A Network Tour to Flight Delay in the US

FENG Wentao, FU Yan, SUN Zhaodong, WANG Yunbei

INTRODUCTION

What we want to solve:

- What will affect on the delay of a flight, could it be departure city, date, hour or airlines?
- Can we predict how long will a flight delay with given certain situations?
- If someone wants to fly from city A to city B, can we give some advice on choosing flights based on the delay rate?

DATA EXPLORATION

Overview of Data set

1. Openflight Data set

   Obtained from: https://openflights.org/data.html
   67,663 routes, 3,321 airports from 224 countries

1. Flight Delay Data set

   Collected by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics, download from Kaggle: https://www.kaggle.com/usdot/flight-delays#flights.csv

   Detailed flight information (including delay time) of over 5,800,000 flights in 2015 in the US

WHOLE WORLD ROUTES DATA

OpenFlight Data Set

Which country is the busiest flight country?

- from ‘equipment’ columns of each route find the aircraft type to estimate passengers
- eg. 380 for airbus a380, with 555 seats

Src_IATA	Src_ID	Des_IATA	Des_ID	Stops	Equipment	Capacity
CDG	1382.0	KUL	3304.0	0	380	555.0

- **US is the country with largest flight passengers!**

Country	Total Flight Passengers
United States	3.536720e+06
China	2.619732e+06
United Kingdom	9.498741e+05
Spain	8.445502e+05
Germany	7.737909e+05

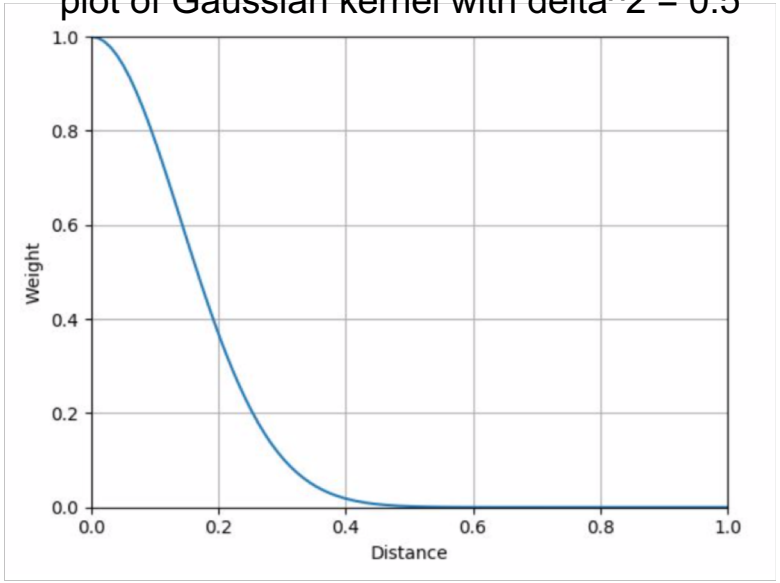
Graph construction

$$DISTANCE = 2 * arcsin\sqrt{sin^2\frac{a}{2} + cos(Lat1) * cos(Lat2) * sin^2\frac{b}{2}} * EarthRadius$$

