SCHEDULED FLIGHT TIMES

MATTHEW THIBAULT

Suppose you buy a non-stop flight from JFK to LAX. Over 6000 other flights from JFK to LAX occur each year on average, making it the 36th most common flight since October 1987. Your ticket says that the flight time is six and a half hours. Does every other flight this year have the same number? Surprisingly, the answer is no; the scheduled flight time follows a complicated pattern. It depends on the airline, date, and time of travel. Figure 1 shows scheduled flight times for American Airlines flights between JFK and LAX, since October 1987; this illustrates the complexity of the problem that we study. Notice the change of density of the scheduled flight times during late 2006 and the start of 2015; this shift seems to be unique to American Airlines flights! Despite this irregularity, there seems to be a regular yearly pattern. For flights from JFK to LAX, scheduled flight times are higher in the winter, and lower in the summer; however, the reverse is true for flights from LAX to JFK! I explain this seasonality later in this document.

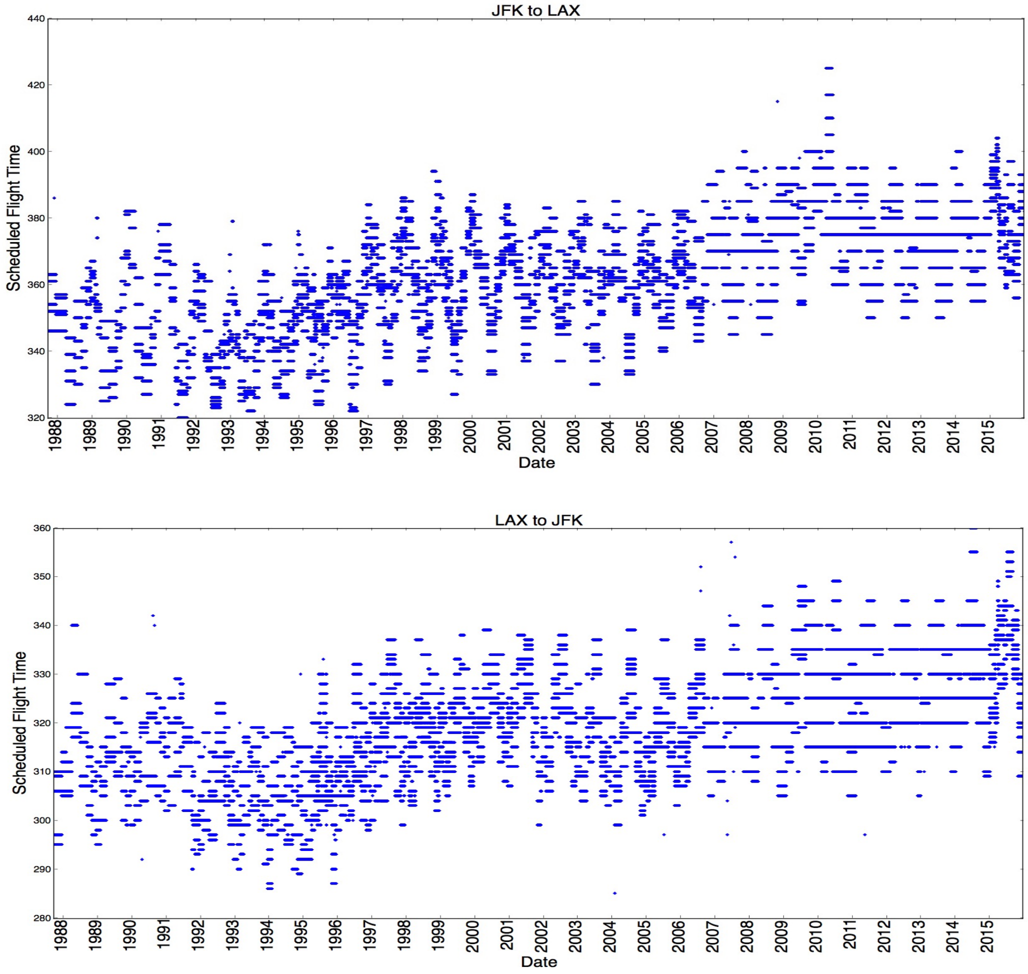

Figure 1.

In this study, I will look at the relationship of scheduled flight times with various factors including departure and arrival times, day of the week, month of the year, distance of the flight, and origin and destination latitude and longitude, for all flights within the US. For this study, I downloaded flight data from the Bureau of Transportation Statistics, available here. I cleaned this data using Python pandas; please see section 3 and the scripts for more details. Subsequently, I uploaded the data to Google BigQuery and obtained cross-sectional data and trends. The queries and their result are in BigQuery.txt and section 1. Finally, I used pandas and scikit-learn to model scheduled flight times during a given month, as a function of distance, and origin and destination coordinates. These results are presented in section 2. Again, please refer to the scripts for further details.

