Airport Delay Prediction using Graph Based Machine Learning

Kayla Sanderson and Nik Gudmundsson

Presentation Overview

- Project Overview
- Introduction to the Problem
- Data Sets
- Data Exploration
- Preprocessing
- Graph Structure
- Methodology
- Results
- Conclusions and Future Work

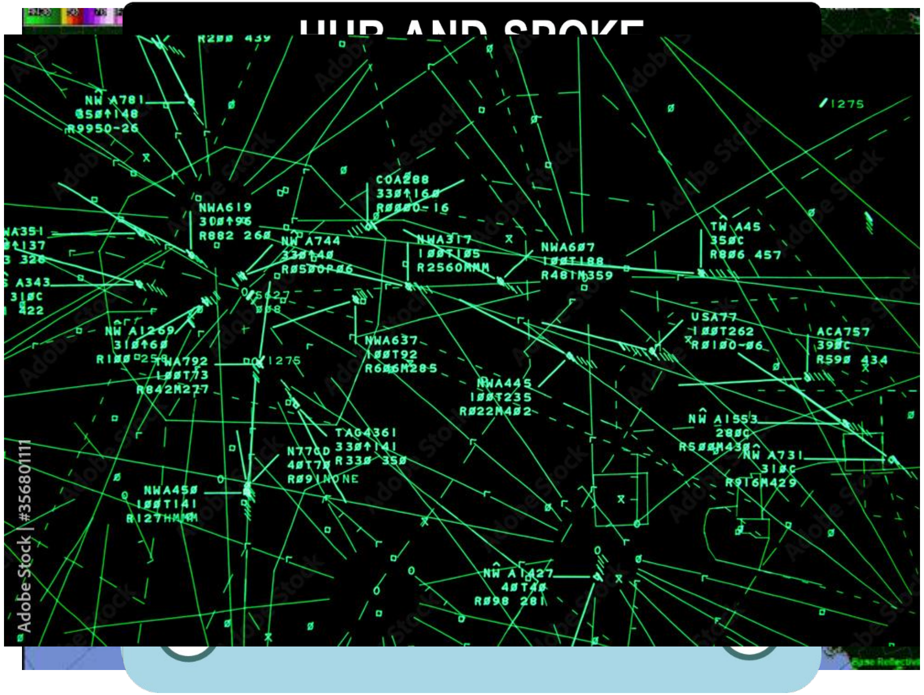
Project Overview

Our Project Goals:

- Replicate ideas and models developed and proposed in “A Geographical and Operational Deep Graph Convolution Approach for Flight Delay Prediction”
- Apply the model on a US based data set as opposed to a Chinese data set to see how well some of these models generalize to and work in a different airspace
- The models we looked into focused on predicting average delay which comes with different applications and motivations than for predicting for individual flights

Motivation

- Airport level delays have a relationship with individual flights
  - Delay causes tend to propagate
- May be possible for carriers and governing bodies to identify points of weakness within the air network
- Relationship between airport level delays and safety
  - Increased ground and air complexity