$$a = Lat1 - Lat2, b = Lon1 - Lon2$$

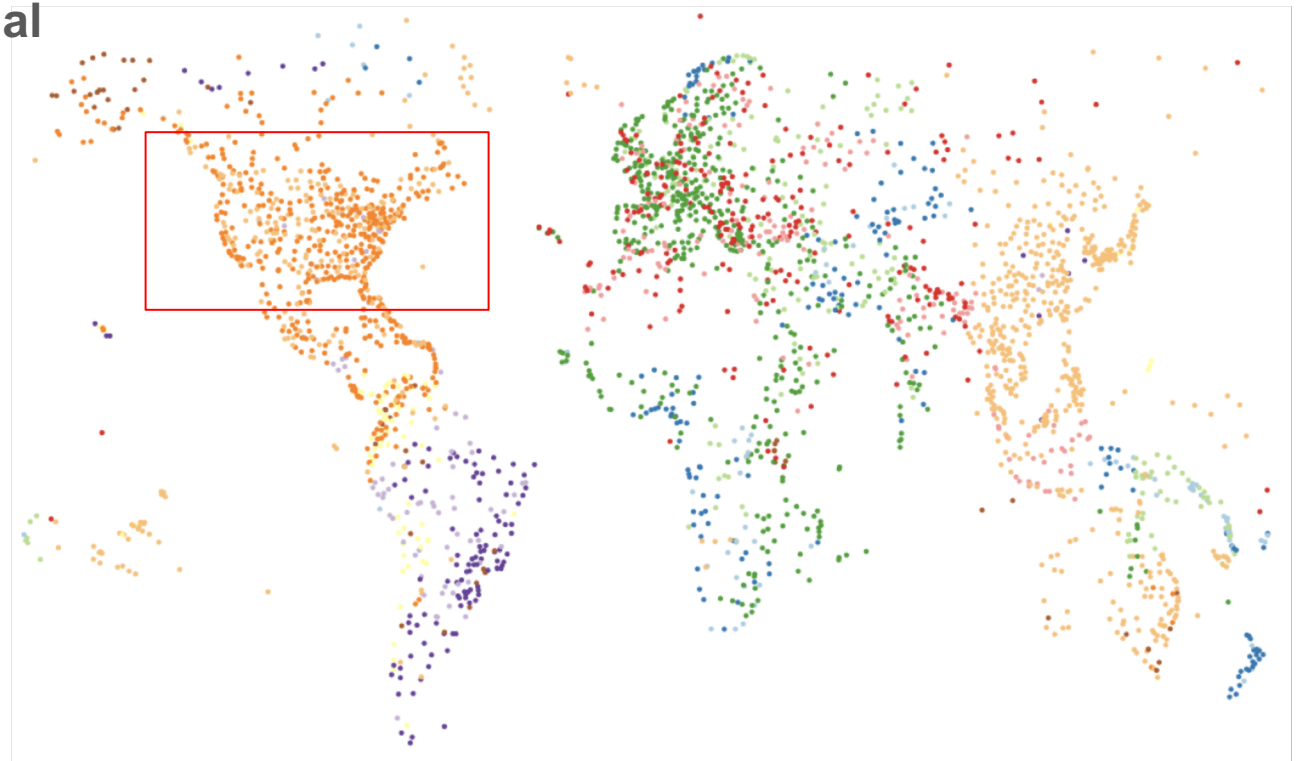
$$W = exp(-\frac{Distance^2}{\delta^2})$$

plot of Gaussian kernel with delta^2 = 0.5



OpenFlight Data Set

- 97 clusters
- The world is connected --
  BUT it is still regional
- zoom in the US



US DELAY DATA

Flight Delay data set (data pre-processing)

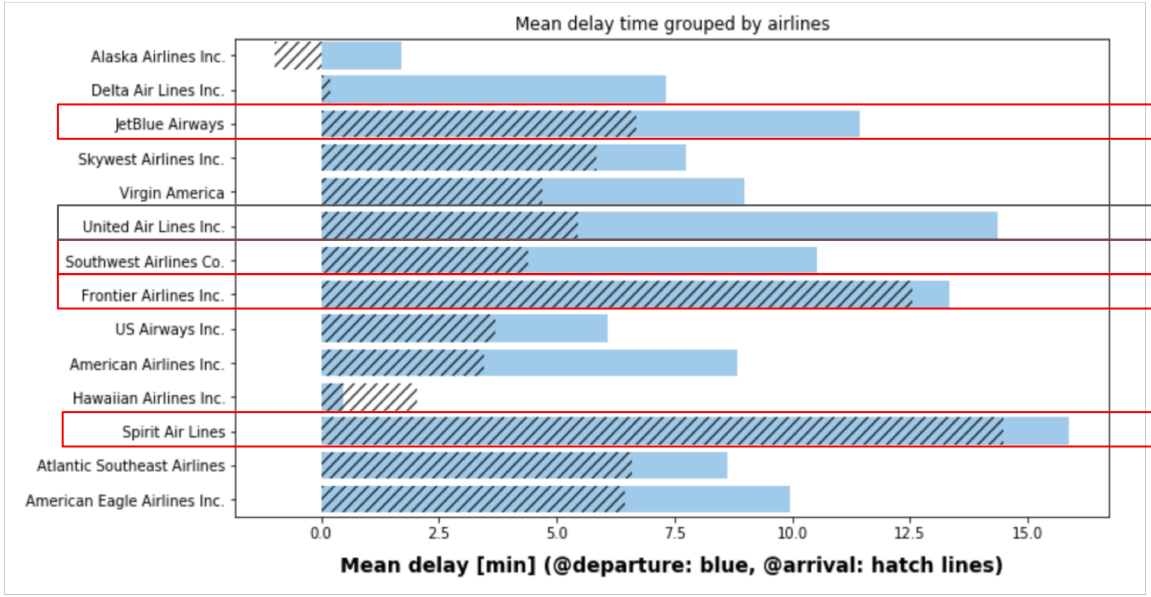
- data cleaning(WHY US)
- data type transformation

DATE	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	SCHEDULED_ARRIVAL
2015-01-01	1	1	4	AS	98	ANC	SEA	00:05:00	04:30:00