Brief Summary of Results

1. Major low-cost airline carriers with a small number of flights have shorter scheduled flight times, up to 3.5 minutes faster than average. Airline carriers with regional airline service have longer scheduled flight times, up to 1.8 minutes slower than average.

2. Scheduled flight times increase during busy departure and arrival times, with average fluctuation of 6 minutes. The effect of the day of the week is minimal.

3. Westbound flights have longer scheduled flight times than eastbound flights, due to the jet stream which flows from west to east. This effect is more pronounced in the winter and less pronounced in the summer, since the jet stream is stronger during the winter.

4. 97.7% of scheduled flight times can be calculated with error less than 20 minutes, using the following model:

$$\text{Sch. Time} = \text{Ground Time} + \frac{\text{Distance}}{\text{Plane speed}} + \frac{\text{Distance} \cdot \cos(\theta)}{\text{Effect of Jet stream}},$$

   where θ is the bearing of the plane.

5. For flights whose destination is not within the Pacific, the model calculates: Plane speed = 490 mph; Ground Time increased from 28 minutes to 42 minutes from Oct. 1987 to Dec. 2015; Effect of Jet stream fluctuates regularly between 50 and 25 mph each year from winter to summer.

6. For flights whose destination is within the Pacific, the model calculates that before 1993: Plane speed = 530 mph, Ground Time = 27.5 minutes, Effect of jet stream fluctuates regularly between 36 and 24 mph each year from winter to summer. After 2011, plane speed dropped to 500 mph, ground time = 24 minutes, and effect of jet stream fluctuates between 40 and 27 mph.

1. Effect of Time, Date, and Carrier on Scheduled Flight Times

After cleaning the flight data, we have over 160 million flight records, with information about the carrier, date, origin, destination, and scheduled departure, arrival, and flight times. To determine the effect of departure and arrival times, day of week, and carrier on scheduled flight times, I performed the following calculation in BigQuery. To get a baseline measurement, take the average scheduled flight time, fixing the flight origin and destination, year, and month. For each flight record, calculate the difference between the scheduled flight time and the average corresponding to that year, month, flight origin, and destination. Finally, group by the dependent variable and take the average. Call this average deviation ΔSchTime. A positive number reflects longer scheduled flight times while a negative number reflects shorter scheduled flight times. Note that since we take the average over the flight records, the resulting average deviation is weighted by the frequency of each flight.

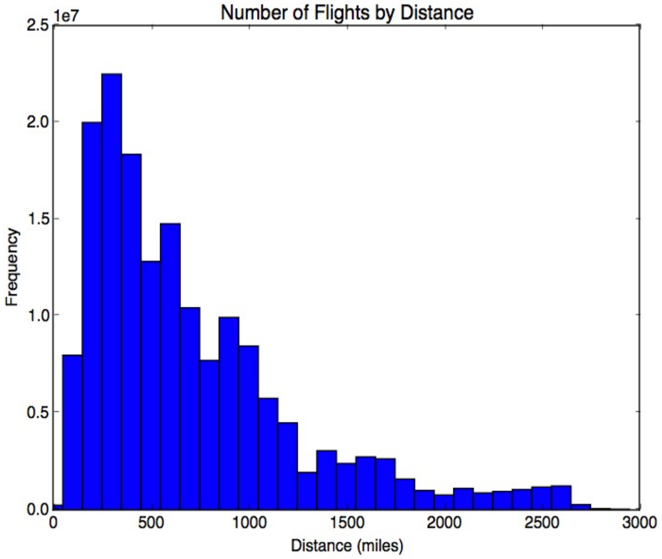

Figure 2.

Figure 2 shows that this is skews the calculation toward flights with distance less than 500 miles. First, we look at the effect of the airline carrier on scheduled flight times.

Carrier	ΔSchTime
Spirit Airlines	-3.547
Frontier Airlines	-1.987
AirTran	-1.801
Virgin America	-1.683
JetBlue Airways	-1.206

Carrier	ΔSchTime
Endeavor Air	0.7930
Independence Air	0.8140
Eastern Air Lines	0.9285
PSA Airlines	1.0304
Pan American	1.7995

Those carriers with shorter scheduled flight times are major carriers with a smaller number of flights each year. Except for Virgin America, these carriers are low-cost airlines.

Both Pan American and Eastern Air Lines went bankrupt during 1991; the increased scheduled flight times may reflect an older fleet or a decline in flights. PSA Airlines, Independence Air, and Endeavor Air primarily offer regional airline service. The smaller planes may explain the longer scheduled flight times.

