Effect of Oil Price on Flights

A fare estimation and correlation analysis

Stanislav Stoytchev
University of Colorado at Boulder
stanislav.stoytchev@colorado.edu

Faisal Alatawi
University of Colorado at Boulder
faisal.alatawi@colorado.edu

Mahmoud Aljarrash
University of Colorado at Boulder
mahmoud.aljarrash@colorado.edu

ABSTRACT

We examine the relationship of the price of oil to the number of flights taken in the US. We find what is the best season to fly and whether flight prices and oil prices are correlated. We also look at what is the delay between the response of flight prices to changes in oil. We then create a proof-of-concept flight price classifier trained with historical data between 1996 - 2009 and leverage it to create a flight price estimator tool.

1 INTRODUCTION

The world today is more connected than ever. The digital age has allowed anyone to contact nearly anyone else with a few simple clicks. One might think this would reduce the number of people flying but the opposite has been observed. There are more people traveling to more places than ever before. A large part of that is due to reasonable transportation costs. Flights are currently some of the cheapest they have ever been historically (with inflation accounted for)[4], and car manufacturers are building and selling more and more cars every year. The economy has been growing for the past few years, and with it air traffic. In this paper, we wish to focus on air travel specifically and how it's volume (and price) is affected by the price of oil.

The price of oil affects many aspects of our world. Food prices, gas prices, cost of manufacturing are all strongly affected by oil. Since planes use kerosene, which is derived from oil, it makes sense that flights would also be affected by oil. In this paper we categorize this relationship. We look at the correlation of flight volume and oil prices. Does the volume of flights increase when there is a drop in the price of oil? Or does the volume stay the same with airlines absorbing fluctuations?

We also look at delays in trying to correlate the two. Since fuel is bought in advance by airlines, we expect there is some sort of delay between a change in the price of oil and the response seen in airfare prices. Categorizing this delay could be useful for predicting how flight prices fluctuate in the short term.
We would also like to explore the seasonality effect on flight volume. Seasonality is a well-known phenomena with cars. People will drive more during the summer, even though gas is more expensive during the summer. Is the same true for flights?

In answering these questions, we account for population as there are far more people now than there were in the 1990s. An interesting question to answer using this information is whether people are flying more now than they were in the past? And by how much?

Finally, create a classifier to estimate the price of a flight given several input parameters. This classifier will be useful to customers looking to travel, transportation companies, and airlines themselves. Flight estimators are already provided by the major airlines but they take into account real flights and try to optimize profit for the companies by leveraging demand. This is why booking a flight late is very expensive. By using historical data to train our classifier, we answer the question of how much booking at the perfect time saves an average customer as well as the estimate of that flight.

2 RELATED WORK

The *National Center of Excellence for Aviation Operations Research* (2014)[3] released a study analyzing the impact of oil to the U.S. economy in 2014. In this study, they looked at the effects of oil prices on airline pricing. It assessed the relation between the oil prices and the air transportation industry based on 3 factors:

(1) Analyzing the operations data
(2) Interviewing domain experts.
(3) Modeling the impact of oil prices on aviation operators using statistical and general equilibrium models.

The study concluded that there is a **strong** correlation between airline ticket prices and oil prices. A 10 percent increase in fuel price would produce a 5.2 percent reduction in the number of flights[3]. In general, the U.S economy was also negatively impacted by higher oil prices.

3 METHODOLOGY

3.1 Presentation Feedback

Upon presenting our project idea, the TAs gave us gave us positive feedback for the data sets we used. We received a question regarding who will benefit from the results of our study - namely consumers, government, and airlines. The TAs expressed enthusiasm for our flight price classifier.

3.2 Tools

Since Python has excellent statistical packages and data handling, we use it for most of our work. NumPy and SciPy are great for analyzing data and matplotlib is excellent at graphing it. For our classifier, we utilize the sklearn package as it has powerful machine learning capabilities and a wealth of documentation. We then wrap the classifier into a Python airfare estimator program.

In addition we used Pandas [5]. Pandas is a data analyzing Python library. It provides many tools to handle and process all sorts of data. It was used extensively to clean, integrate, and process our data sets.