DEPARTURE_DELAY	ARRIVAL_DELAY	AIR_TIME	DISTANCE	time_hour	DEPARTURE_CITY	DEPARTURE_STATE	DESTINATION_CITY	DESTINATION_STATE
-11.0	-22.0	169.0	1448	0	Anchorage	AK	Seattle	WA

Flight Delay data set (Statistic Analysis)

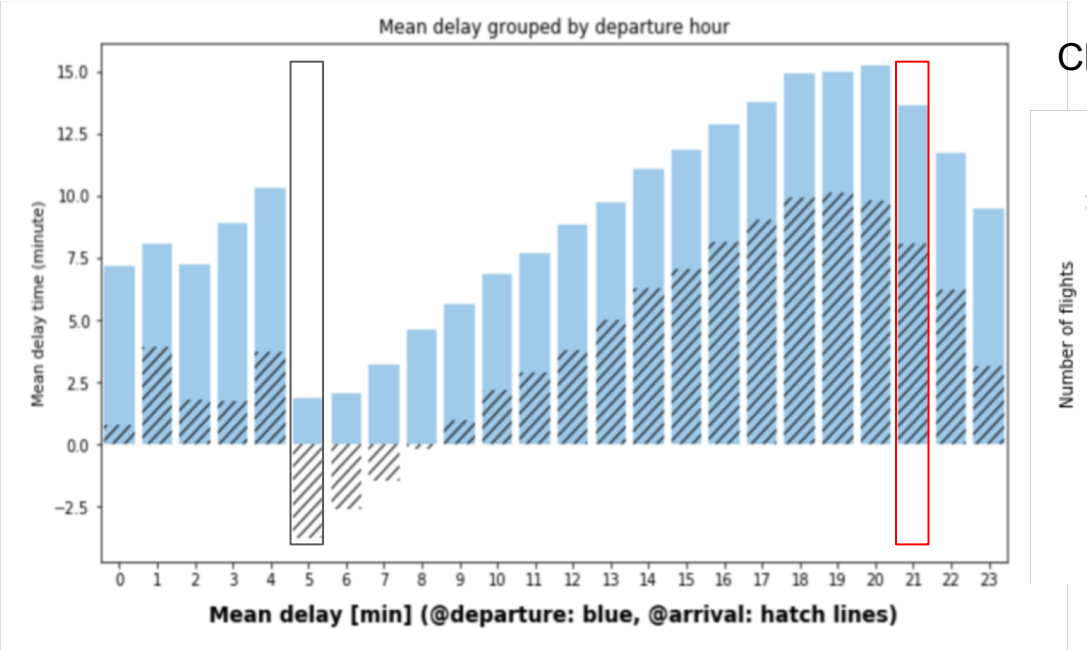
- averaged delay time of different airlines (CONCLUSION)



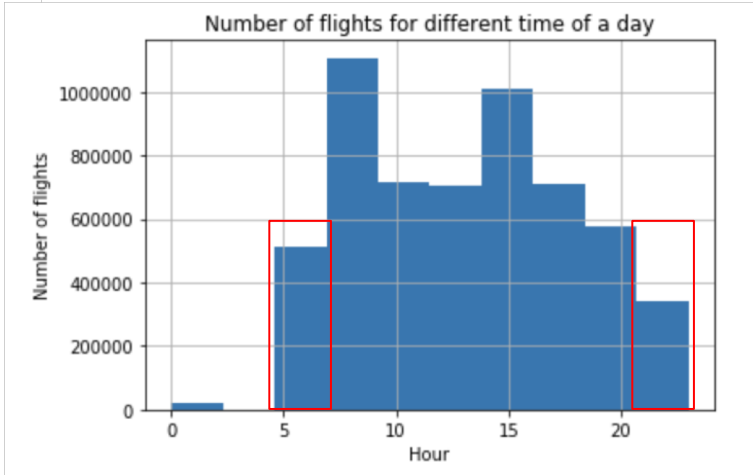
In a hurry:
do not choose low-cost airlines

Flight Delay data set (Statistic Analysis)