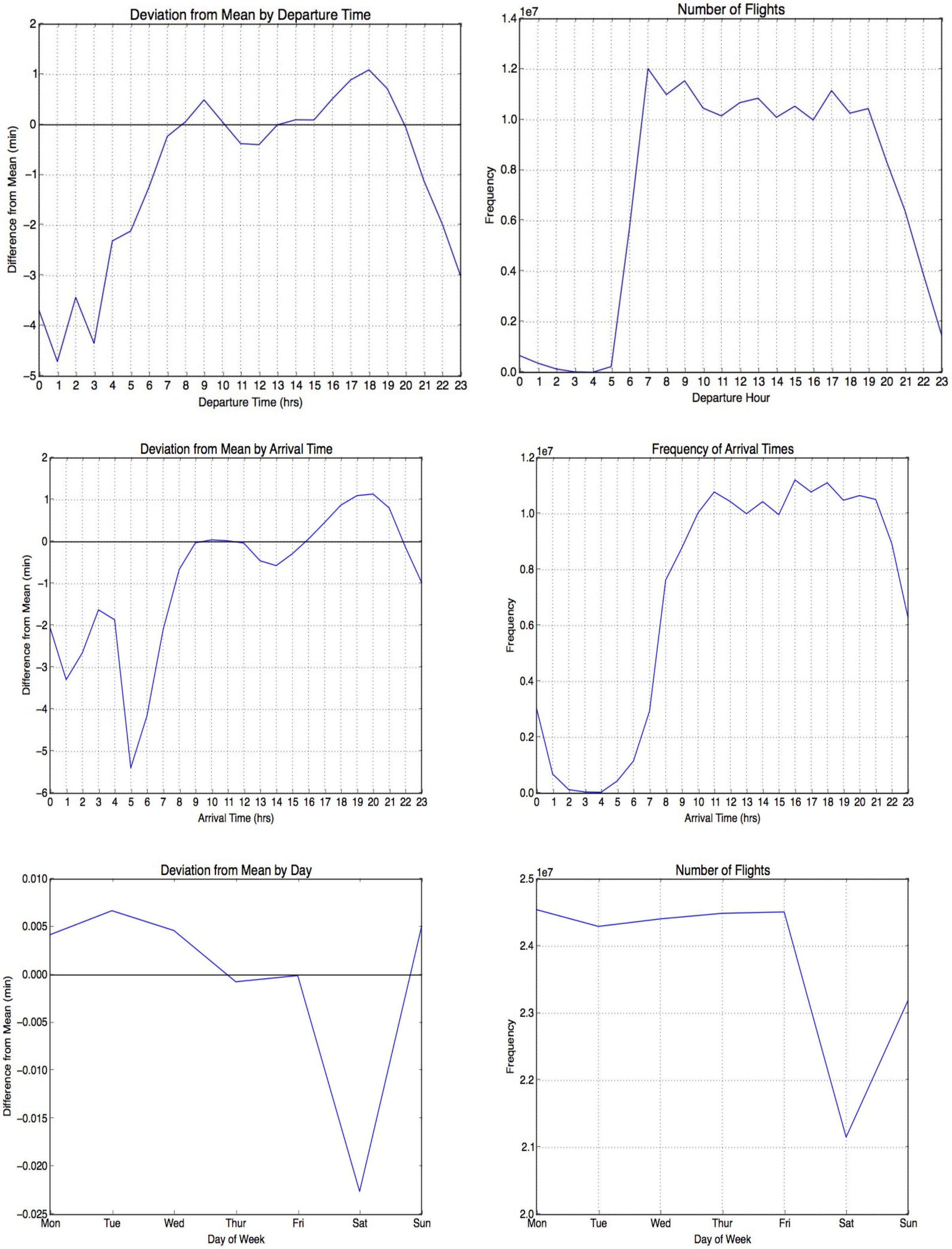
Deviation from Mean by Departure Time
Difference from Mean (min)
2
1
0
−1
−2
−3
−4
−5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Departure Time (hrs)
Number of Flights
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Departure Hour
Deviation from Mean by Arrival Time
Difference from Mean (min)
2
1
0
−1
−2
−3
−4
−5
−6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Arrival Time (hrs)
Frequency of Arrival Times
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Arrival Time (hrs)
Deviation from Mean by Day
Difference from Mean (min)
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.020
−0.025
Mon Tue Wed Thur Fri Sat Sun
Day of Week
Number of Flights
1e7
2.5
2.4
2.3
2.2
2.1
2.0
Frequency
Mon Tue Wed Thur Fri Sat Sun
Day of Week

Figure 3.

The graphs on the left of Figure 3 show the effect of departure time, arrival time, and day of the week on scheduled flight times. The graphs on the right show the frequency of flights by departure time, arrival time, and day of the week. Notice that frequency of flights is correlated with increased scheduled flight times. In other words, scheduled flight times tend to be longer when the airport is busy. This makes sense since at these times, airplanes will take longer to wait for a runway. Popular departure times are 7 am to 8 pm. Popular arrival times are 9 am to 10 pm. Finally, the least number of flights occur on Saturday; this is reflected in the decreased scheduled flight times.

Figure 4 shows the effect of month on scheduled flight times for eastbound and westbound flights. We calculate this in a similar manner as above. However, we consider east and westbound flights separately. When taking the average, we fix the flight origin, destination, and year. Since we only have partial data for 1987, we analyze the data from 1988 to 2015.