4 DATA

We merge and use the the population of the US, oil price, and historical data for flights in the US. More detail is given on each dataset below.

Table 1: The attributes of the airfare dataset [6]

Column Name	Example Value	Description
Year	2014	Data Year
quarter	2	Data Quarter
citymarketid_1	30140	City market ID 1
citymarketid_2	30194	City market ID 2
City1	Albuquerque, NM	City 1
City2	Dallas/Fort Worth, TX	City 2
nsmiles	580	Non-Stop Market Miles (radian)
airportid_1	10140	Airport ID 1
airportid_2	11259	Airport ID 2
airport_1	ABQ	Airport code 1
airport_2	DAL	Airport code 2
passengers	362.417	Passenger Per day
fare	189.977	Overall average fare
carrier_lg	WN	Carrier with the largest fare
large_ms	0.99514	Market share of largest carrier
fare_lg	189.729	Average fare of largest carrier
carrier_low	WN	Carrier with the lowest fare
lf_ms	0.995	Market share of lowest carrier
fare_low	189.729	Average fare of lowest carrier

4.1 Oil Price Dataset [2]

This data set has daily oil prices of Brent crude from 1990-2017. Brent crude is being used by the oil industry as the benchmark price for purchases of oil worldwide [8]. The dataset contains two attributes: date and price.

4.2 Bureau of Transportation Statistics' Domestic Airline Consumer Airfare Report [6]

We extract the average airfare for each route from this data set. Table 1 [7] shows a list of all the attributes of the dataset. "Table 1" contains around 185,000 records from 1996 to 2017.

4.3 US Census Bureau's domestic flights from 1990 to 2009 dataset [1]

This dataset contains over 3.5 million flight records in the US from 1990 to 2009. The dataset aggregates flights between destinations on a monthly basis. It provides: date, origin airport, destination airport, number of passengers and number of seats in the flights. We utilize the distance, destinations, and number of flights in our evaluation. The distance attribute in particular was very helpful for our classifier. The complete list of attributes is given in Table 2.

5 SUBTASKS

Assembling and pre-processing our data sets was the first major task. The resulting pruned superset allowed us to then compute relationships and metrics. We then trained a classifier on the data and examined it's accuracy. Running into some challenges, we then improved our classifier to arrive at the final price estimator. The complete list of tasks is given below:

Table 2: The attributes of the flights dataset [1]

Short Name	Type	Description
Origin	String	Airport 1 code
Destination	String	Airport 2 code
Origin City	String	Origin city name
Destination City	String	Destination city name
Passengers	Integer	Passengers transported
Seats	Integer	Seats available
Flights	Integer	Number of Flights (multiple records for one month)
Distance	Integer	Distance
Fly Date	Integer	Flight date (yyyymm)
Origin Population	Integer	Origin population
Destination Population	Integer	Destination population

- Prune flight data
- Prune oil price data
- Merge oil price into flight data
- Merge population data into flight data
- Account for population growth
- Correlate flight volume to oil price
- Correlate flight price to oil price
- Explore time-delay factor
- Explore seasonality effect
- Create flight price classifier
- Train flight price classifier
- Evaluate accuracy of classifier
- Group data by airport source-destination pairs
- Retrain classifier
- Create price estimator
- Publish findings

6 REVIEW OF PROPOSED WORK

We started this project wanting to analyze the relationship between oil prices and flights. In order to do this, we obtained several data sets that we merged together. The first data set contains flight data from 1996-2009. By cleaning and merging population, oil price, and average flight price into it, we can calculate some relationships. The goal with this data set is to see if flight volume increases with cheaper fuel prices and to examine if a seasonality effect exists. Provided we found a correlation between oil prices and flight volume, we wanted to look at delays for flight prices responding to fluctuations in oil. In parallel to this, we would examine the relationship between flight volume and the U.S population as a larger population is expected to generate more air traffic. The final item with the first dataset is to use the average flight price over the 13 years to see what the correlation is between oil price and the average fare price.