- averaged delay time of a day

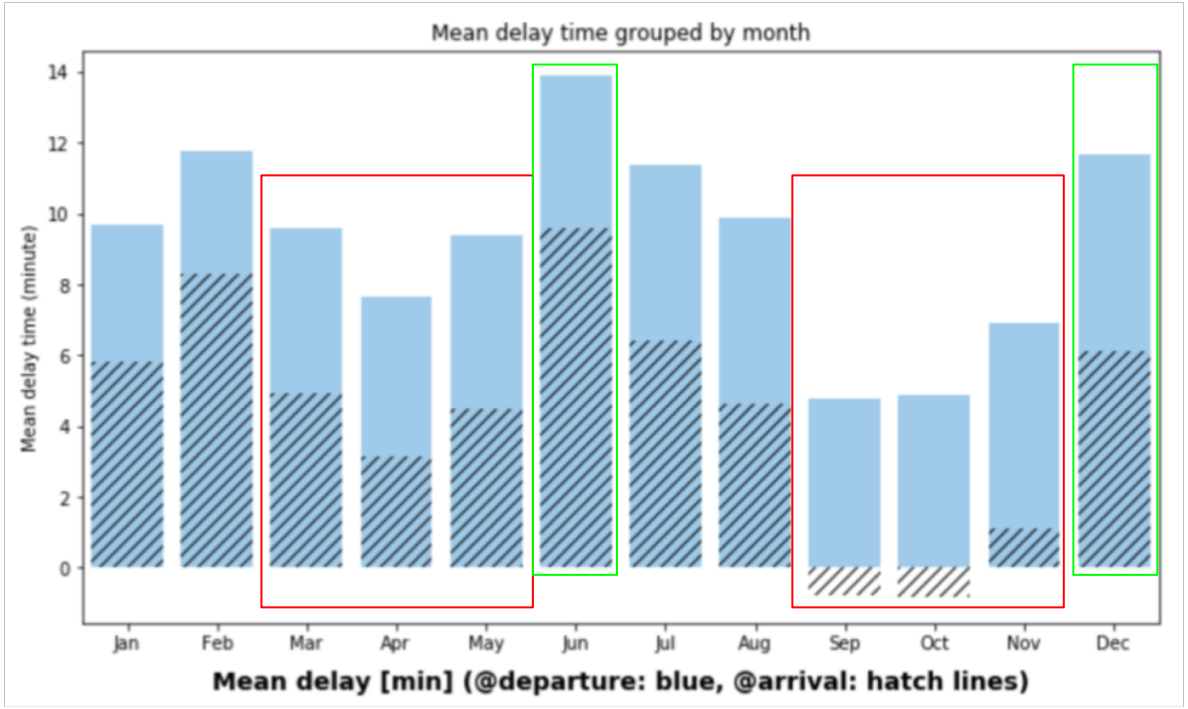


Choose earlier flight (less delay time)



Flight Delay data set (Statistic Analysis)

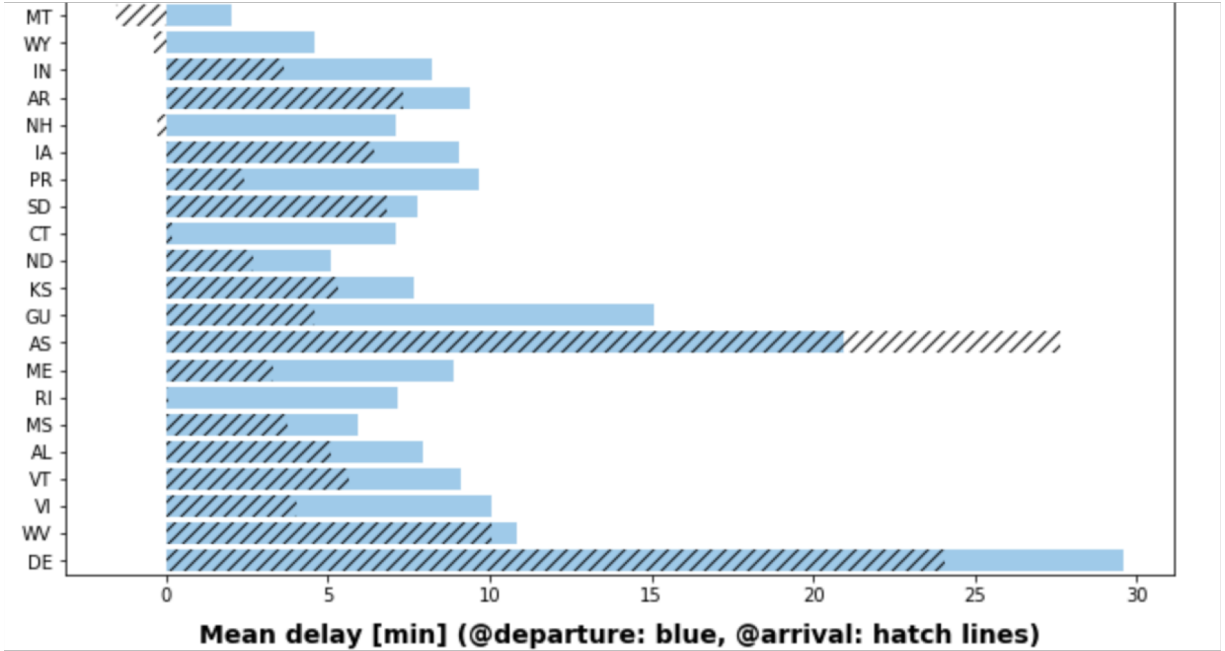
- averaged delay time of different months