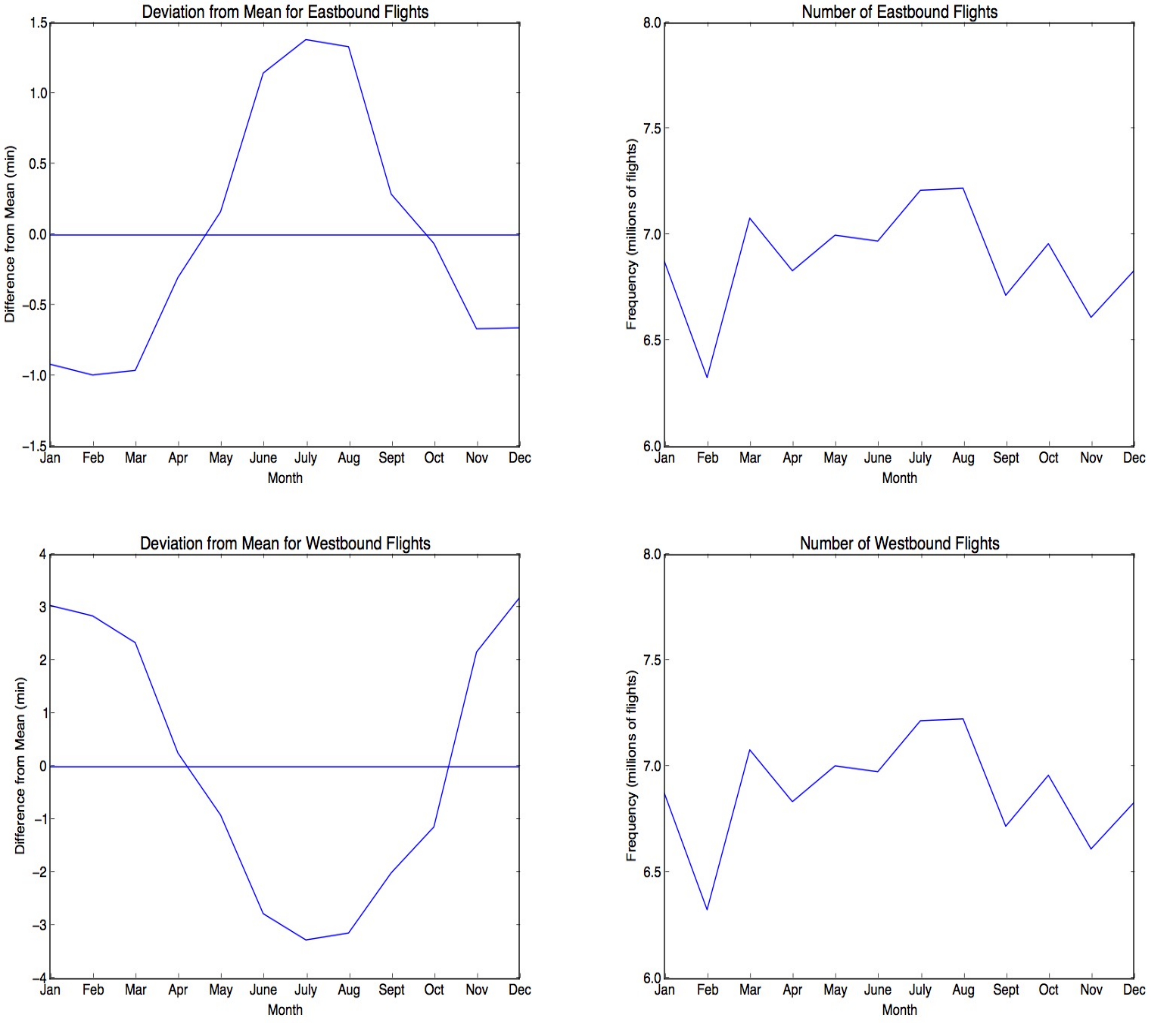

FIGURE 4.

Notice that flights are evenly distributed among the months, with a slight increase in number of flights during the summer months. Thus scheduled flight time does not have a strong correlation with frequency of flights each month. Rather, the plots below verify the pattern observed in the introduction: eastbound scheduled flight times increase in the summer and decrease in the winter, while the reverse is true for westbound flights! In the next section, we will use a model for scheduled flight times to explain this pattern.

The range of fluctuation due to the carrier is around 5.5 minutes; for departure and arrival times it is approximately 6 minutes. For eastbound flights, the range of fluctuation due to the month is 6.5 minutes while for westbound flights it is 2.5 minutes. The effect due to the day of the week is minimal. Looking at the range of scheduled flight times within a day or a month, a BigQuery request similar to those above yields an average range of 10.73 minutes within a day and 13.01 minutes within a month.

2. Modeling Scheduled Flight Times

The following steps are included in the calculation of flight time:

(1) Taxiing from the gate to the runway.
(2) Waiting in line to use the runway.
(3) Ascending to 30000 feet and flying to the destination
(4) Waiting to be cleared, and descending onto the runway.
(5) Taxiing to the gate.

