Forecasting Airport Transfer Passenger Flow Using Real-Time Data and Machine Learning

Xiaojia Guo
UCL School of Management, University College London, London, UK, x.guo.11@ucl.ac.uk

Yael Grushka-Cockayne
Harvard Business School, Cambridge, MA 02163, ygc@hbs.edu
and Darden School of Business, University of Virginia, Charlottesville, VA 22903

Bert De Reyck
UCL School of Management, University College London, London, UK, bdereyck@ucl.ac.uk

Problem definition: Airports and airlines have been challenged to improve decision-making by producing accurate forecasts in real time. We develop a two-phased predictive system that produces forecasts of transfer passenger flows at an airport. In the first phase, the system predicts the distribution of individual transfer passengers' connection times. In the second phase, the system samples from the distribution of individual connection times and produces distributional forecasts for the number of passengers arriving at the immigration and security areas.

Academic/Practical relevance: Our work is the first to apply machine learning for predicting real-time distributional forecasts of journeys in an airport, using passenger level data. Better forecasts of these journeys can help optimize passenger experience and improve airport resource deployment.

Methodology: The predictive system developed is based on a regression tree combined with copula-based simulations. We generalize the tree method to predict distributions, moving beyond point forecasts. We also formulate a newsvendor-based resourcing problem to evaluate decisions made by applying the new predictive system.

Results: We show that when compared to benchmarks, our two-phased approach is more accurate in predicting both connection times and passenger flows. Our approach also has the potential to reduce resourcing costs at the immigration and transfer security areas.

Managerial implications: Our predictive system can produce accurate forecasts frequently and in real-time. With these forecasts, an airport's operating team can make data-driven decisions, identify late passengers and assist them to make their connections. The airport can also update its resourcing plans based on the prediction of passenger flows. Our predictive system can be generalized to other operations management domains, such as hospitals or theme parks, in which customer flows need to be accurately predicted.

Key words: quantile forecasts; regression tree; passenger flow management; data-driven operations.
History: June 15, 2020

1. Introduction

Passengers often experience delays when travelling through airports, especially at immigration and security. These delays are caused in large part by the volatility and uncertainty in arrival patterns of passengers. Airports and airlines have long invested in optimizing and controlling aircrafts' arrivals and departures (Barnhart and Cohn 2004, Lohatepanont and Barnhart 2004, Lan et al. 2006, Atkinson et al. 2016). Once passengers have disembarked, however, airports typically have little knowledge of passengers' whereabouts in the airport. Improved passenger tracking in real time would enable airports to better serve their passengers, stabilize and predict departure times, and plan resourcing needs.

Flights landing or departing from international hubs often carry a high proportion of connecting passengers. For example, flights traveling through Heathrow, the busiest airport in Europe, have at least 30% connecting passengers (Heathrow 2020). Missed connections are the third leading reason for filing a complaint (MacDonald 2016). Better predictions of the time passengers need to traverse the airport can help minimize such missed connections, or if unavoidable, can alert airlines in advance so that their schedules are not affected by unexpected late arrivals. Additionally, predictions of transfer passenger movements can help predict bottlenecks at immigration and security, allowing for preventive actions, thereby avoiding further delays.

Working with Heathrow airport, we develop a two-phased system to produce both individual level predictions, namely the distribution of passengers' connection times, and aggregate level predictions, i.e. the distribution of the number of passenger arrivals at immigration and security. Here, the connection time predicted in the first phase is defined as the time difference between a passenger's arrival at the airport, i.e. when their plane arrives at its gate, and their arrival at the airlines' conformance desk, the last check point before passengers progress to immigration and security. The conformance desk is also where the airlines check whether a passenger has sufficient time to make their connection. Therefore, we can use the predicted distribution of the connection time to calculate the likelihood of a passenger missing their connecting flight.

In the second phase of our model, the system samples from the distribution of each passenger's connection time to predict the distribution of passenger flows, or the number of arrivals at the conformance desk, which could enable managers to make robust resourcing decisions for peak and trough passenger flows (Borndörfer et al. 2007, Solak et al. 2009, Wu and Mengersen 2013) and when there are understaffing and overstaffing costs (Bassamboo et al. 2010, Zychlinski et al. 2019). The system is currently in use at Heathrow airport, supporting the resourcing of immigration desks and security lanes, informing airlines of late passengers, and facilitating improved collaborative decision-making. Although our predictive system is developed for airport decision-making, it can be generalized to

other domains, such as hospitals and theme parks, in which estimates of the number of arrivals are needed for customer flow management.

Passenger delays and passenger flow management have received some attention in the literature (Barnhart et al. 2014, Wei and Hansen 2006). Lack of passenger level data, however, has made it difficult to explore passenger-centric problems. Our paper is the first to study passengers' transfer journeys using airport data, and to provide decision support in real time. Decision-making using real-time information has been studied in other contexts (Bertsimas and Patterson 2000, Mukherjee and Hansen 2009, Jacquillat and Odoni 2015), but most of these studies focus on air traffic issues. Additionally, the integration of machine learning, big data, and real-time decision making has received limited attention in the literature (Shang et al. 2017). Our study is the first to exploit large data sets of flight and passenger information using customized machine-learning algorithms.

A system that can be implemented at the airport to predict transfer passengers' movements should meet three requirements. First, the approach must produce both accurate point forecasts and well-calibrated distributional forecasts. Second, operational plans need to be updated frequently; and thus, any predictive system developed must be capable of generating forecasts rapidly. Third, the model needs to be intuitive in the sense that it enables airport managers to understand the key factors that influence passengers' connection times. This third requirement was essential in order to secure the buy-in of Heathrow teams for scaling up implementation.

Our predictive model of connection times is based on regression trees (Breiman et al. 1984). Regression trees partition the feature space and then fit a simple model (e.g. simple average) to each of the segments (Hastie et al. 2009). The interpretation of the results from a regression tree is intuitive compared to other advanced machine learning methods. Although regression trees have been widely used to make point forecasts in business (Eliashberg et al. 2007, Ferreira et al. 2015, Xue et al. 2015), few have applied it to make quantile forecasts or to generate prediction intervals.

In this study, we use regression trees to categorize passengers, and then generate the probability density function (pdf) of their connection times. Using these individual distributions, we can then calculate the distribution of the number of arrivals at immigration and security within certain time intervals in closed form. However, this is only possible if passenger arrivals are assumed independent, which is not the case in practice, and would make the distribution of passenger arrivals overconfident. Therefore, we add dependences among passenger arrivals to produce well-calibrated distributional forecasts, using simulation. The degree of dependences, which become tuning parameters in our approach, are incorporated into the model using Gaussian copulas.

Several results are derived from our model. First, we compare the performance of our model in forecasting individual connection times and identifying late passengers against several benchmark methods. Given the simplicity of a regression tree model, we anticipated trading off accuracy for

improved interpretability and run time. Surprisingly, however, the regression tree model performed favorably. Second, we compare the performance of our two-phased approach in predicting passenger flows with traditional time series models and Heathrow's legacy systems. Our two-phased approach outperforms the benchmarks in both point forecasting and quantile forecasting. Finally, we report several findings regarding passengers' connection times based on the regression tree.

To evaluate the impact of our predictive system on real-time decision-making, we simulate a newsvendor-based resourcing problem, reflecting understaffing and overstaffing costs at immigration and security. Such costs represent passengers' dissatisfaction on the one hand and resourcing costs on the other. The optimal resourcing level in this problem is a function of the quantile forecasts of passenger flows. Based on this formulation, we show that in comparison to Heathrow's legacy system, our predictive system could reduce the cost at immigration and security by 12% to 54%, depending on the ratio of staffing cost to understaffing cost. Moreover, our predictive system could offer a 11% to 29% cost reduction compared to traditional time series models.

Our work makes the following contributions. First, to our knowledge, we are the first to develop a predictive system using real-time data to forecast passenger movements at an airport. Second, our two-phased approach, generating individual and aggregate level predictions simultaneously, exhibits superior performance over the traditional approach which considers the two forecasting tasks as independent problems. Although various components of our model (e.g. generating aggregate level predictions using individual information or using copulas to introduce dependences in customer flows) have been used in previous studies (e.g. Hoot et al. 2008, Van Brussel 2018), the combined system, and the focus on generating and utilizing distributional forecasts, are new. As a final contribution, our system has been fully implemented at Heathrow, where it is being used to minimize missed connections and reduce congestion at immigration and security areas, resulting in improved passenger satisfaction. When combined, these contributions can assist practitioners and academics looking to use machine learning to solve practical operations management problems.

2. Problem Description

In this section, we describe a typical transfer journey at Heathrow airport, the related decision-making processes, and the forecasting challenges involved.

2.1. The Transfer Journey

The journey for connecting passengers at Heathrow airport is not dissimilar to that at other international hubs. In our study, we focused on the most complex passenger journey at Heathrow, i.e. passengers arriving on international flights, and connecting to a flight at Terminal 5. Typically, once an international flight lands and the passengers disembark, connecting passengers follow signs for flight connections. Passengers arriving at Terminals 2, 3, or 4 first transfer to Terminal 5 using a

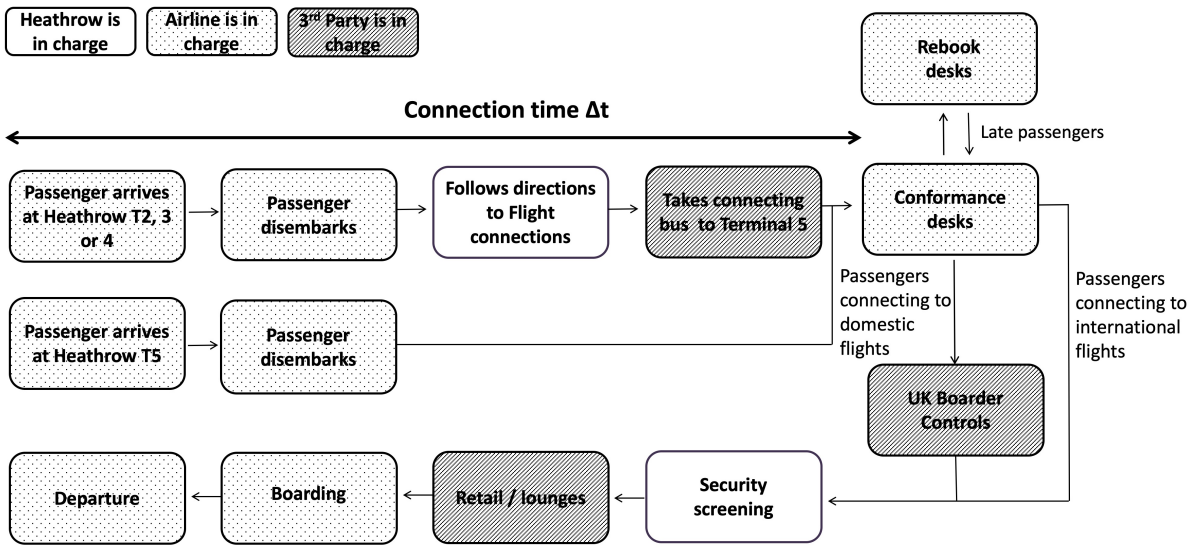
shuttle bus. Upon arrival at Terminal 5, all passengers check in at the airport conformance desks, where their boarding pass is checked to ensure that the passengers are in the right place with enough time to catch their onward flight. If a passenger arrives after a specific cut-off time (typically 30 minutes before connecting flight's departure), they are deemed unlikely to reach their outbound flight in time, and are redirected to a ticket desk for assistance. This transfer journey is depicted in Figure 1.

Passengers connecting to domestic destinations enrol at UK Border Control, and then progress to security screening. Enrolment at UK Border Control is not required for passengers connecting to international destinations. After passing through the conformance desk, these passengers progress directly to the security screening. Finally, after passing airport security, passengers enter the departure lounge, and proceed toward their boarding gates.

2.2. Decision-making Processes and Forecasting Challenges Related to the Transfer Journey

Heathrow's passenger flow manager, an individual working with terminal-based operations, is in charge of resource deployment and optimizing passengers' experience. In cases where the passenger flow manager realizes that there is likely to be passenger congestion, she informs relevant teams, such as the Border Force and airlines, and recommends actions to resolve the issue. Examples of the recommended actions by the passenger flow manager include adding more staff at the Border Force, offloading and rebooking late passengers to other flights, or delaying a flight if the majority of the passengers connecting to the flight are going to be late. Heathrow's security flow manager, an individual in charge of the security screening areas, continually assesses the flow of passengers, and

Figure 1 The connecting journey of international arriving passengers departing through Terminal 5



makes decisions, in real time, on how many lanes are required to be opened and whether staff could be better distributed between terminals (De Reyck et al. 2016). Both the passenger flow manager and the security flow manager operate from Heathrow's Airport Operation Centre (APOC), which consolidates all airport operations including gate management, security, baggage, passenger processes and crisis management in order to improve data sharing and collaborations across teams (Eurocontrol 2010).

Prior to the current study, decisions relating to transfer passengers were not all made with complete information in the center, preventing proactive and truly collaborative decision-making. Without individual level forecasts of connection times, for instance, the passenger flow manager can not precisely identify passengers with high chances of missing their outbound flight in order to take proactive actions. Also, lack of real-time predictions of passenger flows into immigration and security areas limits the passenger and security flow managers' ability to adjust necessary resources. Heathrow's legacy system generated forecasts of transfer passenger flows the day before operation, based on historical patterns, e.g. historical averages of the proportion of transfer passengers and connection times between terminals, and the predicted schedule of the next day's arriving flights. As no real-time passenger information was used, significant variability in arrival times and passenger delays made this approach of limited value. Recognizing this shortcoming, the security flow manager would monitor passengers through CCTV cameras. However, it was often too late to reallocate staff by the time the manager realized there was potential congestion.

To support the decision-making processes mentioned above, two forecasting challenges needed to be solved. The first challenge was to predict the likelihood of a passenger missing her connecting flight. Given these probabilities, airlines, with the support of the passenger flow manager, would be able to help late passengers move faster through the airport or facilitate early offloading and rebooking. The second forecasting challenge was to predict the distribution of passenger flows, or the number of arrivals at the immigration and security areas within certain time intervals, to support resourcing decisions.