- summer and winter have longer delay time -- weather
- June and December have highest delay time -- vacation

Flight Delay data set (Statistic Analysis)

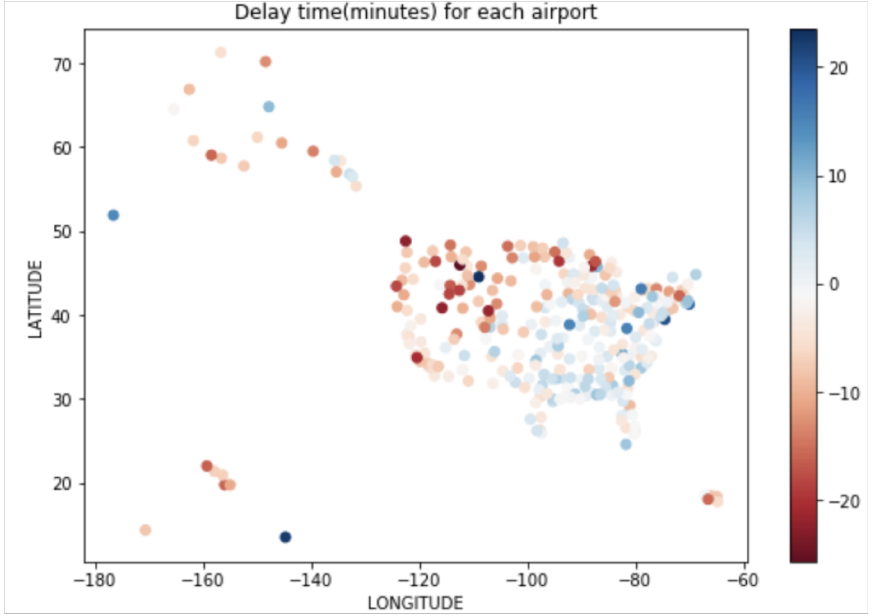
- averaged delay time of different states



the location of airports may also be influential in flights delay.