The second dataset is an aggregate flights history from 1996-2017. Since this dataset contains pricing for individual routes, it lends itself to creating a classifier. By creating a classifier and training it on the flight data, we can then use it to estimate the cost of a

Table 3: The attributes added to Dataset-1

Short Name	Type	Description
ID	String	Unique id.
Year	Integer	Year
Quarter	Integer	Quarter
Oil price	Float	Oil price (quarter)
Flight fare	Float	Air fare (quarter)

plane ticket between two points based on distance, time of year, and current oil price.

7 EVALUATION

7.1 Data Preparation

We aggregated data into two datasets. The first one is based on the US Census Bureau's domestic flights dataset and contains a lot of attributes which can be used as features in our classifier. The second dataset covers a longer time period 1990-2017 and while it doesn't contain as many features, it has cheapest pricing for each route. We named the two datasets Dataset-1 and Dataset-2 respectively, and we will refer to them as such throughout this report.

To create and pre-process the two datasets we used Pandas [5] - a Python data analysis library. Pandas helped us to read the huge TSV file provided by the US Census Bureau's. The first task in preparing Dataset-1 is to combine the rows which represent the information about a route in a month into one row which represents a quarter. We had to combine them because, unfortunately, we had the average airfare in quarters form rather than months or days. Next, we used a function Pandas provides to inner-join Dataset-1 with the airfare report dataset so that we can add an airfare attribute to Dataset-1. We used inner-join to ignore any route that we don't have airfare for. Finally, we integrated the oil price into Dataset-1. Preparing Dataset-2 was a bit easier as we only had to join the various years and quarters of data (the files came separate) and add the oil price feature.

Dataset-1 covers the period from 1996 till 2009. It has over 51,000 data rows about different airline routes in the USA. Dataset-1 has most of the attributes in Table 2 except for the "Fly Date" attribute which was was changed to year and quarter in the final dataset. Table 3 shows the new attributes we added to Dataset-1.

With more than 176,000 route data points, Dataset-2 is significantly larger than Dataset-1 and covers a longer period - 1996 to 2017. We removed unnecessary attributes from the original Bureau of Transportation Statistics' Domestic Airline Consumer Airfare Report data, detailed in Table 1. The result were the 11 attributes described in Table 4.

8 RESULTS & DISCUSSION

With the two datasets ready, we calculated the correlations we were interested in. From Dataset 1, we calculated the correlation between flight volume and oil price to be 0.66 according to the Pearson correlation. This implies flight pricing is positively correlated to oil price (though does not imply causality). This is slightly unexpected as we believed lower oil prices to support more air traffic. The correlation instead implies that generally there is more air traffic

Table 4: The attributes of Dataset-2

Column Name	Example Value	Description
ID	ABEMCO19961	Unique ID
Year	1996	Data Year
quarter	1	Data Quarter
City1	Easton, PA	City 1
City2	Orlando, FL	City 2
nsmiles	906	Non-Stop Market Miles
airport_1	ABE	Airport code 1
airport_2	MCO	Airport code 2
passengers	234.945	Passenger Per day
fare	189.977	Overall average fare
oil_price	18.5569	Average oil price

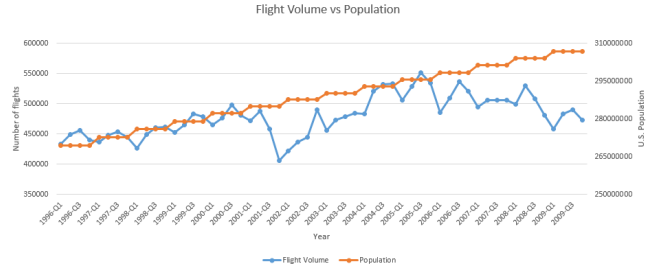
when oil is more expensive. There is some merit to this as oil price is partially dictated by demand, so a higher oil price could be the result of more people (or airlines) wanting to buy the commodity. The graph of this is given below:

Figure 1: Flight Volume and Oil Price

It is interesting to note that the graph shows the seasonality effect. We can see that most people fly during quarters 2 and 3 (summer) every year. This makes quarters 1 and 4 a preferable time to fly as plane tickets may be cheaper. Due to the quarter scale of our data, we are unable to offer advice on specific months or days.