There are two major constraints to consider: the availability of real-time data and the time needed to adjust plans. For example, it usually takes a security flow manager more than half an hour to move staff between terminals. To guarantee that a manager has enough time to make adjustments, the forecasts should be provided as soon as new data comes in. This requires an effective procedure that can produce accurate forecasts ahead of time.

3. The Predictive Model

A typical approach for solving the two forecasting challenges described in Section 2 would be to produce two independent forecasts: one from a classification model that identifies late passengers,

and one from a time series model that predicts passenger arrivals (Wei and Chen 2012, Milenković et al. 2018). In this paper, we propose a two-phased approach, with the output from one model serving as input to the second. As we describe below, the two-phased approach produces better forecasts of passenger flows compared to a single-phase model, by allowing the use of real-time information, calibration between the phases using hyper parameters, and the use of correlations between passenger movements.

In the first phase, we predict the distribution of passengers' connection times, Δt, the time between a passenger's arrival at the airport and their arrival at the conformance desk, where passengers have to show up 30 minutes before their next flight. In the second phase, using the distributions of individual connection times, we generate distributional forecasts of the number of passenger arrivals at the conformance desk within certain time intervals.

Our two-phased approach generates better forecasts of passenger flows for two reasons. First, by utilizing distributions of individual connection times, real-time passenger and flight information, such as a passenger's travel class and the inbound flight's stand type, can be incorporated when predicting passenger flows. Second, although the calibration of the distributional forecasts depends largely on the performance of the model in the first phase, these forecasts can be calibrated using a hyper-parameter. In our model, this hyper-parameter is the correlation between passengers' arrivals at the conformance desk, and is tuned using a validation set. Intuitively, arrivals of passengers travelling together, or from the same inbound flight, are likely to be highly correlated. The traditional time series models are fit to the aggregate level data, i.e. the time series of the number of arrivals at the conformance desk, making it difficult to incorporate correlations between individual passengers. In addition, a single predictive framework can better support airport collaborative decision-making. For example, managers now can easily see the impact of a decision, such as delaying an outbound flight with many delayed transfer passengers, on both missed connections and on the number of passenger arrivals at immigration and security.

In the rest of this section, we first describe the data used to train and test the predictive model. We then follow with details on how we evaluated the model's forecasts. Finally, we explain how we developed the model. All the code developed in this study is available from the authors.

3.1. Data Processing

In our model, we use both flight and passenger data. Flight data includes information on departures, arrivals, and aircraft features (e.g., aircraft body type). Passenger data includes individual passenger information, such as their travel class. Historical data for all of 2015, a total of 3,762,690 records, was used to train and test the model. This data was collected from three data sets: Heathrow's Business Objective Search System (BOSS), Baggage Daily Download (BDD), and Conformance data. Details of the databases used in our study can be found in De Reyck et al. (2016).

The target variable of our study — passengers' connection times Δt — was calculated as the time between the arrival time at the airport, measured by the on-chock time of the aircraft, and the arrival time at the immigration and security areas. According to Heathrow, the service time and queues at the conformance desk are negligible, so we use the time a passenger scans their boarding pass at the conformance desk to approximate their arrival time at the immigration and security areas. The data was cleansed for errors on connection times, such as rerouted passengers (0.24%), negative connection times (1%), and extremely long connection times greater than the 99% quantile. In total, we removed 84,339 records, and the resulting data set contains 3,678,351 passenger records and 32 variables. The median and mean of Δt are 27.0 minutes and 30.5 minutes, respectively. Passenger records of the first 80% of the days in the data set formed our training and validation set. The remaining 20% was used as our out-of-sample testing set.

3.2. Model Variable Selection

It is well known that variable selection and creating new input features are key parts of building an accurate machine learning model (Domingos 2012). In this study, we create eight new features relying on Heathrow experts' knowledge of the aviation domain and the connecting passengers' journey. We anticipate that other airports are likely to find most of these variable relevant as well, as procedures do not vary substantially. The new features we included were:

Inbound flight region and outbound flight region. Where a passenger is travelling from and to (Europe, East Asia, North America, rest of the world) may have an effect on the time needed to traverse the airport. Passengers travelling from or to European countries, for instance, may be more familiar with Heathrow, resulting in shorter connection times. Also, long-haul passengers may have more hand luggage and therefore need more time to disembark.

Punctuality of the arriving flight. Punctuality is defined as a flight's actual on-chock time minus its scheduled arrival time. This variable can be negative when a flight arrives ahead of schedule. Passengers on late flights might collectively be affected, e.g. due to a lack of gate availability.

Hour of the day the arriving flight lands at the airport. A passenger can move faster through the airport during hours when the airport is not busy.

Perceived connection time. This feature is calculated as the connecting flight's scheduled departure time a passenger has at the time they depart from the original airport, minus the arriving flight's actual on-chock time. The perceived connection time may influence the stress level of a passenger trying to make a connection, thereby influencing the speed with which they move through the airport.

Arriving and connecting flight load factor. The load factor is calculated as the ratio of the actual number of passengers to the capacity of the flight. Passengers arriving on a flight that is relatively full may need more time to disembark from the aircraft.

Day of the week. Day of the week has been shown to be a significant factor on passenger delays in the literature (Barnhart et al. 2014).

We considered a total of 40 predictors, given in Table A.1 in the Appendix. In the end, we only used 17 predictors, including the six newly created features described above, because the others (1) were not available in real time, (2) were too specific or entailed too many categorical levels, or (3) did not improve the model's accuracy when included. Summary statistics for the 17 predictive variables can be found in Table A.2 and Table A.3 in the Appendix.

3.3. Model Accuracy

In this study, we generate distributional forecasts of passengers' connection times and the number of passenger arrivals at the conformance desk. Thus, our objective is to minimize the error between the distributional forecasts and the realizations. When evaluating the performance of our model and the benchmarks, we measure the accuracy of both distributional forecasts (probabilities and quantiles) and point forecasts.

The primary scoring rule we use for evaluating distributional forecasts is the pinball loss function (Jose and Winkler 2009), computed for multiple quantiles that roughly describe the entire distribution. The pinball loss is a piecewise linear function and is negatively-oriented (lower is better). According to the pinball loss, if a realization falls above a reported quantile, say, the 0.05-quantile, the quantile's loss is its distance from the realization multiplied by its probability of 0.05. Otherwise, if the realization falls below the reported quantile, the quantile's loss is its distance from the realization multiplied by one minus its probability (0.95 in the case of the 0.05-quantile). The results in the pinball loss function penalizing low-probability quantiles more for overestimation than for underestimation and vice versa in the case of high-probability quantiles. In the example of the 0.05-quantile, we would penalize more when the realization falls below this low-probability quantile than above it.

Given the realization y_i of the ith observation, the pinball loss of the p-quantile (Q_p) is

$$PL(Q_p, y_i) = \begin{cases} p(y_i - Q_p) & \text{for} \quad Q_p \leq y_i \\ (1-p)(Q_p - y_i) & \text{for} \quad Q_p > y_i \end{cases}$$

The different values p and 1-p for the slope of the loss function when the quantile is below and above the realization, respectively, reflect the desired imbalance in evaluating quantile forecasts (Gneiting 2011). For the median ($p = 0.5$), the loss function is symmetric with an equally weighted loss for a realization falling below or above its reported quantile. For the tails of the distribution, the function is asymmetric with a higher weighted loss when the realization falls below a reported low-probability quantile ($p < 0.5$) or above a reported high-probability quantile ($p > 0.5$).

The pinball loss has two additional appealing properties: (1) it is a strictly proper scoring rule for multiple quantiles (Grushka-Cockayne et al. 2017), i.e. it incentivizes reporting of true beliefs, and

(2) the pinball loss function has an economic interpretation. The newsvendor problem, which we will use to make resourcing decisions in Section 6, is an affine transformation of the pinball loss (Jose and Winkler 2009). The lower the pinball loss of the quantile forecasts, the higher the payoff in a newsvendor problem.

The pinball loss is also closely related to the Continuous Ranked Probability Score (CRPS), which is used for density forecasts. As the number of quantiles goes to infinity, the limit of the pinball loss summed over all quantiles is equal to the CRPS (Grushka-Cockayne et al. 2017). Compared to the CRPS, however, the pinball loss has fewer assumptions and is easier to implement (Hong et al. 2016). The logarithmic score, an alternative scoring rule that also works in the density space, may return infinity scores when the realization is outside of the predicted distribution's support, making it difficult to evaluate the average performance of a model.

In this study, we score the median and the 0.05, 0.25, 0.75, and 0.95 quantiles as they are commonly used for calculating confidence intervals, p-values and interquartile ranges, as well as plotting standard boxplots. Moreover, these five quantiles are often considered as critical quantiles and are reasonably spaced out to represent the entire distribution.

We also check the calibration of the distributional forecast, i.e. the consistency between the distributional forecasts and the observations (Gneiting and Raftery 2007), using the probability integral transform (PIT). The PIT of a distributional forecast's F is $p = F(y_i)$, the cumulative distribution function (cdf) F evaluated at the realization y_i. If the predicted distribution is consistent with observations, the PIT should follow a uniform distribution. If the predicted distribution is overconfident or underconfident, the density of its PIT is bathtub-shaped or hump-shaped, respectively. A bathtub-shaped density indicates that the predicted distribution is too narrow, with too many realizations falling farther out in the distribution's tails than expected; a hump-shaped density indicates that the predicted distribution is too wide, with too few realizations falling out far enough (Lichtendahl et al. 2013).

Once we have estimated the distribution of Δt, we can evaluate the model's ability to identify late passengers, using the logarithmic loss (log loss) and brier score to test the probability of a passenger being late. The log loss is calculated as $-(y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$, and the brier score is calculated as $(p_i - y_i)^2$, where y_i is a binary indicator of the ith realization, and p_i is the model's predicted probability of a passenger being late.

The point forecast of connection times and passenger flows we score is the mean, and we use the root mean squared error (RMSE) to measure its accuracy, defined as $\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n}$, where y_i is the realization, $\hat{y}_i$ is the prediction, and n is the number of predictions. The mean is the optimal point forecast under RMSE. If the mean absolute error or mean absolute percentage error is preferred as the primary accuracy measure, one could instead use the median or mode as the point forecast (Gneiting 2011, Jose 2016).

3.4. Phase One: The Regression Tree Model for Connection Times

In phase one, we estimate the distribution of passengers' connection times, Δt, using a regression tree. We fit the tree by minimizing the mean squared error, while the tuning parameters are selected by minimizing the pinball loss. The simplicity of the regression tree approach is not only useful for rapid prediction, but can also yield intuitive explanations as to why observations are predicted in a particular manner. An overview of the steps taken in Section 3.4 and 3.5 is shown in Figure A.1 in the Appendix.

Regression trees are widely used for generating point forecasts (James et al. 2013); however, relatively few studies have applied them to produce entire distributions. We use regression trees to generate distributional forecasts as follows. After fitting a tree to the logarithmic transform of the connection times,[1] each of the leaves in the tree represents a passenger segment, containing a number of observations. We find that the majority of the leaves have connection times, in their original scale, that follow right-skewed distributions. Fitting distributions that can capture right skewness, such as the log-normal and gamma distributions, however, produce underconfident forecasts in the test set, with a hump-shaped PIT. Pure empirical distributions, on the other hand, produce overconfident forecasts, because the distribution is limited to the range of data in the training set. Instead, we apply kernel density estimation (Wand and Jones 1994) to approximate the empirical distribution of the connection times within each leaf, which generates a continuous version of the empirical distribution as a compromise between a parametric distribution and a pure empirical distribution.

Given a connecting passenger's information, the regression tree will first determine which leaf (or segment) this passenger belongs to; the empirical distribution attached to this leaf will then provide the quantile forecasts of this passenger's connection time, and the probability of this passenger being late for his onward flight.

To avoid overfitting to the training set, the maximum depth of the tree, N_{max}, and the minimum number of observations, l_{min}, in each leaf are tuned using a two-fold cross validation approach.[2] The tree is fit — for a range of values of the two parameters ($5 <= N_{max} <= 25$ with an increment of 1; $50 <= l_{min} <= 1000$ with an increment of 50) — to half of the data in the training set, and the pinball loss of the 0.05, 0.25, 0.50, 0.75 and 0.95 quantiles are computed and then averaged in the remaining half. This is done in turn for each fold, and the two pinball scores are averaged. The results of the grid search suggest the optimal N_{max} and l_{min} should be 14 and 200, respectively. The tree with these optimal tuning parameters partitions the transfer passengers in our training set into 2,569 segments, with branches between 8 and 14 levels, and segments containing between 200 and 6,215 passengers.

[1] We found that the distribution of the connection times is highly right skewed, and therefore transforming the data may yield better accuracy results, as suggested in De'ath and Fabricius (2000).

[2] We also ran a 5-fold cross validation, with results ending up similar. With 5-fold, we obtained $N_{max} = 15$ and $l_{min} = 200$.

3.5. Phase Two: Distributional Forecasts for the Number of Arrivals at Immigration and Security

In phase two, we use the leaf distributions from the regression tree to generate predictions for the number of passengers arriving at immigration and security. This will help with predicting congestion in these areas and reallocating resources in advance of such congestion in order to improve the passenger experience. A passenger's arrival at the conformance desk during the ith time interval $t_{iu} - t_{il}$, where t_{iu} and t_{il} are the upper and lower bounds of the time interval, can be considered as a Bernoulli process. Let random variable X_{ij} denote whether or not the jth passenger arrives during this ith time interval. Then, $X_{ij} \sim Bernoulli(p_{ij})$, where p_{ij} is the probability of this passenger arriving during the time interval. Each X_{ij} may have different probabilities p_{ij} due to the segment (leaf of the tree) a passenger is classified into. The number of passengers arriving during the ith interval, $S_i(n)$, can then be calculated as $X_{i1} + X_{i2} + ...X_{in}$, where n is the total number of connecting passengers. If X_{ij} are independently distributed, then $S_i(n)$ follows a Poisson Binomial distribution, or a normal distribution if n is large enough. Then, the mean $\hat{\mu}_{S_i(n)}$ and variance $\hat{\sigma}^2_{S_i(n)}$ of $S_i(n)$ are $\sum_{j=1}^{n} p_{ij}$ and $\sum_{j=1}^{n} p_{ij}(1-p_{ij})$, respectively. The probability of arriving during the ith time interval, p_{ij}, can be calculated as $F_j(t_{iu} - t_{arrj}) - F_j(t_{il} - t_{arrj})$, where t_{arrj} is passenger j's arrival time at the airport, and F_j is the cdf of the predicted distribution of passenger j's connection time.