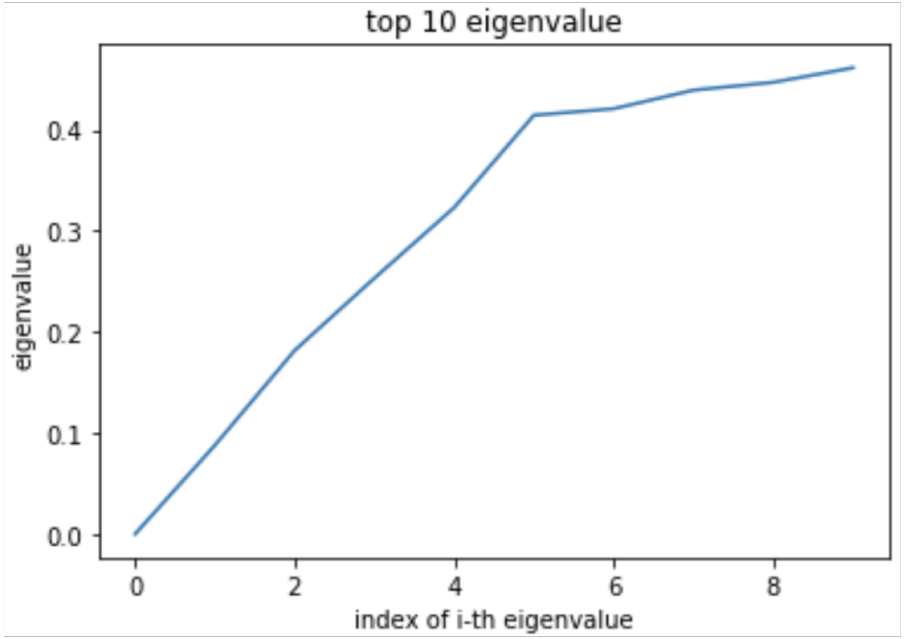
AIRPORT LOCATION AND DELAY

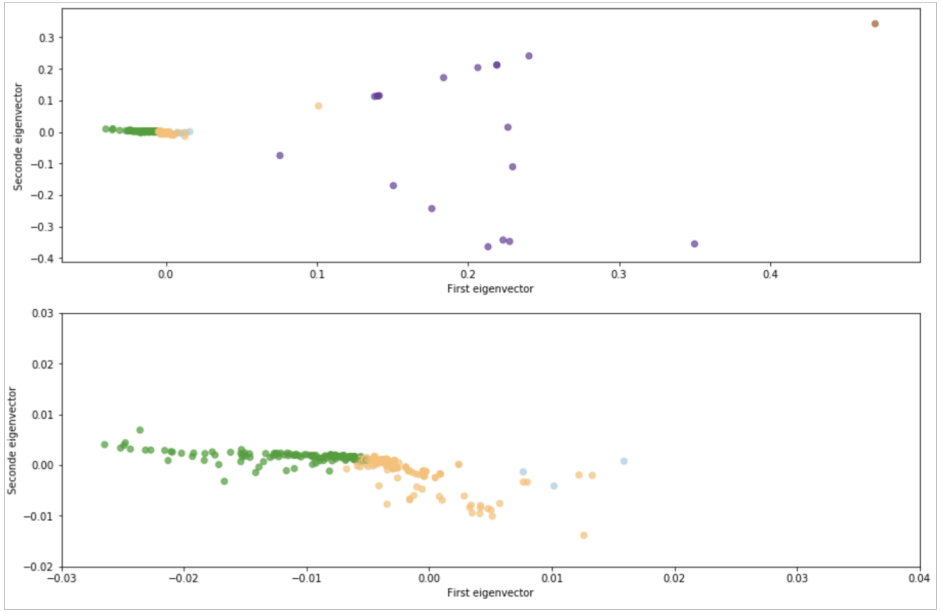
Location and Delay



- Let’s go back to our graph: airport as nodes, Gaussian distance as weight
- Averaged US airport delay on the map
- Different delay patterns in the west and east coast
- delay is related to location
- Delay lower than 10 min is considered as no delay

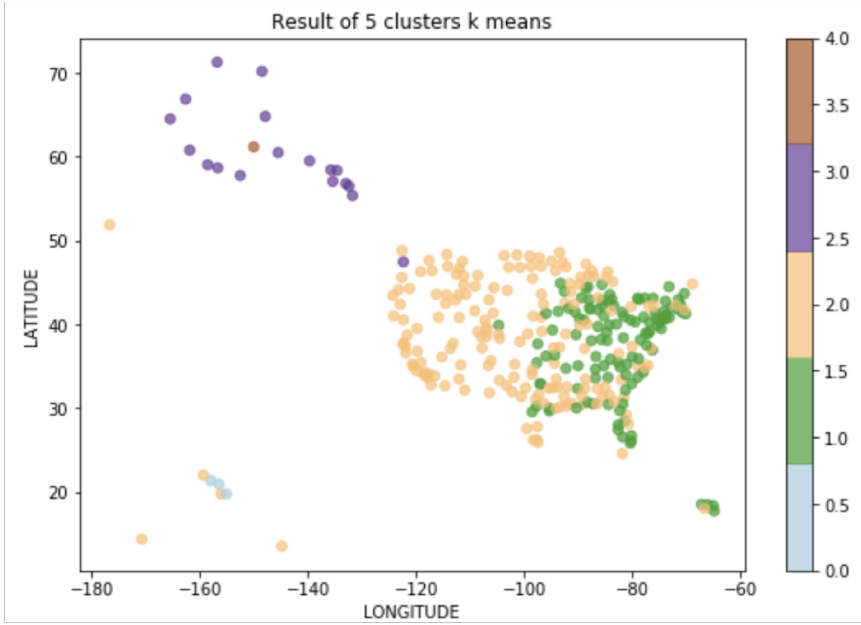
From Spectral Clustering to Delay Analysis