Next, we looked at how population affects flight volume. We calculated a correlation coefficient of 0.62, meaning a positive correlation. This is consistent with our expectations of more people leading to an increase in flight volume. The graph is shown in Figure 2.

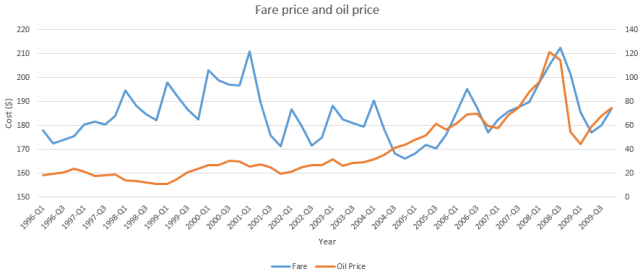
Figure 2: Flight Volume and U.S. Population



An observation here is that the flight volume decreases around 2007 despite an increasing U.S population. This is the result of the housing recession that hit the economy in 2006-2012. It can be seen that the airline industry was significantly affected by this crisis despite not having much to do with housing.

Next, we examine flight pricing. To calculate a correlation, we used the average price of a flight for each quarter. Inflation was accounted for in these numbers. The correlation coefficient we came up with was 0.27. This implies a weak positive correlation meaning oil price does not have a major influence on flight pricing. The graph is shown below:

Figure 3: Flight Ticket Pricing and Oil Price



We should note that there is not a clear time delay visible for when flight prices respond to changes in oil pricing. This could be due to the resolution of our data or due to oil price not having a huge impact on flights pricing. It is possible airlines maintain a margin to account for fluctuating fuel costs, however without more detailed data we will not be able to find out. Our best estimate is airline fares respond within 6 months of major oil fluctuations.

8.1 Classifier

For our classifier, we decided to use the Python sklearn package. After training on the data, the classifiers produce a coefficient of determination $R\hat{2}$. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a $R\hat{2}$ score of 0.0.

We originally created and trained a classifier using the Python sklearn package and Dataset-2. Our first attempt produced a very poor result with a determination coefficient $R\hat{2}$ of 0.03 when the distance attribute was used. We used three different linear classifiers (Ridge, Lasso, Linear Regression) though all had similar results. Including the other attributes did not improve this score significantly. This implied we are missing some information or attribute. After additional research, we found out that airports charge various fees for landing and stay. In order to account for this, we needed to have the classifier take note of the routes.

Since we do not airport fee information in our data, the best way to account for it was by encoding the source and destination airports as a pair. At first we tried to use "label" encoding where we encode the airport codes into values. However, this confused our classifier and the results were abysmal. Our next attempt was to use "one-hot" encoding. This type of encoding adds a large number of binary rows to the dataset (equal to number of routes). Training the classifier on the new dataset makes it only train on historical data for one source-destination route at a time. This increased our score to 0.36 (average for all routes), still low but a significant improvement over our initial attempt.

We then decided to use Dataset-1 as it had more features that could be used for training. We extracted pricing information for each route from Dataset-2 and merged it into Dataset-1. The attributes we used for training are:

- Origin
- Destination
- Passengers
- Seats
- Flights
- Distance
- Origin Population
- Destination Population
- Quarter
- Oil price

This classifier achieved a score of 0.33 (using the same three classification methods). 0.25 of this came solely from the distance attribute, correctly implying that distance has a strong influence on ticket price. The rest of the attributes on their own ranged between 0.01 to 0.04 score with oil price having nearly no effect at 0.002. This was a marked improvement from Dataset-2 but still not sufficient for price estimation.

Applying the "one-hot" binary encoding to the source-destinations in Dataset-1 improved the classifier score to 0.65. We should note that this score is the average for all routes - some routes had classifier scores in the 0.8 range while others were barely usable.