As not all passengers travel independently but often in groups, it is important to consider dependences between passengers' connection times and arrivals at the conformance desk. In that case, the mean of the distribution for passenger flows stays the same as above, and the variance of the distribution becomes $\sum_{j=1}^{n} p_{ij}(1-p_{ij}) + 2\sum_{i=1}^{n}\sum_{l=i}^{n} Cov(X_{ij}, X_{il})$, with $Cov(X_{ij}, X_{il})$ capturing the fact that the arrivals of passengers travelling together are correlated. Incorporating correlations between passenger arrivals is also a convenient way of recalibrating predictions: if predictions assuming independent arrivals are overconfident, i.e., the prediction intervals are too narrow, incorporating positive correlations among passenger arrivals will make the predictions well-calibrated. If the predictions are underconfident, i.e., the prediction intervals are too wide, negative correlations among passenger arrivals will again make the predictions well-calibrated.

Although we can still derive the mean and variance of the distribution for passenger flows, incorporating correlated arrivals makes it difficult to derive analytical expressions for the entire distribution and its quantiles. Instead, we run simulations to generate distributional forecasts or quantile forecasts, with the X_{ij} correlated through Gaussian copulas, a multivariate cdf defined as $C_R(\mathbf{U}) = \Phi_R(\Phi^{-1}(U_1), \Phi^{-1}(U_2), ..., \Phi^{-1}(U_n))$, where Φ^{-1} is the inverse cdf of a standard normal distribution, and Φ_R is the joint cdf of a multivariate normal distribution with zero mean vector and covariance matrix $\mathbf{R}$. The term $\mathbf{U} = (U_1, U_2, ..., U_n)$ is a vector of uniformly distributed random variables, with

$(U_1, U_2, ..., U_d) = (B_{i1}(X_{i1}), B_{i2}(X_{i2}), ..., B_{in}(X_{in}))$, where B_{ij} is the cdf of the Bernoulli distribution with parameter p_{ij}, the probability of passenger j arriving during the ith time interval.

Unfortunately, our data does not show which passengers were travelling together. Therefore, we assume that passengers arriving on the same flight are correlated with each other in a homogenous way, using ρ. We also assume the same correlation parameter ρ for all flights. As a result, the off-diagonal elements in $\mathbf{R}$ are all equal to ρ if the passengers arrive on the same flight, and zero otherwise. The correlation coefficient ρ can be viewed as a tuning parameter, chosen using a grid search to minimize the average pinball loss using the two-fold validation process.

To compute the distribution of passenger flows, we use the following procedure: First, in each iteration of the simulations for a time interval i, we randomly generate a uniformly distributed vector between 0 and 1, $(U_1, U_2, ..., U_d)$, from the Gaussian copula. Next, to obtain the binary variables $(X_{i1}, X_{i2}, ..., X_{in})$, we apply the quantile function (or the inverse cdf) of their corresponding Bernoulli distributions to the vector of probabilities generated in the first step. Finally, for time interval i, we calculate the number of passenger arrivals by summing $(X_{i1}, X_{i2}, ..., X_{in})$. The distribution of the passenger flow is then approximated by the empirical distribution constructed by the number of arrivals obtained from the simulations.

4. Results

Next we present results derived from our predictive model. We first report on the accuracy of our model on the 20% test set in predicting connection times and passenger flows compared against several benchmarks. Next, we highlight a few key findings from the regression tree model.

4.1. Accuracy of the Distributional Forecasts for Individuals' Connection Times

We first compare the accuracy of our regression tree method in predicting the 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles of individuals' connection times. Among the methods that can generate distributional forecasts, four are widely used by the machine learning community: linear regression, quantile regression, quantile regression forests, and gradient boosting machines (Hong et al. 2016). Quantile regression and gradient boosting machines estimate different quantiles independently, possibly resulting in a lack of monotonicity in the estimated quantile function. This longstanding problem is also known as the quantile crossing problem (Bassett and Koenker 1982). Chernozhukov et al. (2010) propose a method of rearranging the curve into a monotone curve. This, however, requires the estimates of thousands of quantile regressions, making the method computationally expensive. Linear regression and quantile regression forests, on the other hand, are able to produce monotone quantiles. Linear regression is also easy to fit and usually runs fast. However, it does not perform well with nonlinear relationships and complex interactions. Quantile regression forests are a generalization of the random forests and are competitive in terms of predictive power (Meinshausen 2006). They are, however, time consuming and typically treated as a black box.

We compare our regression tree model with these four methods and a naïve forecast on the test set. The naïve model predicts passengers' connection times based on their arrival terminals, the most important predictor given by our regression tree. Specifically, given a new passenger's arrival terminal, the quantiles of her connection time are predicted as the quantiles of the connection times of all passengers arriving at the same terminal in the training set. The linear regression and the quantile regression are fit to all 17 variables that were selected as predictors in Section 3.2 and their interactions. We fit these two models to the data, with the goal of predicting the logarithmic transformation of the connection times. For the linear regression, the distribution of a passenger's connection time is predicted as a log-normal distribution with the mean and standard deviation set to the point forecast, and the standard error of the forecast from the linear regression, respectively. The standard error of the forecast generated from a linear regression is typically calculated as the square root of the residuals' variance plus the variance contributed by the regression coefficients. For the quantile regression, the exponential of the predictions generated from the model are the quantile forecasts of connection times. We also found that the LASSO penalized quantile regression (Friedman et al. 2010) produces more accurate forecasts than the quantile regression without penalty terms. The LASSO linear regression is not considered here because it is not typically used for producing prediction intervals due to the difficulties in estimating standard errors of the coefficients, which are part of the uncertainties captured by prediction intervals (Goeman 2010, Goeman et al. 2018).

The other two methods, quantile regression forests and gradient boosting machines are tree-based models. A quantile regression forest fits independent trees and constructs conditional distributions from these trees. Gradient boosting is an ensemble technique in which the regression trees are not fit independently, but sequentially, learning from one tree to the next. Since the quantile regression forest is extremely time consuming and requires processing with a large amount of memory, we fit the model to the seven key factors identified by our regression tree to balance accuracy and computational cost. Details of these factors will be discussed in Section 4.3. For the gradient boosting machine, we use all 17 variables as predictors. We tune the number of trees for both models to avoid overfitting. The quantile regression forest model is fitted to the log transform of connection times. The gradient boosting machine is the only model that is fitted to the original connection times, as we found the log transform of the target variable makes the model performs worse than without the transformation.

As shown in Table 1, among all six models, our regression tree has the lowest out-of-sample average pinball loss (2.74) and is best on four of the five quantiles in the test set. Not surprisingly, the naïve model performs the worst with an average pinball loss of 3.29. The quantile regression forest, which is often considered as an advanced machine learning method with high accuracy, performs worse than the third best model, the LASSO penalized quantile regression. The performance of the quantile

Table 1 Accuracy of forecasts on connection times in the test set

	RMSE of mean	Pinball losses					
		$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	Avg.
Naïve model	15.93	0.92	3.40	4.90	4.73	2.48	3.29
Linear regression	14.12	0.82	2.79	4.12	4.24	2.24	2.84
LASSO quantile regression	14.10	0.79	2.80	4.09	4.03	2.10	2.76
Quantile regression forest	14.13	0.81	2.83	4.12	4.07	2.16	2.80
Gradient boosting machine	14.20	0.79	2.82	4.08	**4.00**	2.09	2.76
Regression tree	**14.00***	**0.77*****	**2.77*****	**4.07***	4.02	**2.08**	**2.74*****

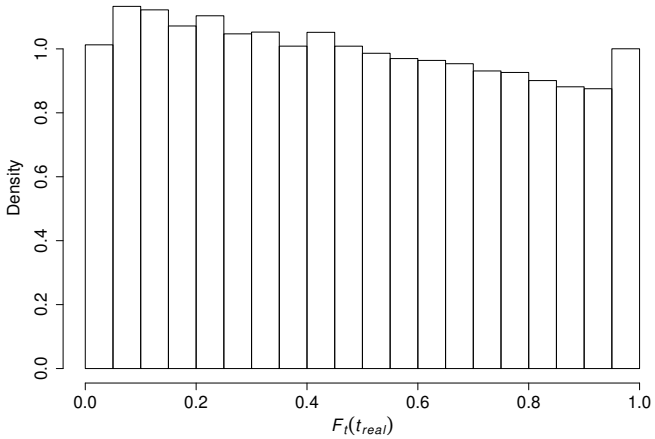
Values in bold indicate the lowest errors.
The symbol *** and * indicate the difference between the regression tree model and the second best model in each column is significant based on a t-test at the 1% and 10% level, respectively.

regression forest could be due to the fact that only seven variables were used to train the model to reduce computing time.

The predictions generated from our regression tree model are also well-calibrated, evidenced by a roughly uniform empirical density of the PIT (Figure 2), the density of the predicted cdf F_t for passengers' connection times evaluated at the realizations t_{real}. The accuracies of the predicted distributions evaluated by other performance measures (the CRPS and the log scores) can be found in Tablel A.4 in the Appendix. In terms of point forecasting, our regression tree model has the lowest RMSE. The mean is used as point forecast for the naïve model, linear regression, quantile regression forest and our regression tree model. The 0.5-quantile is used as point forecasts for LASSO quantile regression and gradient boosting machine as these methods are designed to predict quantiles.

Among all six models in the Table 1, only linear regression and our regression tree model can generate forecasts of the entire distribution and therefore can be used in our two-phased approach.

Figure 2 Empirical PIT densities from the distributional forecasts of connection times generated by the regression tree



The current implementation of quantile regression forests in leading machine learning tools does not provide complete leaf-distributions as outputs. For the gradient boosting machine and the quantile regression, thousands of models need to be fit and stored to construct an empirical distribution and simulate from it, making the model-training process inefficient. The crossing problem inherent to these two models may also make it difficult to construct empirical distributions.

Since we use the predictions of passengers' connection times to identify late transfer passengers, we also compared the accuracy of our regression tree model with several benchmarks in the context of a classification problem. We first calculate the time difference between a passenger's scheduled departure time and their arrival time at the conformance desk. A passenger is considered to be late if this time difference is less than 30 minutes. Table 2 shows the out-of-sample average log loss and brier score of our regression tree model and benchmark models on predicting probabilities of being late. The benchmark models considered include the naïve model, linear regression, LASSO Logistic regression, random forest with binary target variable, gradient boosting machine with binary target variable, and classification tree. The naïve forecast is defined as the proportion of late passengers in the training set.

Among all seven models in Table 2, the classification tree is the most accurate in log loss, while the regression tree that predicts the distribution of passengers' connection times is the most accurate according to the brier score. We note that the differences between our method and the classification tree are not statistically significant. Given the results above, our regression tree model performs favourably in both identifying late passengers and predicting distributions of connection times, making it the preferred model to be used in the predictive system's first phase.

4.2. Accuracy of Aggregate Forecasts for Arrivals at Immigration and Security

Based on the predicted number of passenger arrivals, airports can decide how many immigration desks and security lanes to open throughout the day. These two decisions are based on the passenger flows into the immigration and security areas, respectively. As discussed in Section 2.1, only passengers

Table 2 Accuracy of forecasts of the probability of being late in the test set

	Log Loss	Brier Score
Naïve model	0.3030	0.07092
Linear regression	0.0396	0.00971
LASSO logistic regression	0.0436	0.01143
Random forest	0.0585	0.00965
Gradient boosting machine	0.0311	0.00760
Classification tree	**0.0295**	0.00742
Regression tree	0.0304	**0.00736**

Values in bold indicate the lowest errors.
The difference between the best and second best model in each column is not statistically significant.

connecting to domestic destinations need to go though immigration desks. Therefore, we generate two sets of forecasts of the passenger flows, one for those connecting to domestic destinations, and the other for those connecting to international destinations. Since Heathrow's security resources are planned in 15 minutes intervals, we use same windows to construct passenger flow patterns. In Section 6 we will examine the implication of the quality of the forecasts on the resourcing decisions. To generate forecasts from our two-phased approach, we use the two-fold cross validation approach to search for the correlation coefficient ρ of the Gaussian copula,[3] with the goal of of minimizing the average pinball loss over five quantiles during business hours of 5:00 am to 10:00 pm daily.

We construct a naïve model as a benchmark to predict passenger flows. For each day in the training set, we calculate the number of passengers connecting to domestic (international) destinations that arrive at the immigration (security) areas during each 15-minute interval. These numbers of passengers across all days in the training set are then used to construct the naïve distribution of the domestic (international) passenger flows.

Univariate time series models have long been used to forecast short-term passenger or customer arrivals (Taylor 2008, Wei and Chen 2012). Among these models, the Seasonal Autoregressive Integrated Moving Average Model (SARIMA) emerges as a benchmark since it can handle complex seasonality and is often accurate in forecasting arrivals (Taylor 2012, Hyndman and Athanasopoulos 2018).[4] We fit a SARIMA model to the time series of the number of arrivals at the conformance desk, with a frequency of 15 minutes, i.e. with 96 periods for a 24-hour day. We consider the following variables calculated for the current period (t), and for the previous three 15-minute periods ($t-1$, $t-2$ and $t-3$), as covariates in the SARIMA model: (1) the number of passengers arriving at the airport and traveling to domestic destinations, (2) number of passengers arriving at the airport and traveling to international destinations, (3) number of passengers arriving at Terminal 5, and (4) the number of passengers arriving in economy class.We consider only the previous three 15-minute periods as 90% of the passengers in the data set spent less than 45 minutes to make their connections.