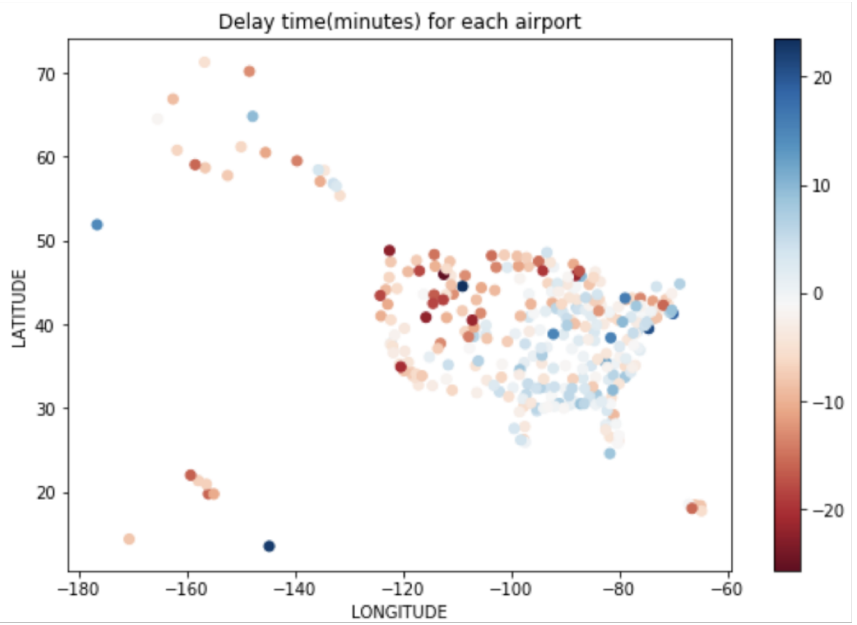




5 clusters

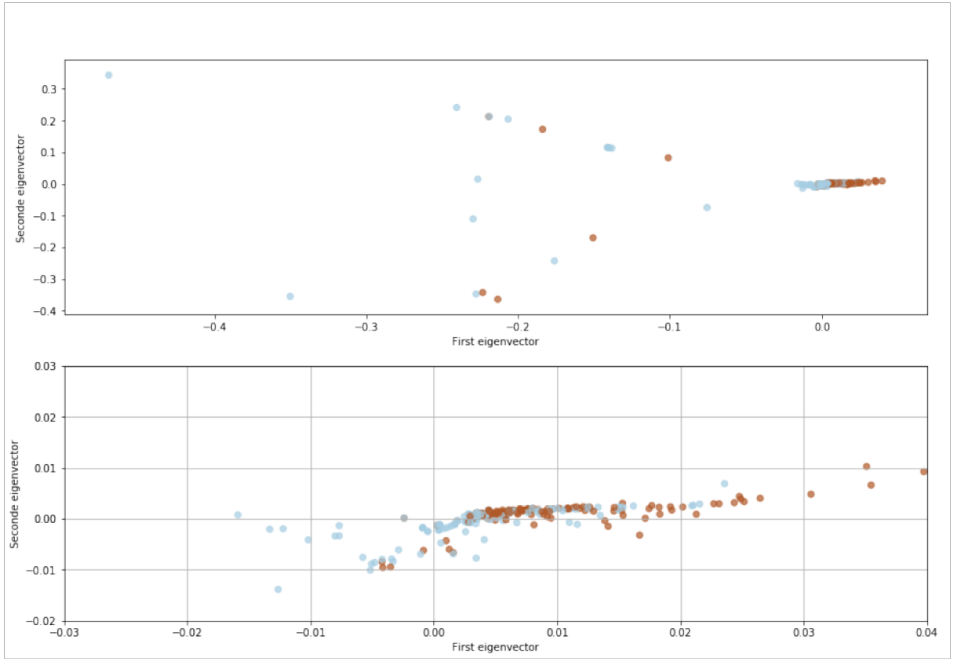
From Spectral Clustering to Delay Analysis