8.2 Fare Estimator

For our final piece, we created a Python program that leverages the classifier to estimate the price of a plane ticket. The user enters two airport codes, the current price of oil, and quarter they wish to fly in. The tool then uses output from the classifier to predict how much such a ticket will cost if booked under ideal conditions.

9 CHALLENGES AND LIMITATIONS

9.1 Resolution

While processing the data, we ran into an issue - all flight data was quarterly for both of our compiled datasets. Because of this, we weren't able to accurately calculate how long it takes for flight prices to be affected following a change in oil pricing. We found that the two are correlated but the exact delay for the correlation is unknown. Our best estimate is that air flight fares react to oil pricing within 6 months.

Due to the quarterly resolution of our data, we could only see the seasonality effect at a quarterly level. Originally we had hoped to be able to accurately predict the best and worst times to travel down to the month (or even day) resolution but our data does not support this.

The resolution also affected our classifier as a better resolution would have likely increased our accuracy and allowed our fare prediction tool to be far more specific. Fare prices vary wildly from

day to day (weekends and holidays are significantly more expensive) and having the capability to predict that would be useful.

9.2 Data

Originally Dataset-2 seemed like it would have all the information we could ever want to train an accurate fare classifier and create a fare estimation tool. After multiple attempts training different classifiers, we were unable to obtain good results. This necessitated moving to Dataset-1 and extrapolating pricing for all routes from Dataset-2 as Dataset-1 had more features to train. We were still unable to create a great classifier (0.33 score) as the airport costs (and possibly other variables) were not present in any of the data. By treating each route individually, we greatly improved the accuracy of the classifier (0.65 score over all known routes) as airport costs were now accounted for. However, this created limitations - we would only be able to predict prices for existing routes.

Our data only spans 1996 - 2009. Because of this, we do not have a lot of data for each route. Routes with more data result in better accuracy while routes with less data suffer - if our classifier score is under 0.5 for a given route, we do not use it. Additionally, our tool only can predict fares for routes that already exist in our database. Any airports or routes not in our database are not currently supported.

9.3 Classifier

In addition to the challenges with training our classifiers, we attempted to create a support vector machine (SVT) classifier. Due to the large amount of data, we were never able to complete training it. Fit time increases faster than exponentially with large data and our millions of records were able to overwhelm even capable desktops. Since the classifiers we tried did not achieve high scores until we started using "one-hot" encoding, our classifier can only predict fares for already existing routes.

10 FUTURE WORK

In the future, we would like to look around for a data set that is better suited for estimating fares. The data sets we used in this investigation provided ample statistics on flight volume in the continental US but did not contain pricing data broken down into it's components. For example, we did not have airport fees, type of plane, a detailed time resolution, or other fees we may not know about.

While we trained (or attempted to train) a number of classifiers, we did not find one with a perfect fit. Given the breadth of options sklearn offers, we would like to try a bayesian, quadratic (or other non-linear) classifier, and a neural net.

11 CONCLUSION

With some of the cheapest flights in history, now is arguably the best time to fly. More airline carriers are flying to more destinations than ever before, producing record air traffic. With an ongoing oil glut projected to keep the price of oil low (compared to the last 15 years), this is unlikely to change soon. In this study, we characterize the relationship between the price of oil and flights. We also look at how volume and pricing get affected by oil. Specifically, we answered the following questions:

- Does flight volume increase with cheaper fuel prices? No, positive correlation of 0.66 says they tend to rise and fall in sync.
- Are flight prices correlated to fuel prices? Yes, positive 0.273 correlation indicates they move in sync.
- What is the delay between oil price increases and flight price increases if any? We were unable to see a delay, perhaps due to our quarterly scale.
- What is the correlation between fuel prices and flight volume? Positive 0.66 correlation.
- What is the correlation between population increases and flight volume? Positive 0.62 correlation. Population likely increases flight volume (correlation does not imply causation)
- What is the best season to fly? Quarter 4 and Quarter 1.

We then used our data to train a classifier. The classifier trained on passengers, seats, number of flights, distance, origin population, destination population, quarter, and oil price. The classifier was provided the fare for each route as the learning variable it would later try to predict. We encoded the source and destination airports using the "one-hot" encoding in order to preserve airport fees as we did not have these fees as additional training features. While we had a number of issues with the classifier, in the end we were able to create classifier that achieved an $R\hat{2}$ score of 0.65 averaged over all flight routes.