Logarithmic and square root transformations are often applied to time series to stabilize count data and enforce positive predictions (Taylor 2012). We apply square root transformation in our study as the data contains many zeros. To identify seasonality in the data, we inspected the autocorrelation function (ACF, Venables and Ripley 2002) and the periodogram (Bloomfied 2004). Only daily seasonality (hour of the day) can be easily observed from the results. Although we did not find evidence of weekly seasonality (day of the week), we also tested models with weekly seasonality since it was reported to be an important factor on passenger delays in Barnhart et al. (2014).

[3] We also tested 5-fold approach, and the optimal ρ ends up the same.

[4] We also experimented with other time series models, including the Exponential Smoothing State Space Model with Trigonometric Seasonality, Box-Cox Transformation, ARMA Errors, and Trend And Seasonal Components (TBATS). All these models underperformed compared to the SARIMA model.

The SARIMA model is fit to the first 80% of the days of data, and used thereafter to produce one-period-ahead forecasts for the remaining 20%. When we predict the number of passengers arriving at the conformance desk during the interval $(t, t+1]$, we use the actual number of arrivals at the conformance desk up to time t, and assume the information of passengers arriving at the airport during $(t, t+1]$ is also available so we can calculate the covariates listed above. The final SARIMA model is SARIMA$(5,0,1)(0,1,0)^{96}$.

Two benchmark models are included to simulate Heathrow's legacy systems. In the legacy systems, individual transfer passengers' information was not used. Rather, the number of passengers transferring to domestic (or international) destinations on an arrival flight was estimated as the total number of passengers on the flight, which is known by the airport, multiplied by the estimated percentage of passengers transferring to domestic (or international) destinations on similar flights in the training data. The estimated percentage for an arrival flight was calculated as the percentage of passengers transferring to domestic (or international) destinations on the flights in the training set that arrived from the same region, at the same terminal, and run by the same carrier. Passengers arriving on the same flight were assumed to have the same connection time. The distribution of their connection time was estimated as the empirical distribution of the connection times of passengers arriving at the same terminal, and during the same hour of the day in the training set.

To compute distributions of passenger flows, we use the same simulation approach described in Section 3.5. We test two versions of Heathrow's legacy systems: a static version, which uses flights' scheduled arrival times to approximate passengers' arrival times at the airport, and a dynamic version, which uses their actual arrival times at the airport. It should also be noted that Heathrow's legacy systems only generated point forecasts. Here, however, we generalize their methods to produce also distributional forecasts.

Next we measure the accuracy of the models in predicting passenger flows between 5:00 am and 10:00 pm. The point forecasting errors and pinball losses are presented in Table 3 and Table 4. The

Table 3 Accuracy of the predicted flow of passengers connecting to domestic destinations in the test set

	RMSE	Pinball losses					
	of mean	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	Avg.
Naïve model	14.97	1.07	3.82	5.45	4.91	1.86	3.42
SARIMA with covariates	13.46	1.06	3.51	4.72	4.15	1.57	3.00
Static legacy system	15.50	1.03	3.65	5.19	4.94	2.58	3.48
Dynamic legacy system	12.25	0.82	2.90	4.01	3.63	1.56	2.59
Linear regression with copula	7.66	0.62	2.02	2.58	2.10	0.72	1.61
Regression tree without copula	7.42	0.97	2.06	2.45	2.11	1.07	1.73
Regression tree with copula	**7.42***	**0.60****	**1.91*****	**2.44*****	**2.01****	**0.71**	**1.53*****

Values in bold indicate the lowest errors.
The symbol ***, **, and * indicate the difference between the regression tree model with copula and the second best model in each column, excluding the regression tree model without copula, is significant based on a t-test at the 1% , 5% and 10% level, respectively.

Table 4 Accuracy of the predicted flow of passengers connecting to international destinations in the test set

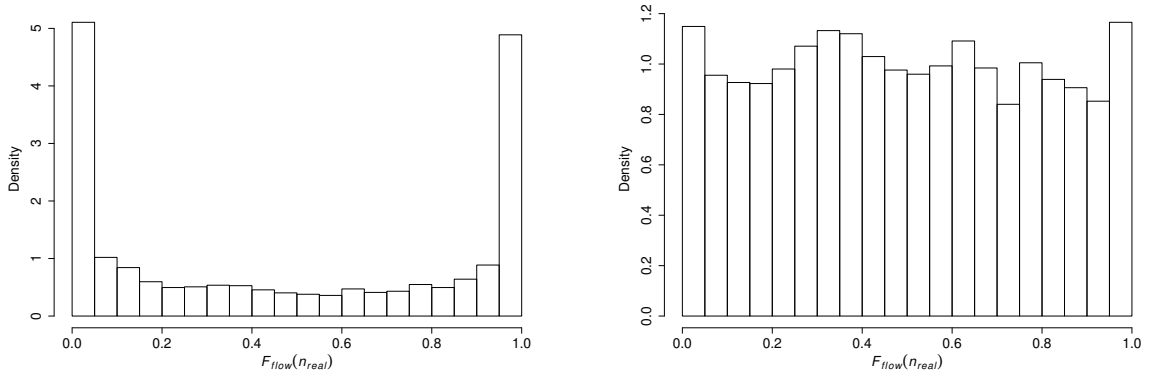
	RMSE	Pinball losses					
	of mean	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.95}$	Avg.
Naïve model	53.09	4.77	15.63	19.83	15.69	5.16	12.22
SARIMA with covariates	39.40	3.42	11.16	14.82	12.56	4.56	9.30
Static legacy system	53.37	5.17	16.16	20.81	17.28	5.84	13.05
Dynamic legacy system	41.66	4.61	13.45	16.43	13.29	4.58	10.47
Linear regression with copula	**22.33**	2.13	6.43	8.05	6.70	2.38	5.14
Regression tree without copula	22.41	3.79	6.83	7.94	7.02	4.14	5.94
Regression tree with copula	22.41	**2.04****	**6.21****	**7.93***	**6.60**	**2.35**	**5.03***

Values in bold indicate the lowest errors.
The symbol ** and * indicate the difference between the regression tree model with copula and the second best model in each column, excluding the regression tree model without copula, is significant based on a t-test at the 5% and 10% level, respectively.

results in Table 3 are for domestic destinations, and the results in Table 4 are for international destinations. The last three models in the table apply the two-phased approach using linear regression and regression tree as their first phase model to predict connection times. These models are much more accurate in both quantile forecasting and point forecasting than the other models. Our regression tree model with copula-based simulations outperforms the other models on all five quantiles. It is also the most accurate in generating point forecasts for the domestic passenger flow. The two-phased approach with linear regression as the first phase model performs slightly better in point forecasting of the international passenger flow. The difference between the two RMSEs, however, is not statistically significant.

The passenger flows predicted by the regression tree using independent conformance desk arrivals (i.e. without using copulas) are overconfident. This is also the case when connection times are predicted by the linear regression. Figure 3a shows the PIT density for the forecasts generated from our

Figure 3 Empirical PIT densities from the distributional forecasts of the flow of passengers connecting to international destinations generated by our model



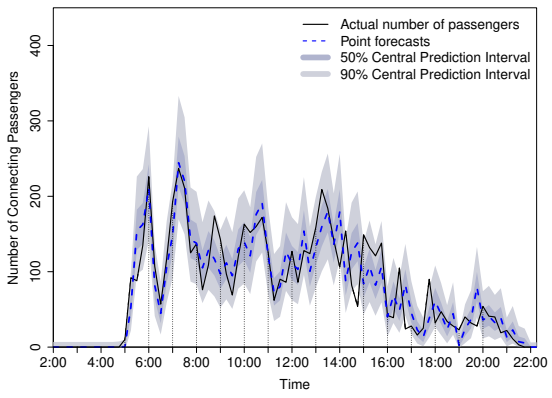
(a) Regression tree with independent simulations

(b) Regression tree with copula-based simulations

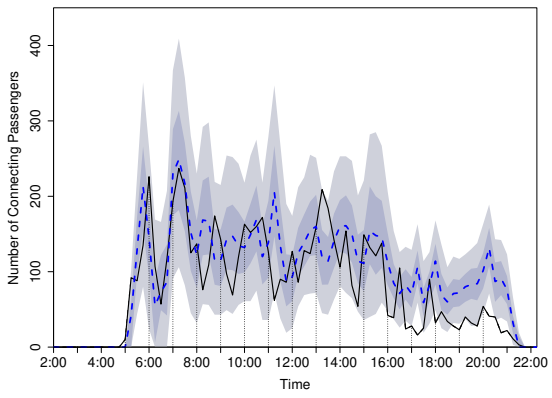
model without using copulas for the flow of passengers connecting to international destinations. As shown in the figure, too many realizations fall in the tails of the predicted cdf. Incorporating dependences between passengers' arrivals using Gaussian copula results in a more uniform PIT density (Figure 3b). The PIT densities for the flow of passengers connecting to domestic destinations exhibit similar patterns.

Figure 4 presents the prediction intervals for the flow of passengers connecting to international destinations on a randomly selected day from our test set. These prediction intervals are generated by the SARIMA model with covariates (Figure 4a), the dynamic legacy model with copula-based simulations (Figure 4b), and the regression tree model using independent and copula-based simulations, Figure 4c and Figure 4d, respectively. It is easy to observe the accuracy of the forecasts from the two-phased models. The prediction intervals produced when assuming independence (Figure 4c) are clearly much narrower than when using copula-based simulations (Figure 4d). Such narrow intervals result in overconfidence and misjudged decisions regarding peak activity. Similar patterns hold when

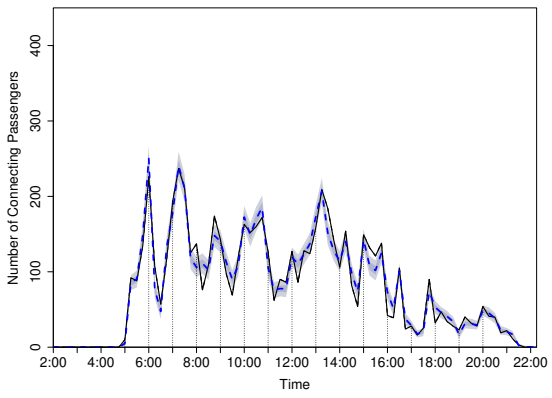
Figure 4 Forecasts and Actuals of the flow of passengers connecting to international destinations at the conformance desk, on a random test day.



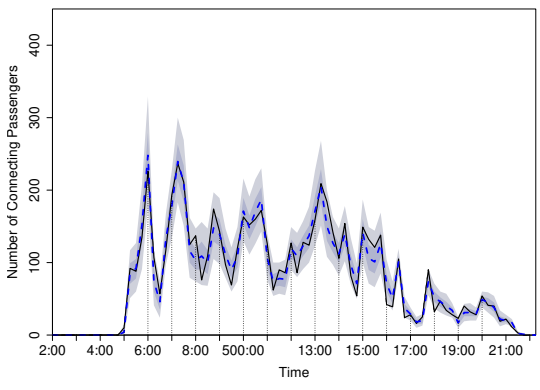
(a) SARIMA model with covariates



(b) Dynamic legacy system



(c) Regression tree with independent simulations (PIT shown in Figure 3a)



(d) Regression tree with Copula-based simulations (PIT shown in Figure 3b)

using alternative scoring rules. The CRPS and log scores of the models' distributional forecasts are shown in Tablel A.5 in the Appendix.

4.3. Key Findings from the Model

As the tuned regression tree model has more than 2,000 leaves, we use a pruned tree for deriving insights. The pruned version of the tree partitions the passengers into only a few segments, making it easier to explain why a passenger's connection time is predicted in a particular manner.

We find when the tree is fit to different subsamples of the data, and with different settings of the tuning parameters, the pruned tree, which is essentially the first four levels of the full tree, always relies on the same features, and the sequences of these features chosen to be split are the same. We also find that the coefficients of variation of the variables' cutting points that appear in the first four levels are within a reasonable level (5%). In addition, the pruned tree is visually manageable so that we can easily derive insights from it. When generating forecasts from our predictive system, we use the full tree because of its accuracy. The pruned tree was only applied to generate insights.[5]

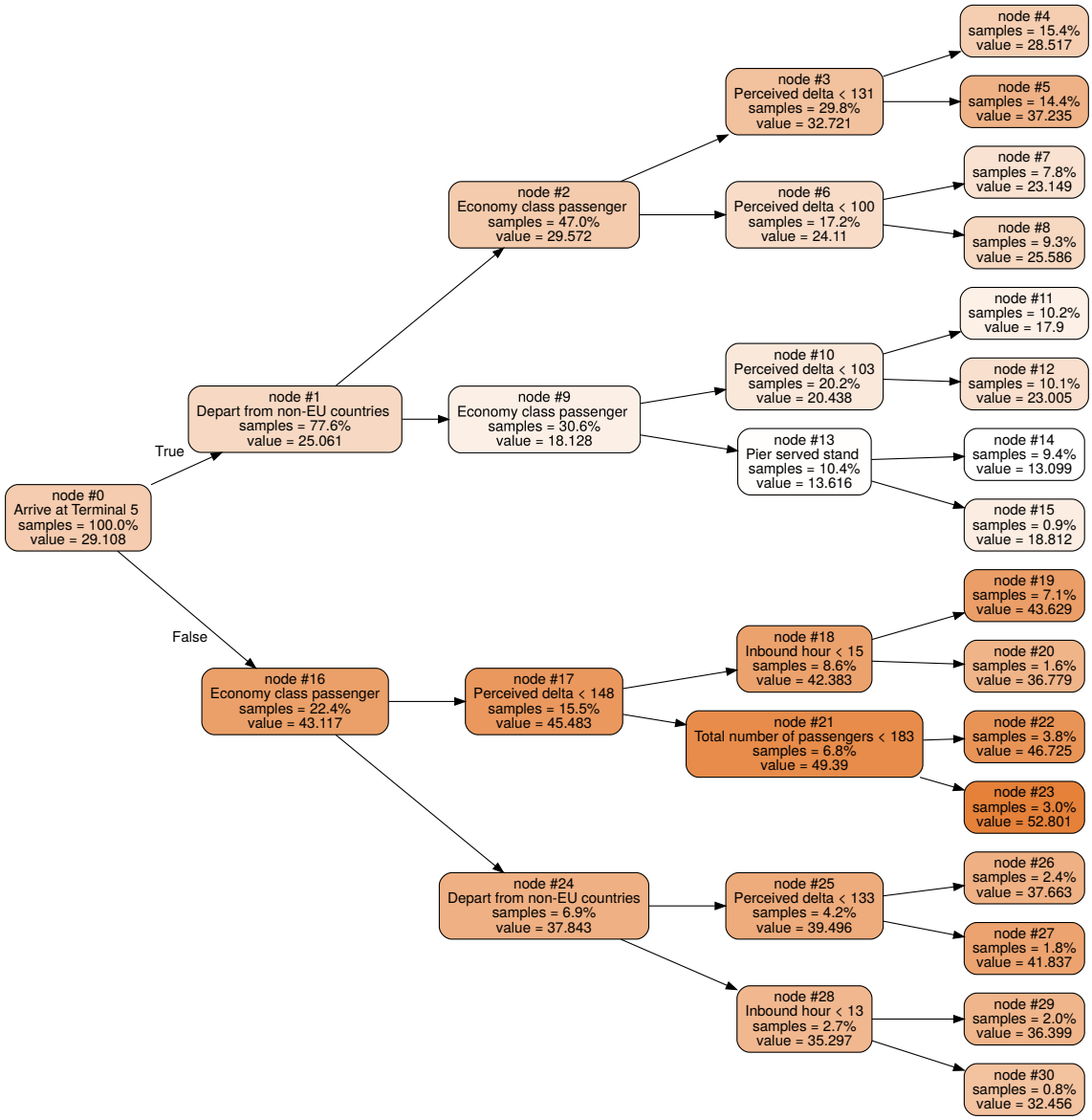
The first four levels of the tree have partitioned the passengers into 16 segments. Figure 5 visualizes the pruned tree trained to the entire training set. Each node of the tree presented in the figure gives information on the name of the feature split in this node, percentage of passengers falling in this node, and the average connection time of these passengers. A summary of the 16 segments, represented by the 16 leaf nodes in Figure 5, is provided in Table A.6 in the Appendix. The empirical distributions estimated by kernel smoothing for leaves are shown in Figure 6. Using the results summarized in the table and the figures, we highlight three key findings regarding passengers' connection times.

