. Sci. 50 (1), 132–149.
Liu, Y., Hansen, M., Gupta, G., Malik, W., Jung, Y., 2014. Predictability impacts of airport surface automation. Transp. Res. Part C 44, 128–145.
Mangortey, E., Gilleron, J., Dard, G., Pinon-Fischer, O.J., Mavris, D.N., 2019a. Development of a data fusion framework to support the analysis of aviation big data. In: AIAA Scitech 2019 Forum, p. 1538.
Mangortey, E., Pinon-Fischer, O.J., Puranik, T.G. and Mavris, D.N., 2019b. Predicting the occurrence of weather and volume related ground delay programs. In: AIAA Aviation 2019 Forum, p. 3188.
Manley, B., Sherry, L., 2010. Analysis of performance and equity in ground delay programs. Transp. Res. Part C: Emerg. Technol. 18 (6), 910–920.
Mukherjee, A., Grabbe, S., Sridhar, B., 2014. Predicting ground delay program at an airport based on meteorological conditions. In: AIAA Aviation Technology, Integration, and Operations Conference, AIAA Paper 2014–2713, June 2014.
Mukherjee, A., Hansen, M., 2007. Dynamic stochastic model for single airport ground holding problem. Transp. Sci. 41 (4), 444–456.
Mukherjee, A., Hansen, M., Grabbe, S., 2012. Ground delay program planning under uncertainty in airport capacity. Transp. Plan. Technol. 35 (6), 611–628.
Pozdnoukhov, A., Matasci, G., Kanevski, M., Purves, R.S., 2011. Spatio-temporal avalanche forecasting with Support Vector Machines. Nat. Hazards Earth Syst. Sci. 11, 367–382.
Vossen, T., Ball, M.O., Hoffman, R., Wambsganns, M., 2003. A general approach to equity in traffic flow management and its application to mitigating exemption bias in ground delay programs. Air Traffic Control Quart. 11 (4).
Wang, Y., Kulkarni, D., 2011. Modeling weather impact on ground delay programs. SAE Int. J. Aerosp. 4 (2), 1207–1215.
Wilks, D., 2011. Chapter 8: Forecast Verification, Statistical Methods in the Atmospheric Sciences, third ed. Elsevier.
Yan, C., Vaze, V., Barnhart, C., 2018. Airline-driven ground delay programs: a benefits assessment. Transp. Res. Part C: Emerg. Technol. 89, 268–288.