Thus one can model scheduled flight time as $f(\text{origin airport}) + g(\text{distance}) + h(\text{destination airport})$. Let's look at the effect that distance has on scheduled flight time, during January 2015.

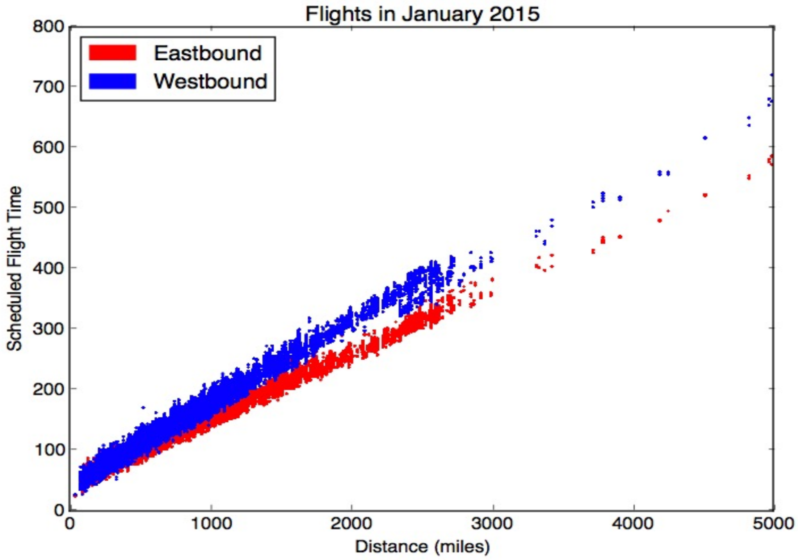

Figure 5.

There are two lines of different slope in this diagram! Using the longitude of each airport, we mark the flights that travel westbound and eastbound. Comparing flights of similar distances, note that westbound flights are slower than eastbound flights. This occurs because of the jet stream, which flows from west to east.

Regarding the flow as a force acting on the plane, we see that the force reduces the velocity of the plane by an amount directly proportional to the cosine of the bearing of the plane, θ. We use a planar triangle to estimate $\cos(\theta)$ as follows:

$$\cos(\theta) = \frac{-\Delta\text{Long}}{\sqrt{(\Delta\text{Long})^2 + (\Delta\text{Lat})^2}}, \text{ where}$$

$\Delta\text{Long} = \text{Long}_{\text{dest}} - \text{Long}_{\text{origin}}$ and $\Delta\text{Lat} = \text{Lat}_{\text{dest}} - \text{Lat}_{\text{origin}}$.

Putting everything together, we have:

$$\text{Sch. Time} = \alpha + \beta \cdot \text{dist} + \gamma \cdot \text{dist} \cdot \cos\theta + \epsilon, \quad \text{where:}$$

$$\alpha = f(\text{origin airport}) + h(\text{destination airport}) = \text{time spent on the ground.}$$

$$\frac{1}{\beta} = \text{plane speed;} \qquad \frac{1}{\beta + \gamma\cos(\theta)} = \text{plane speed, adjusted for the jet stream.}$$

$$\epsilon = \text{Error}$$

Finally, we use ordinary linear regression to determine the best fit constants α, β, γ. We arrive at:

$$(\alpha, \beta, \gamma) = (41.64482, 0.12277, 0.014338), \qquad r^2 = 0.98486.$$

The standard error for α, β, and γ are 0.0225, 2.25×10^{-5}, and 1.52×10^{-5}. Thus each of these constants α, β, γ are significant.

Our linear regression suggests that we wait 41.6 minutes in steps 1, 2, 4, and 5 above. The speed of the plane is $1/\beta$ miles per minute, which evaluates to 489 miles per hour. Finally, the jet stream reduces the speed of the plane by $1/\beta - 1/(\beta + \gamma) = 51$ miles per hour.

By measuring and plotting the error terms from our linear regression, we see that the error has a long left tail. Outputting the origin and destination of the flights with error value less than -25 minutes, primarily produces westbound flights to Hawaii or another Pacific island. So, we run regressions on these flights separately; the resulting predictions are plotted below. The solid lines plot predict flight times given distance, if we travel directly east or west. The dashed line plots predict flight times for westbound flights to Pacific islands. Errors from each regression are plotted in Figure 7. Performing the regression for each month from October 1987 to December 2015, we find that for 97.7% of flights, this regression calculates scheduled flight times within 20 minutes.