Key Finding 1: Key Factors The key factors that impact passengers' connection times are (1) whether or not the passenger arrives at Terminal 5, (2) whether or not the passenger arrives from an European Union (EU) country, (3) perceived connection time, (4) whether or not the passenger is in economy class, (5) hour of the day the arriving flight lands at the airport, (6) arriving flight's stand type (pier serviced or remote stand), and (7) the total number of passengers on the arriving flight. Although our predictive model contains 17 predictors, the first four levels of the tree only use seven of them. Therefore, we treat only these seven features listed above as the key factors that impact passengers' connection times.

Key Finding 2: Expected Connection Times Passengers arriving at Terminal 5, from EU countries, in business or first class, whose arriving flight parks at a pier-served stand, take the shortest amount of time to make their connections (Segment 14 in Figure 5 and 6). We also find that passengers arriving at Terminal 5 and departing from EU countries have shorter connection times

[5] The average pinball loss of the reduced version of the tree on the test set increases from 2.74 to 2.86. The RMSE for the point forecasts increases from 14.0 to 14.4.

Figure 5 The first four levels of the tree trained to the entire training set



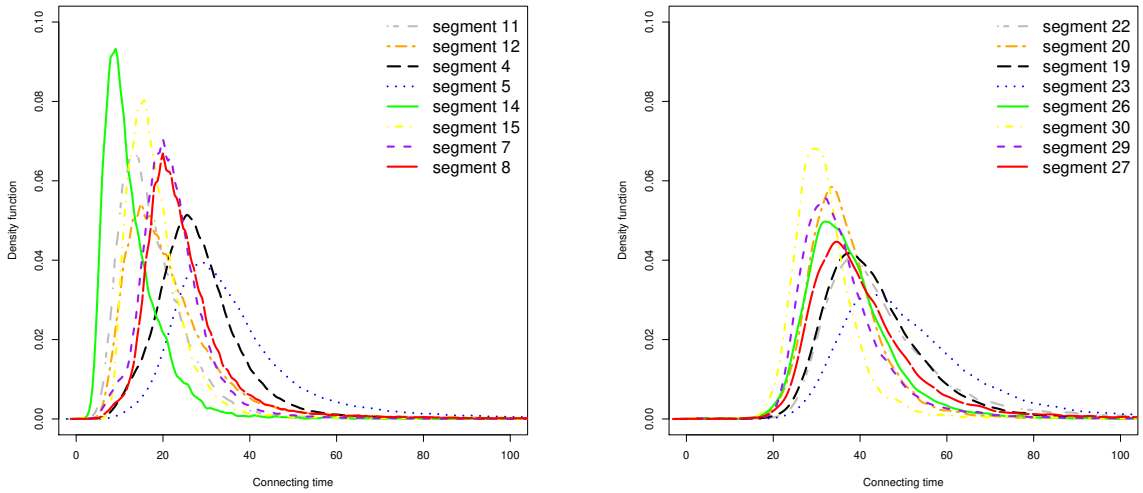
The value shown in each node (or segment) is the mean of the connection times.

compared with other passengers. This is regardless of the value of the rest predictors. According to Table A.6, these passengers form the first four segments that have the lowest connection times.

Passengers arriving at Terminal 2, 3 or 4, in economy class, with perceived connection time greater than 148 minutes, whose arriving flight has more than 183 passengers, take the longest time to make their connections (Segment 23 in Figure 5 and 6).

Key Finding 3: Uncertainty of Connection Times In general, the uncertainty and average value of passengers' connection times within a segment are positively correlated. For most segments shown in Figure 6b, not only are the average connection times longer, but the variances among the passengers are also larger compared to those in Figure 6a. The uncertainty, however, does not

Figure 6 Smoothed empirical distributions of the connection times in the 16 segments[a]



(a) Segments with low average connection times

(b) Segments with high average connection times

[a] The numbered nodes in the figures are the leaf nodes shown in Figure 5.

increase by as much as the average, which can be seen from the coefficient of variation, defined as the ratio of the standard deviation to the average connection time. Therefore, when taking the average connection time into account, the higher the connection time, the more predictable the journey.

5. Real-Time Implementation

With the predictive model at hand, we worked with Heathrow's APOC to develop an application for forecasting individual connection times, and the flow of transfer passengers into the immigration and security areas in real time. We use these predictions to also calculate the probability of a passenger missing her connection, and the expected number of late passengers for each outbound flight. The application generates forecasts on a rolling basis and updated every 15 minutes. A prototype of the predictive system was first tested at Heathrow in 2016. In order to get buy in from the various APOC stakeholders, the project team spent a day presenting the model to each group and discussing the implications and new possibilities made available by the improved predictions.

5.1. Real-Time Input Data

Real-time flight-level information was exported from the Airport Flight Operations System, also known as IDAHO. Real-time passenger level information was obtained from Passenger Transfer Message (PTM) files. These files are sent by the airline when a flight takes off from an origin airport. PTM files are currently the only real-time data source that contains passenger level information. How far in advance of the flight's arrival these files get sent by the airline has a significant impact on the accuracy of our predictions; for instance, missing PTMs can cause forecasts to underestimate

the passenger flow. According to Heathrow, PTMs for 88% of the passengers are received more than 90 minutes prior to arrival. Therefore, to ensure that there is sufficient data to enable the model to provide accurate forecasts, the forecasting window chosen for the application was 90 minutes.

It should also be noted that the actual on-chock time, the time when an aircraft is parked at gate, will not be available if an aircraft is en-route. In this case, we used estimated on-chock time. If we had neither of the actual or estimated on-chock time, we used the actual time of ground handling, the estimated time of ground handling, or the scheduled time of ground handling. If none of these five fields was presented in IDAHO, we dropped the record.

A few variables in our historical data set were not available in real-time. Some of these variables, such as stand number and runway number which may indicate the distance between gate and transfer bus station, could be significant predictors in predicting connection times. Once these variables become available in the future, the predictive model should be retrained and reassessed. Other variables that are currently important, such as stand type, may become less significant predictors.

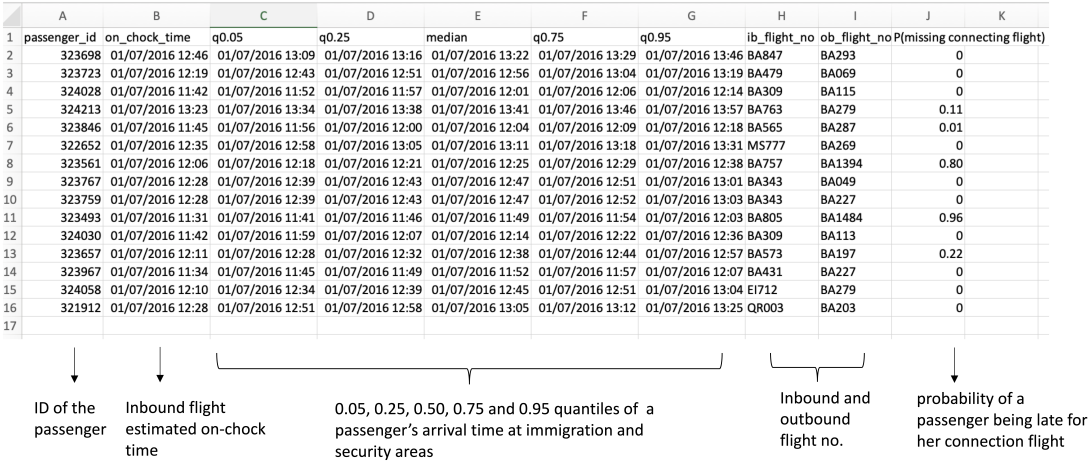
5.2. Forecasting Application

A prototype for real-time forecasting of passenger connection times was developed using a `Python` GUI scripting interface. Forecasts were generated every 15 minutes, on a rolling basis, for the next 90-minute time window. At the start of each iteration, the application collected real-time information of passengers who have arrived in the previous 150 minutes or will arrive in the next 90 minutes. The application ran for approximately three minutes on a typical Heathrow machine, and generated as output several CSV files containing individual level and aggregate level forecasts. Plots were also generated to visualize the forecasts of transfer passenger flows.

In the application, users can set the granularity of the passenger flow and the forecasted time-horizon. Forecasts with different granularities are generated for different purposes. For example, the forecasts of passengers arriving in every 5 minutes provide detailed flow profiles, while the forecasts of passengers arriving in every 15 minutes can be used to adjust resourcing plans.

Figure 7 shows an example of the output for the forecasts of individual connection times generated at 12:00 on July 1, 2016. Each row in the file represents the forecast for one passenger. In addition to the quantile forecasts of passengers' arrival time at immigration and security areas, this file also contains probabilities of these passengers being late for their connecting flights. Here, a passenger is considered to be late if they arrive at the conformance desk later than 30 minutes before the scheduled departure time of the connecting flight. Based on the predicted probabilities, Heathrow and airlines can easily identify which passengers are at risk of missing their onward flight. Given this information, they would be able to help late passengers move faster through the airport and facilitate early rebooking. The threshold time of rebooking late passengers may vary among airlines. In those cases, managers can easily set different thresholds in the system.

Figure 7 Output from the application: individual connection times

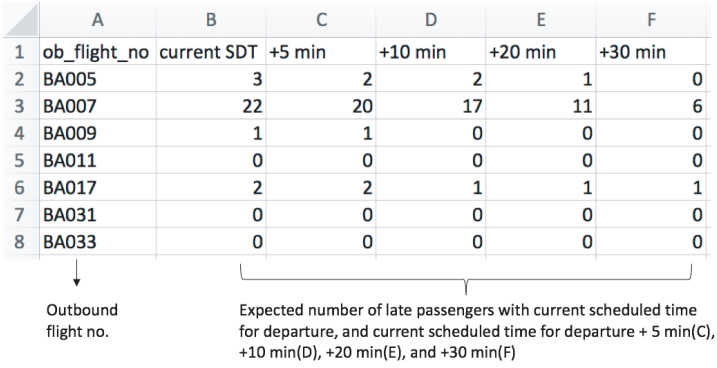


	A	B	C	D	E	F	G	H	I	J	K
1	passenger_id	on_chock_time	q0.05	q0.25	median	q0.75	q0.95	ib_flight_no	ob_flight_no	P(missing connecting flight)	
2	323698	01/07/2016 12:46	01/07/2016 13:09	01/07/2016 13:16	01/07/2016 13:22	01/07/2016 13:29	01/07/2016 13:46	BA847	BA293	0	
3	323723	01/07/2016 12:19	01/07/2016 12:43	01/07/2016 12:51	01/07/2016 12:56	01/07/2016 13:04	01/07/2016 13:19	BA479	BA069	0	
4	324028	01/07/2016 11:42	01/07/2016 11:52	01/07/2016 11:57	01/07/2016 12:01	01/07/2016 12:06	01/07/2016 12:14	BA309	BA115	0	
5	324213	01/07/2016 13:23	01/07/2016 13:34	01/07/2016 13:38	01/07/2016 13:41	01/07/2016 13:46	01/07/2016 13:57	BA763	BA279	0.11	
6	323846	01/07/2016 11:45	01/07/2016 11:56	01/07/2016 12:00	01/07/2016 12:04	01/07/2016 12:09	01/07/2016 12:18	BA565	BA287	0.01	
7	322652	01/07/2016 12:35	01/07/2016 12:58	01/07/2016 13:05	01/07/2016 13:11	01/07/2016 13:18	01/07/2016 13:31	MS777	BA269	0	
8	323561	01/07/2016 12:06	01/07/2016 12:18	01/07/2016 12:21	01/07/2016 12:25	01/07/2016 12:29	01/07/2016 12:38	BA757	BA1394	0.80	
9	323767	01/07/2016 12:28	01/07/2016 12:39	01/07/2016 12:43	01/07/2016 12:47	01/07/2016 12:51	01/07/2016 13:01	BA343	BA049	0	
10	323759	01/07/2016 12:28	01/07/2016 12:39	01/07/2016 12:43	01/07/2016 12:47	01/07/2016 12:52	01/07/2016 13:03	BA343	BA227	0	
11	323493	01/07/2016 11:31	01/07/2016 11:41	01/07/2016 11:46	01/07/2016 11:49	01/07/2016 11:54	01/07/2016 12:03	BA805	BA1484	0.96	
12	324030	01/07/2016 11:42	01/07/2016 11:59	01/07/2016 12:07	01/07/2016 12:14	01/07/2016 12:22	01/07/2016 12:36	BA309	BA113	0	
13	323657	01/07/2016 12:11	01/07/2016 12:28	01/07/2016 12:32	01/07/2016 12:38	01/07/2016 12:44	01/07/2016 12:57	BA573	BA197	0.22	
14	323967	01/07/2016 11:34	01/07/2016 11:45	01/07/2016 11:49	01/07/2016 11:52	01/07/2016 11:57	01/07/2016 12:07	BA431	BA227	0	
15	324058	01/07/2016 12:10	01/07/2016 12:34	01/07/2016 12:39	01/07/2016 12:45	01/07/2016 12:51	01/07/2016 13:04	EI712	BA279	0	
16	321912	01/07/2016 12:28	01/07/2016 12:51	01/07/2016 12:58	01/07/2016 13:05	01/07/2016 13:12	01/07/2016 13:25	QR003	BA203	0	
17											

In the output shown in Figure 8, we grouped passengers by their outbound flights, and calculated how many of them are expected to be late based on the simulation results. We also calculated the expected number of passengers that would be still at risk if the airline delayed the departure time by 5, 10, 20, and 30 minutes. These forecasts can help improve the predictability and stability of outbound flights' scheduled departure time. During the aircraft turn around, its scheduled departure time can be adjusted, but only a few times. Many airlines consider passenger delays when they amend flights' scheduled departure times. In this case, predictions of passenger delays ahead of time can help airlines make accurate adjustments. Moreover, if our predictions are accurate, no further changes are to be expected because of late transfer passengers; and therefore, the stability of the scheduled departure time can be improved.