We leveraged our classifier to create a price fare estimator for users. Inputting a few simple arguments such as source, destination, oil price, and quarter allows the tool the calculate an estimate for what such a flight would cost when booked under ideal conditions. While our tool does not support every airport in the world (most US airports are supported), it is useful as a proof of concept and a great first attempt. Given additional time and more detailed cost data for flights, our tool can be expanded to be very accurate in estimating current and future flight paths.

12 TEAM WORK AND CONTRIBUTIONS

The team held weekly face-to-face meetings along with numerous virtual meetings. Meetings generally consisted of current status of items, integration of the work, a discussion of current and future issues, and a discussion about what needs to be done by next meeting. A brief summary of contributions is given below:

- Raw Data Collection (All): In the first phase of the project, everyone searched for reliable sources for the project's datasets and possible data mining ideas. We came up with multiple great ideas for projects though could not find datasets for most of them. Eventually we settled on the flight datasets described in Section 2.
- Data Preparation/Processing:
  - Data Cleaning: Mahmoud pruned datasets and filled missing values.
  - Data Integration: Alatawi worked on integrating the US Census Bureau's domestic flights' dataset (Section 2.3) with the oil prices dataset to produce Dataset-1 and Steve worked on integrating the Bureau of Transportation Statistics' Domestic Airline Consumer Airfare Report (Section 4.2) with the oil prices dataset. Alatawi also worked on

calculating correlation between different attributes in the datasets.
  - Data Selection: During the analysis, we voted as a team for the attributes that we were going to keep.
  - Data Transformation: One of the challenges we faced is that each dataset had different resolution for time. Mahmoud worked converting the data to use a common time scale - quarters.
- Classifier: Alatawi created the initial classifier. However, it wasn't accurate so all team members jumped in and trained different classifiers with a varying list of features. None of the ones tested achieved a good score with any combination of features used.
- Improved Classifier: The idea to use "one-hot" encoding for routes came up during a meeting after the team had tested many classifiers unsuccessfully. Alatawi helped to encode Dataset-1 to the desired format and Mahmoud helped with training the classifier on the new data.
- Fare Estimation Tool: Mahmoud created the tool and integrated it with the classifier.
- Data Presentation: Steve worked on creating the different correlation graphs and their interpretation. Steve also created the presentations and compiled this report with help from all team members.

REFERENCES

[1] US Census Bureau. [n. d.]. US domestic flights from 1990 to 2009. ([n. d.]). http://academictorrents.com/details/a2ccf94bbb4af222bf8e69dad60a68a29f310d9a.

[2] Federal Reserve Economic Data. [n. d.]. Crude Oil Prices: Brent - Europe. ([n. d.]). https://fred.stlouisfed.org/series/DCOILBRENTEU.

[3] John Hansman. [n. d.]. The Impact of Oil Prices on the Air Transportation Industry. ([n. d.]). http://www.nextor.org/pubs/NEXTOR-II-Oil-Impact-3-2014.pdf.

[4] Consumer Price Index. [n. d.]. Average Price of Flights per year. ([n. d.]). http://cpi.mooseroots.com.

[5] NUMFocus. [n. d.]. Pandas a Python Data Analysis Library. ([n. d.]). http://pandas.pydata.org/index.html.

[6] Bureau of Transportation Statistics. [n. d.]. Domestic Airline Consumer Airfare Report. ([n. d.]). https://www.transportation.gov/policy/aviation-policy/domestic-airline-consumer-airfare-report.

[7] Bureau of Transportation Statistics. [n. d.]. Domestic Airline Consumer Airfare Report (METADATA). ([n. d.]). https://www.transportation.gov/office-policy/aviation-policy/domestic-airline-consumer-airfare-report-metadata.

[8] Wikipedia. [n. d.]. Brent Crude. ([n. d.]). https://en.wikipedia.org/wiki/Brent_Crude.