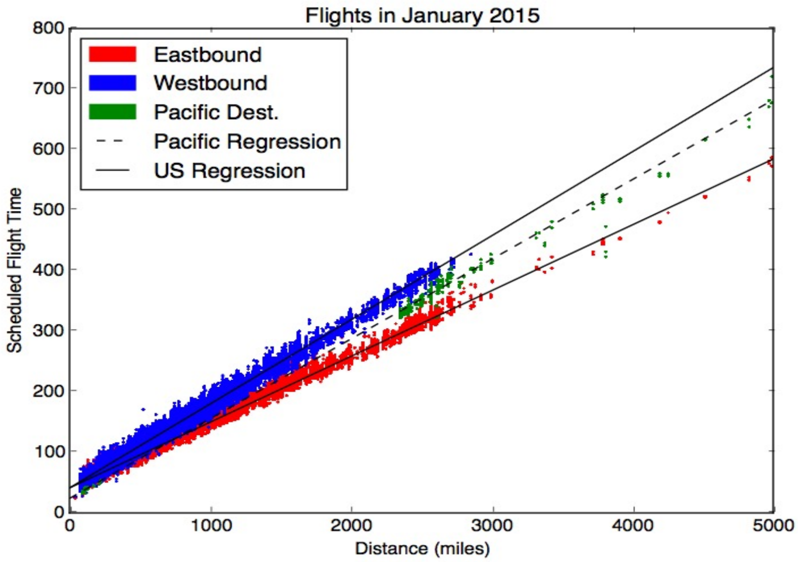

Figure 6.

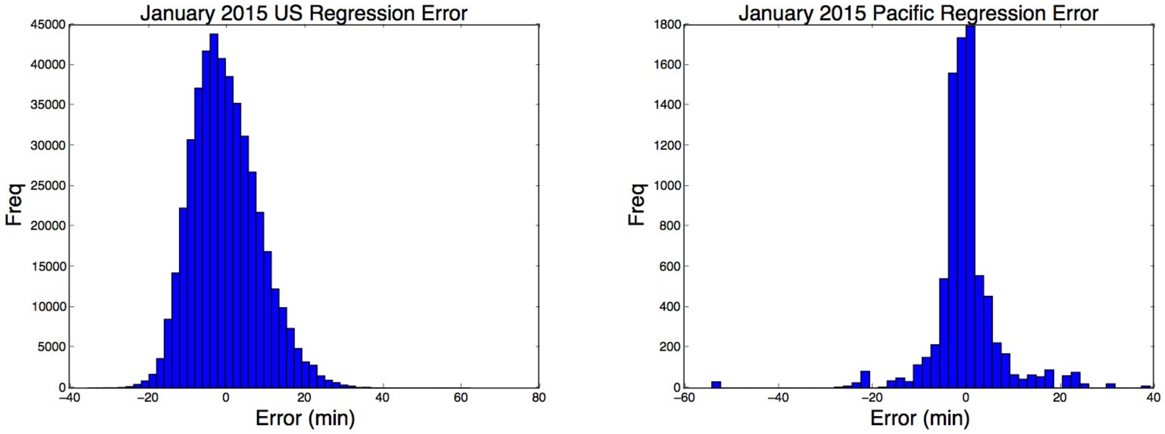

Figure 7.

Below, we plot ground time, plane speed, effect of jet stream, and the correlation coefficient calculated from the regression, for each month from October 1987 to December 2015.