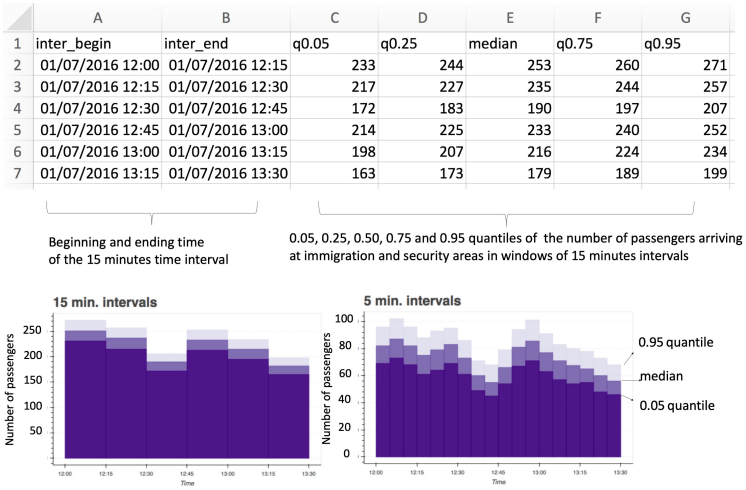
In the output file of passenger flow forecasts (Figure 9), each row represents the forecast for a 15-minute time interval. The 90% prediction intervals and the median of the 5-minute and 15-minute

Figure 8 Output from the application: expected number of late passengers for each outbound flight



	A	B	C	D	E	F
1	ob_flight_no	current SDT	+5 min	+10 min	+20 min	+30 min
2	BA005	3	2	2	1	0
3	BA007	22	20	17	11	6
4	BA009	1	1	0	0	0
5	BA011	0	0	0	0	0
6	BA017	2	2	1	1	1
7	BA031	0	0	0	0	0
8	BA033	0	0	0	0	0

Figure 9 Output from the application: arrivals at immigration and security



	A	B	C	D	E	F	G
1	inter_begin	inter_end	q0.05	q0.25	median	q0.75	q0.95
2	01/07/2016 12:00	01/07/2016 12:15	233	244	253	260	271
3	01/07/2016 12:15	01/07/2016 12:30	217	227	235	244	257
4	01/07/2016 12:30	01/07/2016 12:45	172	183	190	197	207
5	01/07/2016 12:45	01/07/2016 13:00	214	225	233	240	252
6	01/07/2016 13:00	01/07/2016 13:15	198	207	216	224	234
7	01/07/2016 13:15	01/07/2016 13:30	163	173	179	189	199

passenger flows are also visualized in the output figures. These real-time predictions allow the dynamic planning of immigration and transfer security resourcing, by applying the real-time demand forecasts to Heathrow's existing lane planning tool.

5.3. Implementation and Impact

Our predictive system has been implemented at Heathrow since 2017, integrated into the APOC's Dynamic Model of Operations system. The probability of a passenger missing her connection appears in a "Connection at Risk" table, and the passenger flow at the Conformance desks is shown in a "Transfer Security Flow" table. The forecasts of the number of late passengers for each departing flight are provided to APOC's connections team who liaise with the airline.

The predictive system is based on the Azure Machine Learning platform. The regression tree is retrained daily with rolling five years of historical data. On the day of operation, the real-time data is read from Heathrow's SQL server into Azure platform in every 15 minutes. The forecasts for the next 90 minutes are then saved in CSV format to an Azure Blob storage, which is designed to store excessively large quantities of data files.

The Flow Managers in the APOC are currently using the predictive system to optimize the operation for smoother and more predictable service. As commented by Florian Guillermet (Guillermet 2017), the executive director of the Single European Sky Air Traffic Management Research (SESAR) project, "These predictions allow stakeholders – the airport and airlines – to take strategic decisions, such as holding an aircraft at the gate so that delayed passengers can board and the consequences that this may have on traffic." Passenger experience has been improved through reduced queuing, as capacity and resourcing along the journey more closely match with the dynamic demand. In addition,

the managers are able to identify passengers who are at risk of missing their connection, and work with Heathrow and airline teams to assist and expedite their journeys.

An accuracy test of the expected passenger flows has been conducted over July and August 2017. The RMSE of the forecasts generated from the new predictive system is 39.2, which is 25% lower in comparison to Heathrow's previous system. Encouraged by the good results, Heathrow's APOC has expanded the usage of the predictive system to enhance other airport services. Demand forecasts for bags and passengers with restricted mobility have now all been developed using predictive techniques. In addition, real-time prediction of the direct departure flows has been developed to support the optimization of direct security operation. The team is now looking at how these techniques could be applied to other parts of the passenger journey such as surface access and trolley operations.

The encouraging results of our study have also changed the way that the major European airports think about data and machine learning techniques. Robert Graham, the head of airport research at Eurocontrol, stated that "This study has become a reference in SESAR Total Airport Management and is used by a number of major European airports. The study is: (i) groundbreaking, changing the way that the European Air Traffic Management players think about data, data science, and collaborating on sharing data; (ii) demonstrates how decision making can be better informed by the flow of data and the use of predictive algorithms, and (iii) brings state-of-the-art thinking in machine learning, applied to a problem of crucial importance to airports around the world into the airport operations domain."

6. Backtesting: A Numeric Evaluation of the Improved Predictions' Impact

To evaluate the benefits of having more accurate forecasts, we conduct a backtesting study over the 20% test set, examining the optimal staffing decision made when relying on forecasts from the various models. We first formulate the resourcing decision-making problem for the immigration desks.

For simplicity, we assume managers need only one period (or 15 minutes) to reallocate resource, and therefore, decisions on resourcing plans are made dynamically at the beginning of each time period for the next time period. For example, the staffing plan for interval $(t, t+1]$ is made at time $t-1$, based on the actual number of passengers who have not been served until $t-1$, and the predicted passenger arrivals in $(t-1, t]$ and $(t, t+1]$. Results are similar when we assume managers need more time to reallocate resource, and decisions need to be made at time $t-2$, $t-3$, etc.

Staffing the immigration desk carries a cost, and understaffing also carries a cost in terms of dissatisfied passengers and breach of targeted maximum queuing time goals. Therefore, the resourcing problem can be formulated as a newsvendor problem. A similar formulation was applied in Bassamboo et al. (2010) in the context of call centers. Suppose we are at time $t-1$, the optimization problem

for the staffing plan in time interval $(t, t+1]$ is as follows: find the number of immigration desks to be opened, $b^*_{t+1} \geq 0$, that minimizes:

$$\Pi(b_{t+1}) = c_{IU} E(y_t + N_{t+1} - \mu b_{t+1})^+ + c_{IS} b_{t+1} \tag{1}$$

where c_{IU} is the immigration understaffing cost per passenger per 15 minutes, y_t is the number of passengers in line in the beginning of $(t, t+1]$, N_{t+1} is the number of passengers connecting to domestic flights that arrive in $(t, t+1]$, μ is the service rate at each immigration desk, and c_{IS} is the cost of opening a desk per 15 minutes. We assume term y_t equals $\left(x_{t-1} + E(N_t) - \mu b^*_t\right)^+$, where x_{t-1} is the realized number of passengers in line in the beginning of $(t-1, t]$ under optimal decisions $b^*_1, b^*_2, ...b^*_{t-1}$, and $E(N_t)$ is the expected number of passengers who connect to domestic flights and arrive at the immigration area during $(t-1, t]$. The mean value of our prediction on the passenger flow described in Section 3.5 is used as $E(N_t)$.

The optimization problem in Eq. (1) is an instance of the familiar newsvendor problem: optimize the production quantity μb_{t+1} with a unit production cost of c_{IS}/μ and a unit understocking cost c_{IU}. Thus, we obtain the newsvendor-based staffing level at the immigration area as the standard critical quantile solution

$$b^*_{t+1} = \frac{1}{\mu}\left(y_t + F^{-1}_{N_{t+1}}\left(1 - \frac{c_{IS}}{\mu c_{IU}}\right)\right) \tag{2}$$

Where $F_{N_{t+1}}$ is the CDF of N_{t+1}, and therefore $F^{-1}_{N_{t+1}}(1 - c_{IS}/(\mu c_{IU}))$ is the $(1 - c_{IS}/(\mu c_{IU}))$ quantile forecast of N_{t+1} provided by the predictive system. The term $(1 - c_{IS}/(\mu c_{IU}))$ is often referred to as the critical fractile in a newsvendor setting. Note that passengers that are left over from interval t must be processed during interval $t+1$. Since b^*_{t+1} is directly linked to the remaining passengers from the previous period (y_t), this solution guarantees in theory that passengers do not need to wait for more than two periods, i.e. 30 minutes. The decision variable b_{t+1} in the above problem is continuous; however, the number of lanes to be opened should be an integer. Therefore, we round b^*_{t+1} up and down, and retain the option that achieves lower expected cost.

Similarly, the optimization problem in time interval $(t, t+1]$ at the transfer security area is as follows: find the number of scurity lanes to be opened, $q^*_{t+1} \geq 0$, that minimizes:

$$\Pi(q_{t+1}) = c_{SU} E(z_t + \mu b^*_{t+1} + G_{t+1} - \eta q_{t+1})^+ + c_{SS} q_{t+1} \tag{3}$$

where c_{SU} is the understaffing cost per passenger per 15 minutes in the security area, z_t is the number of passengers left over from previous time periods, μb^*_{t+1} denotes the number of passengers who join the queue after passing through immigration during $(t, t+1]$, η is the service rate of each security lane, and c_{SS} is the cost of a lane per 15 minutes. Variable G_{t+1} is the number of passengers connecting to international flights that arrive during $(t, t+1]$. These passengers do not need to go

through immigration desks. We assume $z_t = (m_{t-1} + \mu b_t^* + E(G_t) - \eta q_t^*)^+$, where m_{t-1} is the realized number of passengers left over at time $t-1$ under optimal decisions $q_1^*, q_2^*, ...q_{t-1}^*$ and $b_1^*, b_2^*, ...b_{t-1}^*$, μb_t^* is the number of passengers joining the queue from immigration in $(t-1, t]$, and $E(G_t)$ is the expected number of passengers connecting to international flights and arriving at the security area during $(t-1, t]$. The term $E(G_t)$ is the mean value of our prediction on the flow of passengers connecting to international destinations. The solution for the problem in Eq. (3) is

$$q_{t+1}^* = \frac{1}{\eta}\big(z_t + \mu b_{t+1}^* + F_{G_{t+1}}^{-1}(1 - \frac{c_{SS}}{\eta c_{SU}})\big) \tag{4}$$

Where $F_{G_{t+1}}$ is the CDF of G_{t+1}, and therefore $F_{G_{t+1}}^{-1}(1 - c_{SS}/(\eta c_{SU}))$ is the $(1 - c_{SS}/(\eta c_{SU}))$ quantile of the distribution for G_{t+1}. The decision variable q_{t+1}^* is rounded here as well.

For our empirical evaluation, service rates are assumed to be deterministic. We set μ and η to 12 and 35 passengers per 15 minutes, respectively, based on the information provided by Heathrow. Since we do not want to make any judgement in advance about the staffing and understaffing cost, we allow flexibility and examine the performance under different scenarios. We set the critical fractile levels in (2) and (4) to 0.05, 0.25, 0.5, 0.75, and 0.95 to cover a wide range of scenarios corresponding to different ratios of the cost of staffing to the cost of understaffing. When the 0.05 fractile (or quantile) is used, the unit cost of staffing, c_{IS}/μ, is comparable to the unit cost of understaffing, c_{IU}; when the 0.95 fractile is used, the unit cost of staffing is much lower than the unit cost of understaffing. For each of the critical fractiles, we then add up the staffing cost and understaffing cost based on the realized passenger flows and the decisions made under the predictive models.

