Appendices for "Forecasting Airport Transfer Passenger Flow Using Real Time Data and Machine Learning"

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).

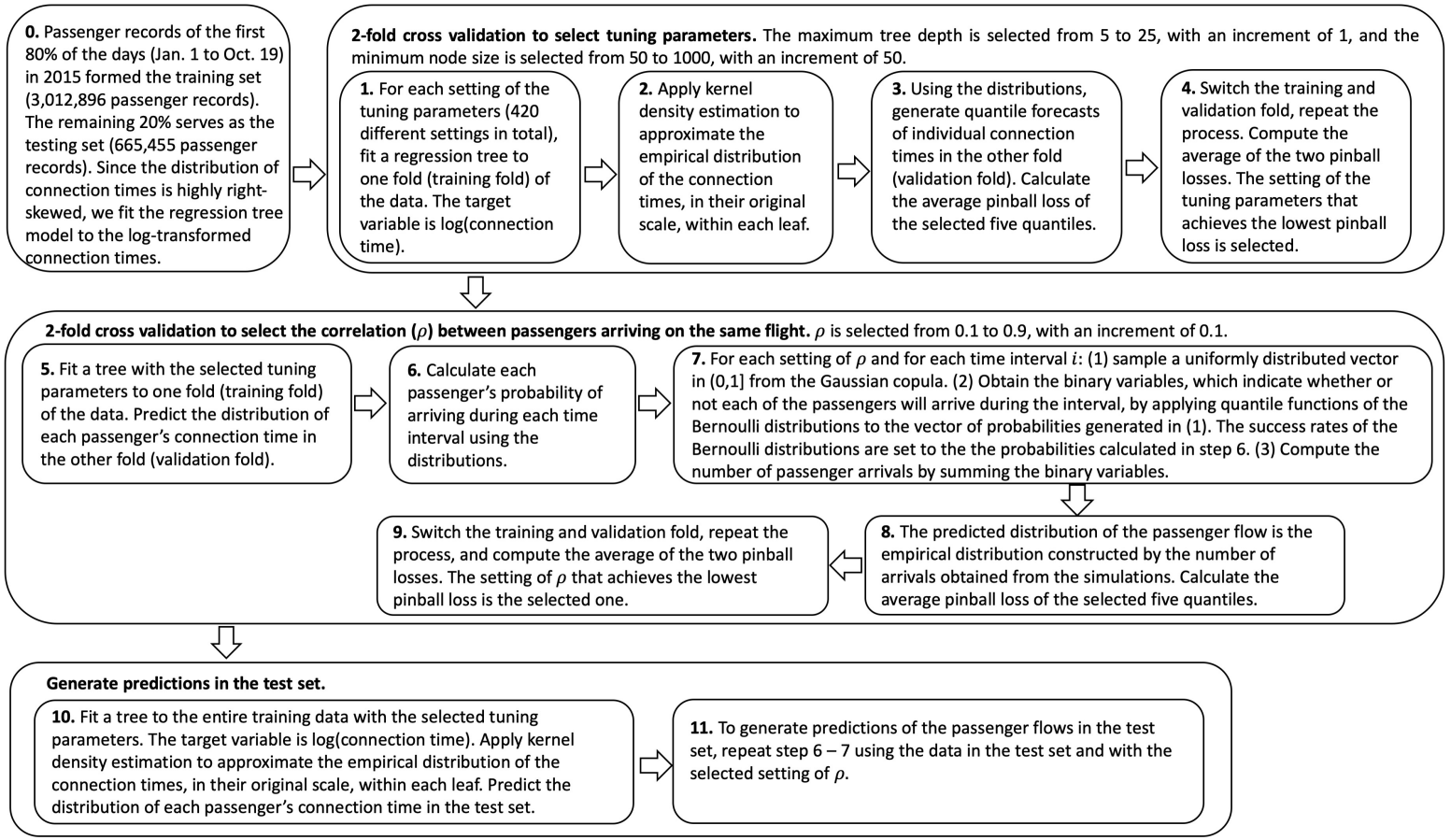
Table A.2 Summary statistics of the numerical predictors

	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 served, the others were remote stand
outbound stand type	90% were pier served, 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.