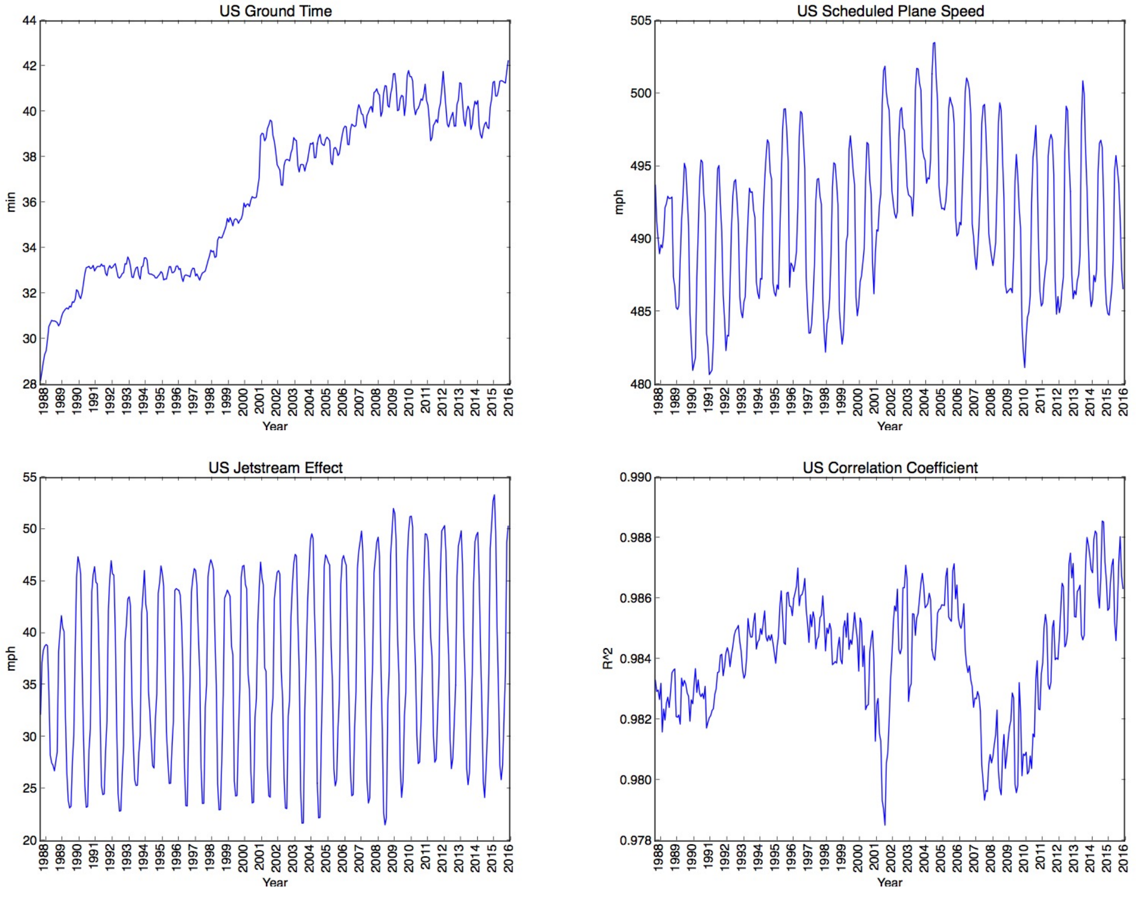

Figure 8.

The scheduled plane speed has held relatively constant at 490 mph since 1987. The scheduled ground time, however, accounts for the increase in scheduled flight times since 1987: it has increased from 28 minutes to 42 minutes over that time period. The correlation coefficient for each month is approximately 0.984, but it had a noticeable dip in mid 2001, and during 2006, and 2007. Notice the effect of the jet stream has consistently fluctuated from 50 mph in the winter to 25 mph in the summer. This agrees with the following statement from the National Weather Service: ”Since these hot and cold air boundaries are most pronounced in winter, jet streams are the strongest for both the northern and southern hemisphere winters.” This explains why westbound flights are faster in the summer while eastbound flights are faster in the winter!

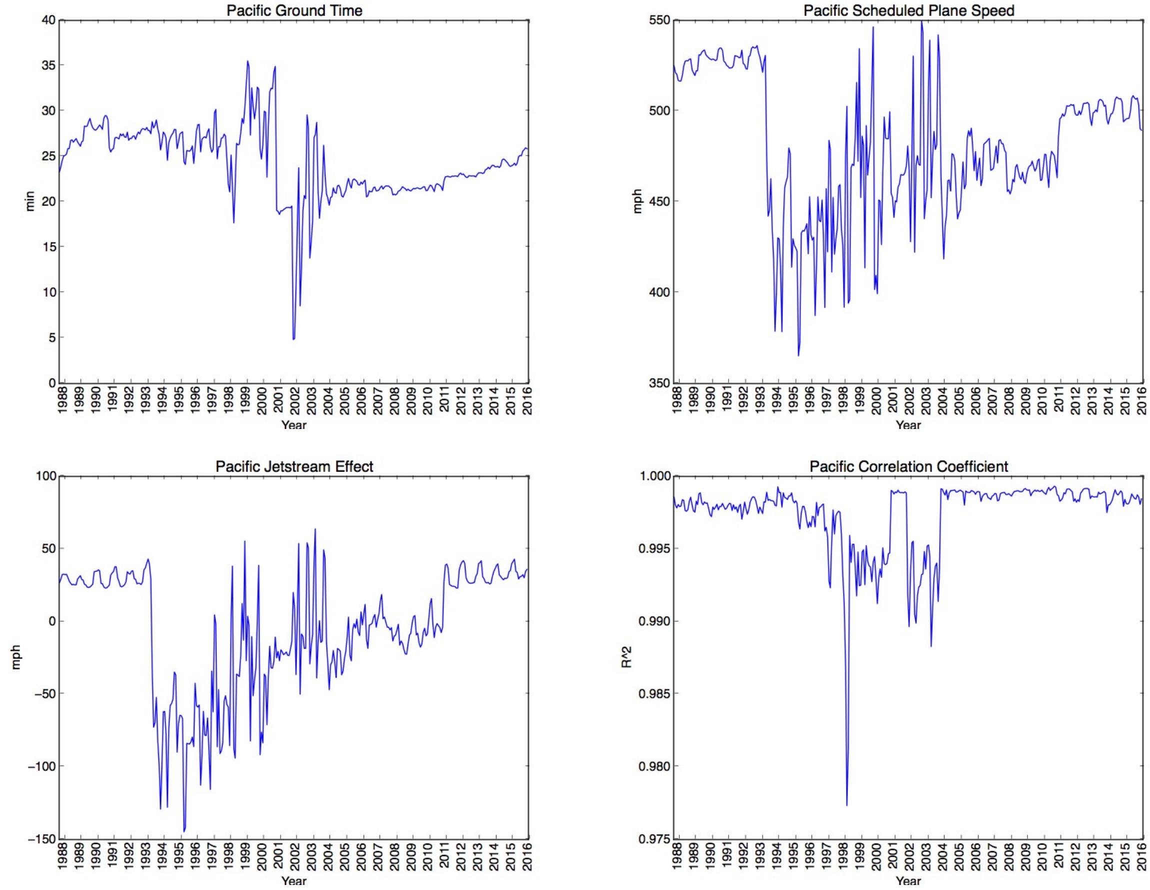

Figure 9.