For the immigration staffing problem, across the five settings of $(1 - c_{IS}/(\mu c_{IU}))$ described above, our two-phased approach reduces cost by anywhere between 20% and 54% as shown in Table 5, when compared to Heathrow's static and dynamic legacy system. In comparison to the SARIMA model with covariates, the two-phased approach reduces cost by 11% to 29%. When testing on the staffing problem at the transfer security area, terms z_t and μb_{t+1}^* in the optimal solution q_{t+1}^* (Eq. 4) depend on the setting of $(1 - c_{IS}/(\mu c_{IU}))$. Therefore, we test the staffing plan at the security area for all 25 different combinations of $(1 - c_{IS}/(\mu c_{IU}))$ and $(1 - c_{SS}/(\eta c_{SU}))$. Percentage improvements in costs using our new predictive system, averaged over the five staffing and understaffing cost ratios at the immigration area, are also presented in Table 5. These results show that, compared to Heathrow's static and dynamic legacy systems, costs are reduced by 12% to 26% when applying our two-phased predictive system. In addition, the cost reduction offered by utilizing our prediction approach compared to the SARIMA model ranges from 10% to 16%.

Capacity, or the maximum number of immigration desks and security lanes that can be opened, is not constrained in the formulation above. In fact, the optimal solutions never exceed the actual

Table 5 Percentage improvements in costs[a] at immigration and security

	Improvements at immigration					Improvements at security				
1 - staffing cost/understaffing cost	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
SARIMA with covariates	11	15	20	23	29	10	11	12	14	16
Static legacy system	22	27	34	41	54	21	22	22	24	26
Dynamic legacy system	20	23	26	27	31	17	15	15	15	12

[a.] The percentage improvement is calculated as $100 * (1 - \frac{\Pi^*_{two-phased}}{\Pi^*_{benchmark}})\%$.

The percentage improvements shown in the table are all significantly greater from zero based on a t-test at the 1% level.

capacity at Heathrow. However, for robustness purposes, we also run simulation with a capacity constraint equal to 80% of the maximum number of immigration desks and security lanes to be opened in the study without constraints. Similar conclusions are drawn as in Table 5.

7. Generalized Application of the Two-Phased Predictive System

According to the airport layout and the description of the passenger transfer journey at Amsterdam Schiphol (Schiphol Airport 2019), Brussels (Brussels Airport 2019), Frankfurt (Frankfurt Airport 2019), and Charles de Gaulle (Paris Aéroports 2019), passengers connecting through European airports seem to have very similar transfer journeys as that described in Figure 1. Therefore, the approach developed in this paper could easily be applied to other airports. Implementation of the approach would require collecting data, training, validating, and testing the model to account for the unique characteristics of each airport.

Most European international hubs are in Schengen countries. All passengers arriving from non-Schengen counties and connecting to Schengen destinations need to go through immigration desks. Instead of predicting two separate passenger flows for those connecting to domestic destinations and international destinations as at Heathrow, these airports could generate forecasts of passenger flows for Schengen destinations and non-Schengen destinations using the two-phased approach, whereby the Schengen region appears as a predictor in the model (similar to the origin or destination region in our current model). In addition, at some of the airports, such as Charles de Gaulle, passengers need to go through border force before taking shuttle service to other terminals. As a result, for these passengers, the time between their departures from the immigration area to their arrivals at the security checking area may also need to be predicted by a separate model. Again, this could be captured by the first phase prediction in our model.

Transfer journeys at international hubs outside Europe are also similar to the journey depicted in Figure 1. For example, based on conversations with representatives from LAX (Los Angeles international airport), international arrivals (with the exception of Qantas passengers) go through immigration and customs before connecting to their domestic flights. These facilities only operate from a few terminals, implying that some passengers have to walk to an adjacent terminal after arrival

(similar to the need at Heathrow to travel to Terminal 5). In some US airports, passengers arriving from international flights also need to pick up their bags from baggage claim and check in them again to their next domestic flight before going through the security check. Therefore, when modelling passengers' connection times in these airports, variables related to the baggage process, such as the number of checked-in baggage, should also be considered as predictors, and the entire transfer journey might need to be broken up into more segments to improve prediction accuracy. Moreover, change in airport layout and passengers' transfer journey would also affect the dependence between passengers' arrivals. The optimal value of ρ used to simulate arrivals in the second phase would need to be reestimated using airport specific data.

Our two-phased approach of predicting arrival times and the number of arrivals can also be generalized to other operations management domains, such as hospitals or theme parks. In a hospital setting, a similar framework can be applied to predict the number of patients in the Emergency Department (ED). Hoot et al. (2008) describes a simulation system for an ED that first fits parametric distributions to patient arrival rates, individual patients' treatment times, wait time for hospital bed, etc. When estimating treatment times, the authors fit separate log-normal distributions for patients who have the same acuity level. Finally, they run simulations based on all the distributions, and generate outputs such as the expected number of patients in the ED.

Our two-phased predictive system can similarly generate predictions of the number of patients in the ED. In the first phase, the system can predict the distribution of time spent in the ED for individual patients, based on their acuity level, their age, the number of patients with higher level of acuity in real-time, etc. Next in the second phase, from these individual distributions, the system can compute the probability of a patient that is still in the ED during different time intervals, and calculate the number of patients as the summation of several Bernoulli variables, with each of them denote whether or not a patient is still in the ED. In this context, distributions and quantiles for the number of patients in the ED can help with resource allocations, and help balance underage and overage costs related to the optimal number of beds (Zychlinski et al. 2019).

A theme park could also potentially utilize a similar predictive system to manage flows of visitors. These visitors' walking times from one attraction to another might be affected by many factors, such as previous location in the park and the number of restaurants along the way. Unlike passengers' connecting journey at airport, however, the next attraction to visit is decided by the visitor. Therefore, if we apply the two-phased approach to predict visitor flows at attraction a, the first-phase output may contain two sets of predictions: distribution of the walking time to attraction a, and the probability of attraction a is the next one to be visited. Based on these predictions, the probability of each visitor j arriving at an attraction during time interval i can be calculated. This probability is similar to the p_{ij} described in Section 3.5, which is the probability of a passenger j arriving at the immigration

and security area during time interval i. Next, by aggregating individual arrivals, the theme park can predict visitor flows at an attraction. These predictions of visitor flows can be applied to set up ride capacities to optimize the number of rides (Ahmadi 1997) or retail profits (Rajaram and Ahmadi 2003). The newsvendor-based staffing decision described in Section 6 can also be applied here to set ride capacities.

Although our predictive system can be adapted to solve other operations problems, there are a few challenges when implementing a similar predictive system in real-time. First, the quality of real-time data, such as how quickly the manager can receive the data and the format of the data, has a significant impact on the quality of the predictions. Second, individual level information and personal data are always difficult to collect and store. Third, the usefulness of the predictive system depends on the decision maker's flexibility of reacting to the real-time predictions. To implement a similar system in other operations domains, firms or organizations may need to build a centralized data system to collect and store data, create potential factors that may affect their customers' behavior using existing data, and train their managers to quickly adjust their decisions according to the real-time predictions.

8. Conclusion and Future Work

In this paper, we offer a first study of passengers' transfer journeys using data provided by Heathrow airport. We develop a two-phased predictive system to provide real-time information about transfer passengers' journeys through the airport. This information is vital for the airport in order to best serve the passengers, the airlines, and their employees. Compared to Heathrow's legacy systems and other benchmarks, our two-phased predictive system performs favorably in both identifying late passengers and predicting number of passenger arrivals at the immigration and security areas.

Our study makes advanced machine learning accessible to managers with no data science background working in data science at the Airport Operations Centre (APOC). It also enables the APOC to move from fragmented data and Excel based reporting to a more sophisticated data science environment. The capabilities demonstrated by this study inspired institutional investment in improving data science skills. Given the fact that APOCs are becoming standard in Europe (Eurocontrol 2010), the impact of this work extends beyond Heathrow, with interest expressed by Aéroports de Paris.

The work here demonstrates the usefulness of the regression tree method. Although simple in nature, it provides accurate predictions and easy interpretations. While other models served only as benchmarks here, we encourage continuous exploration of alternative models for improvement. For instance, although some advanced machine learning methods, such as gradient boosting machine, can generate accurate point and quantile forecasts, how to construct the entire distribution from the model and simulate from the distribution is not trivial. Therefore, developing superior models in forecasting entire probability distributions would be an interesting future direction of research.

A key part of our predictive system is the calibration of the distributions. We make assumptions that passenger arrivals are correlated and use copula-based simulation to make the forecasts well-calibrated. Developing other ways to calibrate the forecasts would be a potential topic for future work. In addition, in the paper, we formulate a relatively simple staffing problem based on a newsvendor objective. For future work, we encourage researchers and practitioners to formulate other decision making problems that make use of distributional forecasts.

Appendix

Table A.1 Descriptions of the variables in the data set excluded variables either did not provide useful information, did not improve accuracy of the model, was not available in real time or had too many levels

Data set	Variable name[a]	Description
BOSS	on chocks time[e]	The time when an aircraft is parked at gate.
	aircraft body	A flight's aircraft body type: W (wide) or N (narrow).
	aircraft type[c]	A flight's aircraft type. There are 23 types in total.
	passenger capacity	The capacity of the flight.
	passenger total	Total number of passengers on the flight.
	passenger transfer	Number of transfer passengers on the flight.
	runway no.[b]	Runway number of the flight. There are four runway numbers in the dataset: 27L, 27R, 09L, and 09R.
	scheduled time[c]	Scheduled arrival/departure time of the flight.
	stand no.[b]	Stand number of the flight. There are 213 stand numbers in total.
	inbound date[c]	The date of a flight arrives at the airport.
	flight no.[c]	There are 692 and 399 unique flight numbers for arriving and connecting flights, respectively.
	origin/destination airport[c]	There are 165 unique origin airports for international arrival flight with passengers connecting through T5, and 143 unique destination airports of flights that have transfer passengers and depart from T5.
BDD	**passenger travel class**	Passengers' travel class on the arriving flight.There are five classes in the data set. We grouped them into two categories: economics (EC) or business and first class (NEC).
	inbound terminal	A passenger's arriving terminal.
	inbound stand type	Stand type of the arriving flight: P (Pier served stand) or R (Remote stand).
	outbound stand type	Stand type of the connecting flight.
	passenger outbound seat[d]	A passenger's seat number on the outbound flight.
Conformance	local conform time[e]	Timestamp of when a passenger arrives at Conformance desk.
	conform location code[d]	Code of the conformance desk.
	conform location descrp[d]	Terminal number, conformance desk number, and international or domestic connecting flight.
Created variables	**inbound region**	The region of the departure airport for the arriving flight. There are four regions: UK, Europe, North America, and the rest of world.
	outbound region	The region of the destination airport for the connecting flight.
	inbound punct	Punctuality of the arriving flight.
	inbound hour	Hour of the day when the arriving flight lands at the airport.
	perceived delta	Time difference between the inbound flight's on-chock time and the outbound flight's scheduled departure time.
	Inbound load[d]	Load factor of the arriving flight. Defined as the ratio of the actual number of passengers to the capacity of the flight for inbound flight.
	outbound load[d]	The ratio of the actual number of passengers to the capacity of the flight for the outbound flight.
	day of the week[d]	Day of the week when the passenger arrives at the airport.

[a] Variables in bold represent the 17 predictors used to train the model. Note eight predictors are from the BOSS data as each row in the table under BOSS are for both arriving and connecting flights.

[b, c, d, e] These variables were excluded because they were not available in real-time (b), had too many levels (c), did not help improve model accuracy (d), or were used only to calculate connection times (e).

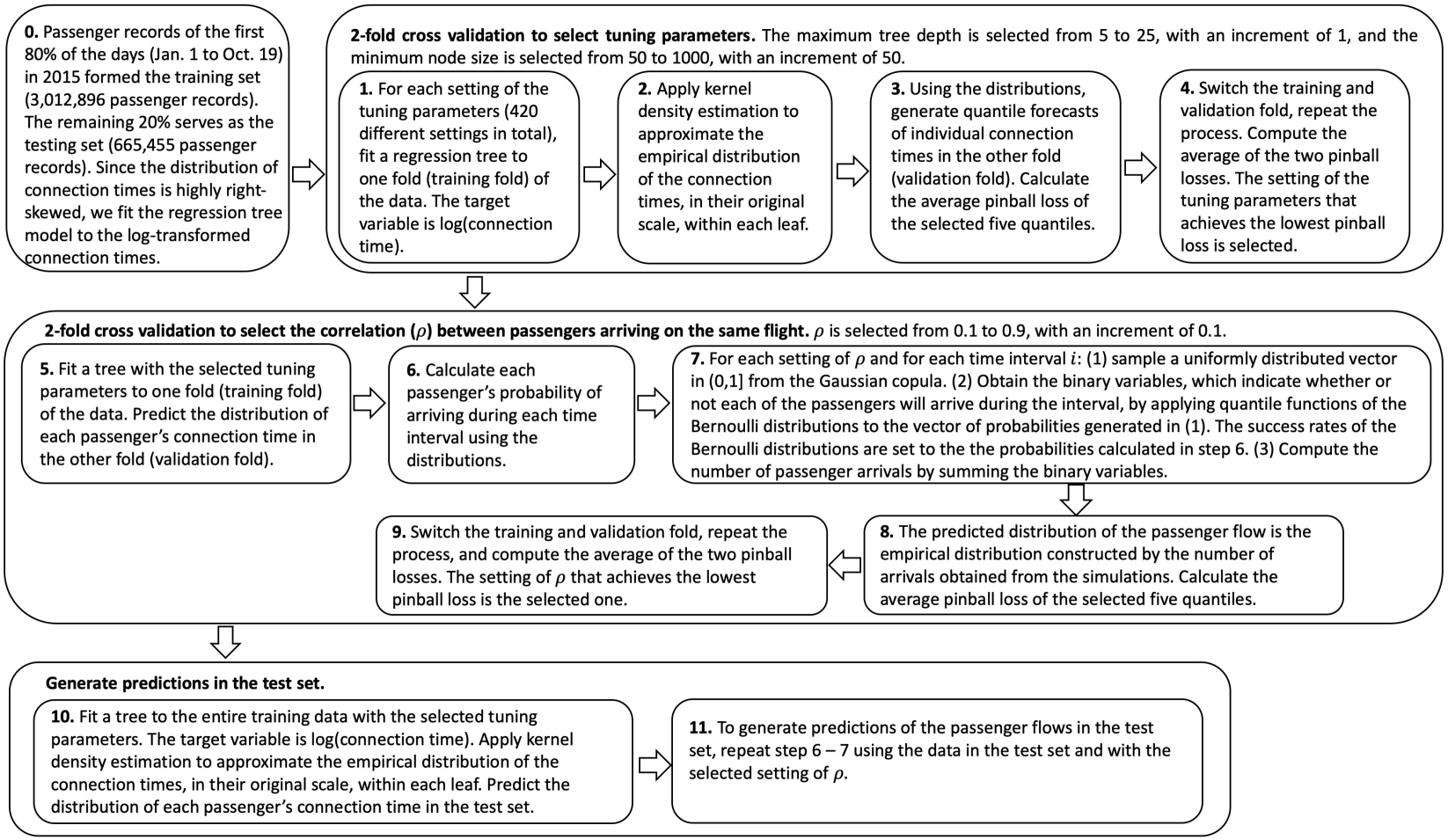
Table A.2 Summary statistics of the numerical predictors

	mean	median	standard deviation
inbound flight passenger capacity	221	189	84
outbound flight passenger capacity	228	205	84
inbound passenger total	176	151	86
outbound passenger total	187	161	90
inbound passenger transfer	62	48	50
outbound passenger transfer	91	76	61
inbound punctuality	42	49	55
perceived delta	132	113	108