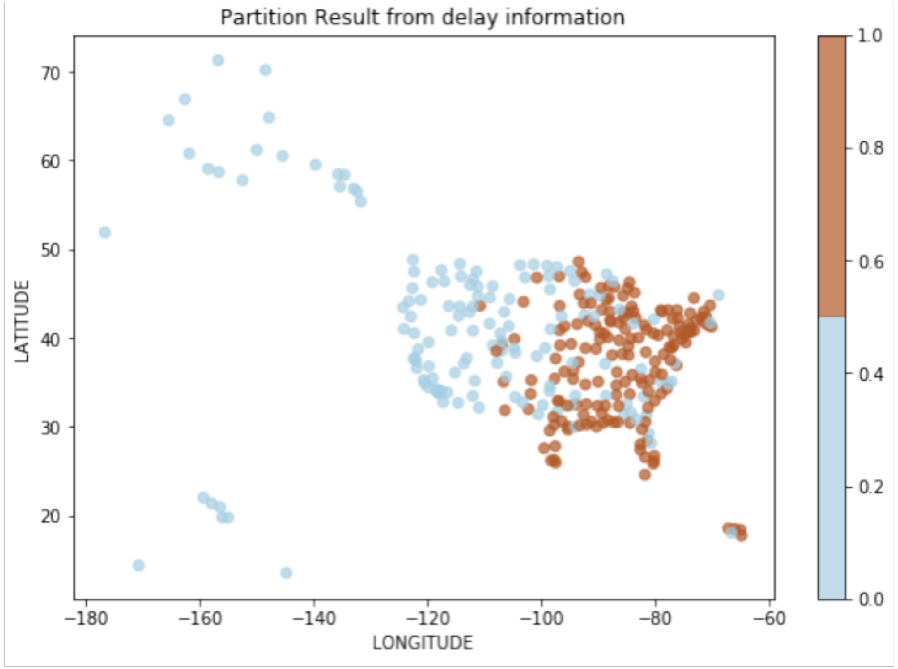




Accuracy 64.3%

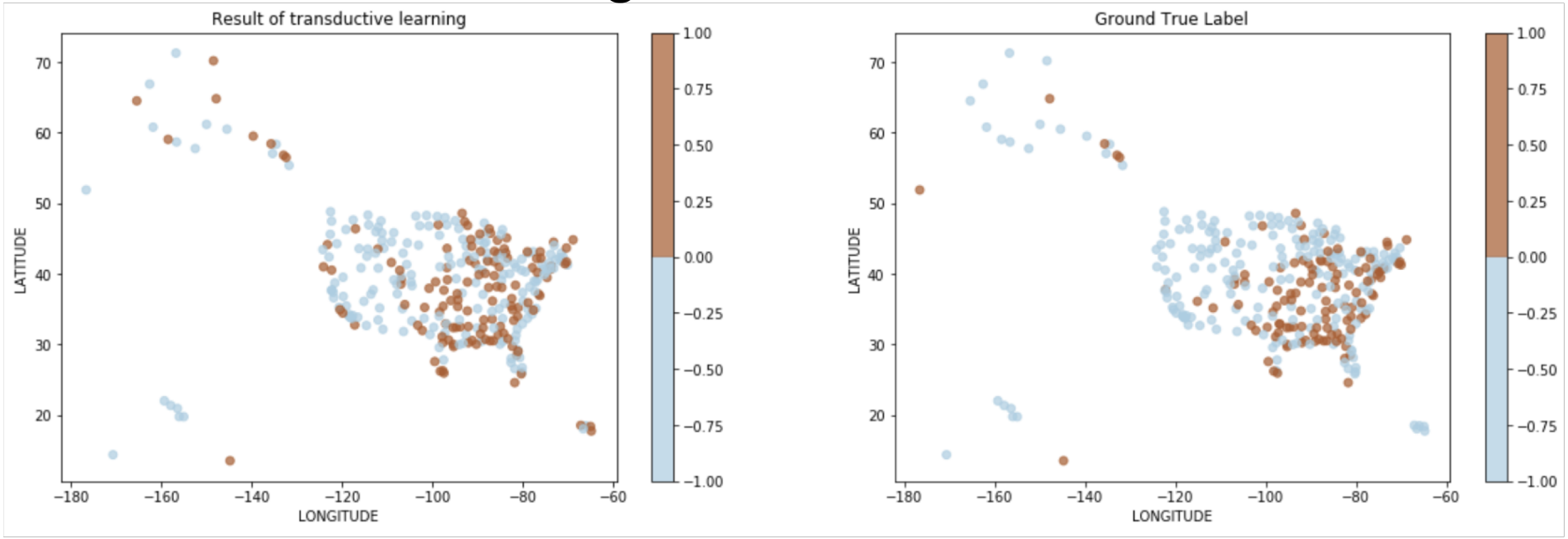
From Delay Patterns to Graph Partition





Accuracy: 69%

Transductive learning



Accuracy 72.3% recovered from 30%
Can we recover worldwide delay based on US delay?
US airports/all OpenFlight airports = 13%; accuracy 56% recovered from 13%
Not wise!

RECOMMENDATION SYSTEM

What if we combine all features … ?

Predict delay

- Data pre-processing
- Feature generation
- Categorical feature
- Model selection
- Results

- What we keep?
- delay between -30 and 30.

- Why we sample these points?
- Reduce the size of dataset.
- Filter outliers

Predict delay

- Data pre-processing
- Feature generation
- Categorical feature
- Model selection
- Results

e.g.
1 January 2015 is:
Thursday (`DAY_OF_WEEK = 4`)
the first day of 2015 (`DAY_OF_YEAR = 1`)
in the first week of 2015 (`WEEK_OF_YEAR` = 1)

Predict delay

- Data pre-processing
- Feature generation
- Categorical feature
- Model selection
- Results

Light Gradient Boosting Machine

Predict delay

- Data pre-processing
- Feature generation
- Categorical feature
- Model selection
- Results

Root Mean Square Error:

- *2.12* on the train set
- *2.31* on the test set

Predict delay - demo

CONCLUSION

- Created our graph using airports as nodes, the distance between airports after Gaussian kernel as weights.
- Spectral clustering and transductive learning to partition the graph and predict delay
- Developed a recommendation system using both geographic and time features