By looking at the plane speed and effect of jet stream, we see that from 1993 to 2011, our model for scheduled flight times for Pacific island destinations needs revision. We focus on the time period before Jan. 1993 and after Jan. 2011. Before 1993, plane speed is 530 mph, ground time is reduced to 27.5 minutes on average, and the effect of the jet stream fluctuated from 24 to 36 mph. After 2011, plane speed for Pacific destinations is 500 mph, ground time is 24 minutes on average, and the effect of the jet stream fluctuates from 27 to 40 mph. This can be reasoned, since flights with Pacific destinations have larger planes, less busy airports, and the effect of the jet stream is reduced nearer to the equator.

3. Downloading and Cleaning the Data

Flight and airport data from the Bureau of Transportation Statistics is available for download here. After filtering the variables I wanted to study, I separated information about distance between the airports into a separate file. Also, I performed the following checks:

(1) Check if flights have the same origin and destination. There are 7 such flights; I removed these flight records.

(2) Check if each flight with given origin and destination has the same distance. Output an error if not. Keep track of the frequency of each distance, and select the majority distance. There are no such errors.

(3) Check that the distance from origin to destination is equal to the distance from destination to origin. There are no such errors.

Finally, using the following variables: Departure Time, Departure Delay, Scheduled Departure, Arrival Time, Arrival Delay, Scheduled Arrival, Actual Flight Time, and Scheduled Flight Time, we check consistency. All variables should not be null. Also, the following equations must hold:

(1) DepTime = SchDep + DepDelay
(2) ArrTime = SchArr + ArrDelay
(3) ActualTime = SchTime + ArrDelay - DepDelay
(4) ActualTime = ArrTime - DepTime, up to multiples of 60 minutes.
(5) SchTime = SchArr - SchDep, up to the same multiple of 60 minutes.
(6) SchTime, ActualTime > 0

We are interested in obtaining correct values of SchDep, SchArr, and SchTime, which we retain for the analysis in the previous sections.

There are 2 entries in which SchArr is null, and 1 in which SchDep is null. In these cases, all other equations hold, so we set SchArr and SchDep to values such that equations (1) and (2) hold.

There are approximately 3.4 million entries in which ActualTime is null or ≤ 0. When ActualTime and SchTime are invalid, many of the other variables are also invalid. In this case, there are not enough equations to verify SchArr and SchDep and thus calculate SchTime; we abandon such cases.

When ActualTime is valid, so are ArrDelay, ArrTime, DepDelay, and DepTime. Then we can use the equations above to calculate and verify SchTime. We handle the cases as follows:

- If SchTime is invalid, then equations (1), (2), and (4) hold. This suggests that DepTime and ArrTime are correct. We set SchDep and SchArr so that (1) and (2) hold, if they are null. Finally, we set SchTime so that (5) holds.

- If SchTime is valid but (5) fails, we perform the following analysis. If (1) and (5) are incorrect, then as the common variable, SchDep is likely incorrect. Similarly, if (2) and (5) are incorrect, then SchArr is likely incorrect. Finally, if (3) and (5) are incorrect, then SchTime is likely incorrect. We correct so that equations (1) - (3) hold.

- If SchTime is valid but (4) and (5) are incorrect, then equations (1) - (3) hold. In this case, It is likely that ActualTime and SchTime are incorrect. We use (5) to set SchTime to the nearest correct value.

Finally, if SchTime is valid and ActualTime is invalid and equation (5) is incorrect, then equations (1)-(4) hold. This suggests that SchTime is invalid. We use equation (5) to update SchTime to the nearest correct value.

At this point, we have performed all corrective actions. We then drop all flight records where SchDep, SchArr, or SchTime are invalid and all records where equation (5) is incorrect. We retain variables that are necessary for analysis of scheduled flight times. Now note that it is possible that SchTime is off by a multiple of 60 minutes. In lieu of using time zone information and accounting for daylight savings time throughout the years, we search for outliers. Organizing flights each month by flight origin and destination, we calculate the 1st and 3rd quartile, and the median scheduled flight time. We flag records which differ from the 1st or 3rd quartile by more than either the interquartile range or 30 minutes. Then we correct the flagged records by changing SchTime by the multiple of 60 minutes which bring it closest to the median. This completes the cleaning process.