Table A.3 Summary statistics of the categorical predictors

	Summary
inbound aircraft body	55% of the flights' body type was "Narrow", the others were "Wide"
outbound aircraft body	51% of the flights' body type was "Narrow", the others were "Wide"
inbound hour	The busiest hours in 2015 were 6:00 - 7:00 and 12:00 - 13:00, both with an average of 28 international flights landing at the airport.
passenger travel class	73% of the passengers traveled in economy class
inbound terminal	65% of the passengers arrived at T5, the others arrived at other terminals
inbound stand type	91% were pier serviced, the others were remote stand
outbound stand type	90% were pier serviced, the others were remote stand
inbound region	39%, 36%, 6%, and 19% of the passengers were from EU countries, North American, Asia, and rest of the world
outbound region	49%, 30%, 5%, and 16% of the passengers traveled to EU countries, North American, Asia, and rest of the world

Figure A.1 Steps taken in Section 3.4 and 3.5 to train the two-phased model and generate predictions from the model



Accuracy Results Based On the Continuous Ranked Probability Score (CRPS) and Logarithmic Score

Although we focus on quantiles when we evaluate the accuracy of distributional forecasts, here we also report the CRPS and logarithmic scores which evaluate the accuracy of probabilities predicted by the distributions. For those methods that only generate independent quantiles, we follow Lichtendahl et al. (2013) and approximate the entire distribution by fitting piecewise-linear cdfs to the five reported quantiles and the minimum and maximum values predicted by the models. The CRPS of a realization, x, and a predicted cdf, F, is defined as $CRPS(F,x) = \int_{-\infty}^{\infty} \big(F(y) - \mathbb{1}(y-x)\big)^2 dy$, where $\mathbb{1}$ is the indicator function that equals 1 if $y - x \geq 0$ and 0 otherwise. The logarithmic score is defined as $LogS(F,x) = -\log(f(x))$, where f is the pdf of the predicted distribution. More details regarding these two accuracy measures can be found in Gneiting and Raftery (2007). As shown in Table A.4 and A.5, the regression tree model performs the best in predicting the distribution of connection times, and the two-phased approach using regression tree in the first phase outperforms the others in predicting passenger flows.

Table A.4 CRPS and log scores of the forecasts on connection times

	CRPS	Log score[a]
Naïve model	7.30	3.80
Linear regression	6.29	3.62
LASSO quantile regression	6.21	3.68
Quantile regression forest	6.26	3.63
Gradient boosting machine	6.19	3.71
Regression tree	**6.10** ***	**3.58** ***

Values in bold indicate the lowest errors. The symbol *** indicates the difference between the regression tree model and the second best model in each column is significant at the 1% level.

[a] Excludes 1.7% passenger records that are outside of one of these models' predicted distributions' support.

Table A.5 CRPS and log scores of the forecasts on passenger flows connecting to domestic and international destinations at the conformance desk

	Domestic		International	
	CRPS	Log score[a]	CRPS	Log score[a]
Naïve model	7.76	4.00	27.83	5.24
SARIMA with covariates	4.86	4.01	15.06	5.09
Static legacy system	7.72	5.72	29.65	5.68
Dynamic legacy system	5.82	4.02	23.70	5.27
Linear regression with copula	3.65	3.23	11.62	4.39
Regression tree without copula	3.79	4.10	12.77	6.38
Regression tree with copula	**3.48****	**3.23**	**11.36***	**4.37**

Values in bold indicate the lowest errors. The symbol ** and * indicate the difference between the regression tree model with copula and the linear regression model with copula in each column is significant based on a t-test at the 5% and 10% level.

[a] Excludes 4.8% and 12.0% time intervals that are outside of one of these models' predicted distributions' support for the domestic and international flows, respectively.

Table A.6 Descriptions of the 16 passenger segments

node number[a]	Average conecting time	Std. of the connecting times	Arriving terminal	Departing region of the arriving flight	Travel class of the arriving flight	Perceived connection time	Stand type of the arriving flight	Hour of the day the arriving flight lands at the airport	Total numbber of passengers on the arriving flight
14	13.1	8.8	Terminal 5	EU	Business/First		Pier served		
11	17.9	8.8	Terminal 5	EU	Economy	< 103			
15	18.8	9.5	Terminal 5	EU	Business/First		Remote		
12	23.0	15.8	Terminal 5	EU	Economy	>= 103			
7	23.1	9.1	Terminal 5	Non-EU	Business/First	< 100			
8	25.6	13.1	Terminal 5	Non-EU	Business/First	>= 100			
4	28.5	10.8	Terminal 5	Non-EU	Economy	< 131			
30	32.5	9.3	Terminal 2/3/4	EU	Business/First			after 13:00	
29	36.4	11.6	Terminal 2/3/4	EU	Business/First			before 13:00	
20	36.8	10.0	Terminal 2/3/4		Economy	< 148		after 15:00	
5	37.2	18.0	Terminal 5	Non-EU	Economy	>= 131			
26	37.7	10.9	Terminal 2/3/4	Non-EU	Business/First	< 133			
27	41.8	15.7	Terminal 2/3/4	Non-EU	Business/First	>= 133			
19	43.6	12.5	Terminal 2/3/4		Economy	< 148		before 15:00	
22	46.7	17.6	Terminal 2/3/4		Economy	>= 148			< 183
23	52.8	18.6	Terminal 2/3/4		Economy	>= 148			>= 183

[a] The numbered nodes are the leaf nodes shown in Figure 5.

References

Ahmadi R. 1997. Managing capacity and flow at theme parks. *Operations research* **45**(1) 1–13.

Atkinson SE, Ramdas K, Williams JW. 2016. Robust scheduling practices in the us airline industry: Costs, returns, and inefficiencies. *Management Science* **62**(11) 3372–3391.

Barnhart C, Cohn A. 2004. Airline schedule planning: Accomplishments and opportunities. *Manufacturing & Service Operations Management* **6**(1) 3–22.

Barnhart C, Fearing D, Vaze V. 2014. Modeling passenger travel and delays in the national air transportation system. *Operations Research* **62**(3) 580–601.

Bassamboo A, Randhawa RS, Zeevi A. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science* **56**(10) 1668–1686.

Bassett Jr G, Koenker R. 1982. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association* **77**(378) 407–415.

Bertsimas D, Patterson SS. 2000. The traffic flow management rerouting problem in air traffic control: A dynamic network flow approach. *Transportation Science* **34**(3) 239–255.

Bloomfield P. 2004. *Fourier Analysis of Time Series: an Introduction.* John Wiley & Sons.

Borndörfer R, Grötschel M, Pfetsch ME. 2007. A column-generation approach to line planning in public transport. *Transportation Science* **41**(1) 123–132.

Breiman L, Stone CJ, Friedman J, Olshen RA. 1984. *Classification and Regression Trees.* CRC press.

Brussels Airport. 2019. Connecting flights. `https://www.brusselsairport.be/en/passengers/your-travel-planner/connections`.

Chernozhukov V, Fernández-Val I, Galichon A. 2010. Quantile and probability curves without crossing. *Econometrica* **78**(3) 1093–1125.

De Reyck B, Guo X, Grushka-Cockayne Y, Lichtendahl Jr. KC, Karasev A, Garside T, Coss N, Tasker F. 2016. Apoc business process reengineering big data study. Tech. rep.

De'ath G, Fabricius KE. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**(11) 3178–3192.

Domingos P. 2012. A few useful things to know about machine learning. *Communications of the ACM* **55**(10) 78–87.

Eliashberg J, Hui SK, Zhang ZJ. 2007. From story line to box office: A new approach for green-lighting movie scripts. *Management Science* **53**(6) 881–893.

Eurocontrol. 2010. The potential role of the Airport Operations Centre (APOC) in the SESAR airport concept. Tech. rep.

Ferreira KJ, Lee BHA, Simchi-Levi D. 2015. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* **18**(1) 69–88.

Frankfurt Airport. 2019. Transferring at FRA. `https://www.frankfurt-airport.com/en/travel/transfer.detail.suffix.html/article/travel/services-a-z/easy-travel/transfer-at-fra.html`.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1) 1.

Gneiting T. 2011. Quantiles as optimal point forecasts. *International Journal of Forecasting* **27**(2) 197–207.

Gneiting T, Raftery E. 2007. Strictly proper scoring rules, prediction, and estimation. *American Statistical Association* **102**(477) 359–378.

Goeman JJ. 2010. L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal* **52**(1) 70–84.

Goeman JJ, Meijer R, Chaturvedi N. 2018. L1 and l2 penalized regression models. *Vignette R Package* ***penalized****. URL https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf* .

Grushka-Cockayne Y, Jose VRR, Lichtendahl Jr. KC, Winkler RL. 2017. Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research* **65**(3) 712–728.

Guillermet F. 2017. Towards total airport management. `http://www.airport-business.com/2017/06/towards-total-airport-management/`.

Hastie T, Tibshirani R, Friedman JH. 2009. *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer.

Heathrow Airport. 2020. Heathrow airport facts and figures. `https://www.heathrow.com/company/about-heathrow/company-information/facts-and-figures`.

Hong T, Fan S, Pinson P, Zareipour H, Troccoli A, Hyndman RJ. 2016. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting* **32**(3) 357–363.

Hoot NR, Jones I Levin SR Zhou C Gadd CS, LeBlanc LJ, Aronsky D. 2008. Forecasting emergency department crowding: a discrete event simulation. *Annals of emergency medicine* **52**(2) 116–125.

Hyndman RJ, Athanasopoulos G. 2018. *Forecasting: Principles and Practice*. OTexts.

Jacquillat A, Odoni AR. 2015. An integrated scheduling and operations approach to airport congestion mitigation. *Operations Research* **63**(6) 1390–1410.

James G, Hastie T, Witten D, Tibshirani R. 2013. *An Introduction to Statistical Learning*. New York, NY: Springer.

Jose VRR. 2016. Percentage and relative error measures in forecast evaluation. *Operations Research* **65**(1) 200–211.

Jose VRR, Winkler RL. 2009. Evaluating quantile assessments. *Operations Research* **57**(5) 1287–1297.

Lan S, Clarke JP, Barnhart C. 2006. Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions. *Transportation Science* **40**(1) 15–28.

Lichtendahl Jr. KC, Grushka-Cockayne Y, Winkler RL. 2013. Is it better to average probabilities or quantiles? *Management Science* **59**(7) 1594–1611.

Lohatepanont M, Barnhart C. 2004. Airline schedule planning: Integrated models and algorithms for schedule design and fleet assignment. *Transportation Science* **38**(1) 19–32.

MacDonald M. 2016. Annual analyses of the EU air transport market. Tech. rep.

Meinshausen N. 2006. Quantile regression forests. *Journal of Machine Learning Research* **7**(Jun) 983–999.

Milenković M, Melichar V Bojović N, Švadlenka L, Avramović Z. 2018. Sarima modelling approach for railway passenger flow forecasting. *Transport* **33**(5) 1113–1120.

Mukherjee A, Hansen M. 2009. A dynamic rerouting model for air traffic flow management. *Transportation Research Part B: Methodological* **43**(1) 159–171.

Paris Aéroports. 2019. Connecting flights. `https://www.parisaeroport.fr/en/passengers/flights/connecting-flights`.

Rajaram K, Ahmadi R. 2003. Flow management to optimize retail profits at theme parks. *Operations Research* **51**(2) 175–184.

Schiphol Airport. 2019. Schiphol airport map. `https://www.schiphol.nl/en/airport-maps`.

Shang Y, Dunson D, Song JS. 2017. Exploiting big data in logistics risk assessment via bayesian nonparametrics. *Operations Research* **65**(6) 1574–1588.

Solak S, Clarke JB, Johnson EL. 2009. Airport terminal capacity planning. *Transportation Research Part B: Methodological* **43**(6) 659–676.

Taylor JW. 2008. A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science* **54**(2) 253–265.

Taylor JW. 2012. Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science* **58**(3) 534–549.

Van Brussel B. 2018. *Simulating the patient flow on the short stay unit.* Master thesis, University of Amsterdam.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S.* Springer-Verlag.

Wand MP, Jones MC. 1994. *Kernel Smoothing.* Chapman and Hall/CRC.

Wei W, Hansen M. 2006. An aggregate demand model for air passenger traffic in the hub-and-spoke network. *Transportation Research Part A: Policy and Practice* **40**(10) 841–851.

Wei Y, Chen MC. 2012. Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies* **21**(1) 148–162.

Wu PP, Mengersen K. 2013. A review of models and model usage scenarios for an airport complex system. *Transportation Research Part A: Policy and Practice* **47** 124–140.

Xue Z, Wang Z, Ettl M. 2015. Pricing personalized bundles: A new approach and an empirical study. *Manufacturing & Service Operations Management* **18**(1) 51–68.

Zychlinski N, Mandelbaum A, Momčilović P, Cohen I. 2019. Bed blocking in hospitals owing to scarce capacity in geriatric institutions—cost minimization via fluid models. *Manufacturing & Service Operations Management* .