Data Sets

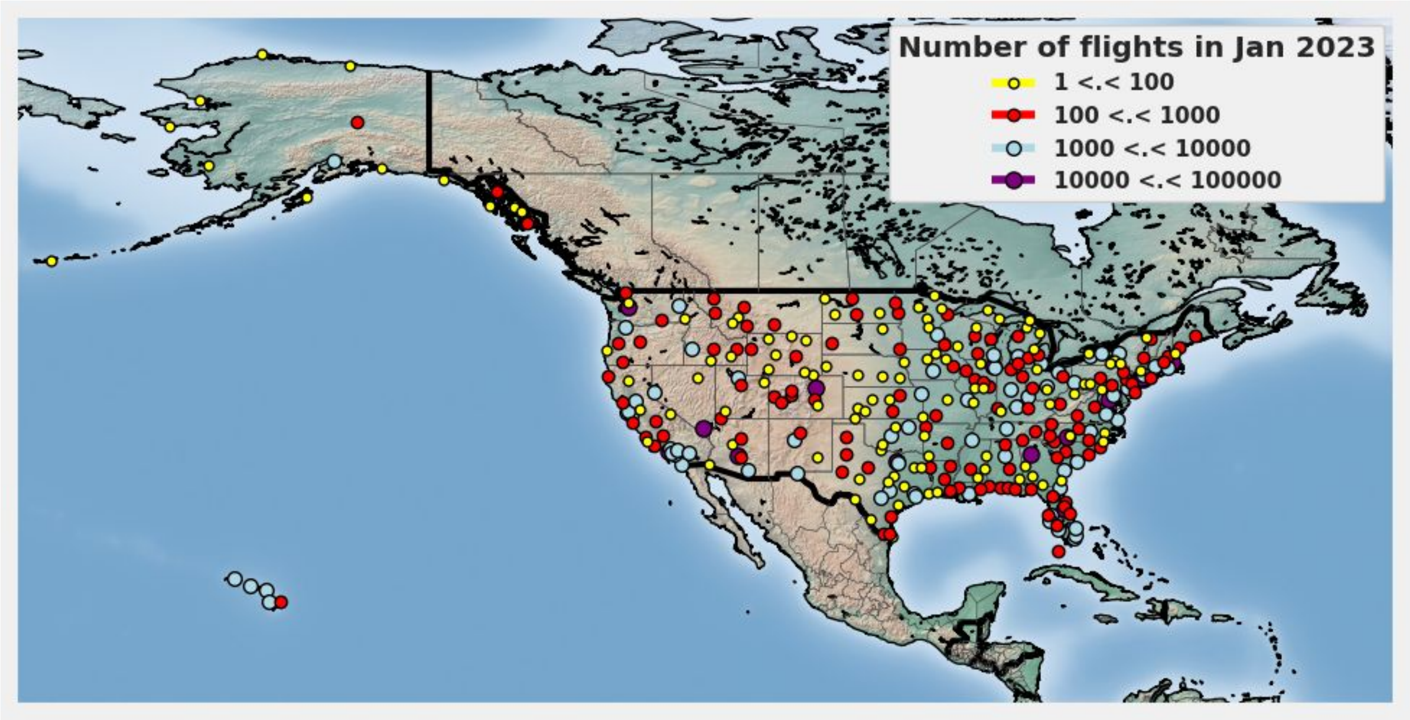
- Reporting Carrier On-Time Performance Data Set
  - Collected by USDOT's Bureau of Transportation Statistics
  - Data set includes information for all reported domestic flights
  - 109 features, of which we identified 22 of interest
  - Data set subsetted to include data from January 2023 due to large volume

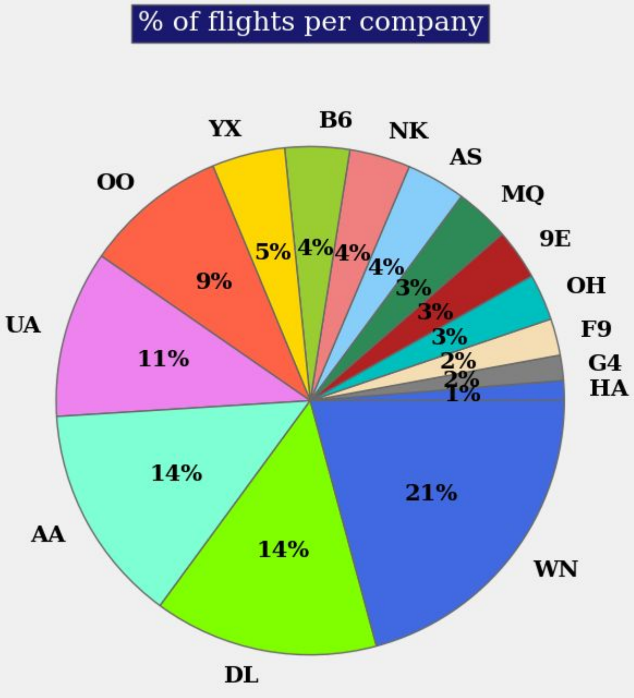
	OP_UNIQUE_CARRIER	ORIGIN_AIRPORT_ID	ORIGIN	DEST_AIRPORT_ID	DEST	CRS_DEP_TIME	DEP_DELAY	CRS_ARR_TIME	ARR_DELAY	CANCELLED	...
958	DL	10140	ABQ	10397	ATL	710	-10.0	1209	-9.0	0.0	...
1059	WN	10140	ABQ	11292	DEN	525	106.0	655	81.0	0.0	...
1079	F9	10397	ATL	11298	DFW	505	7.0	639	-12.0	0.0	...
1005	AA	10423	AUS	13204	MCO	616	-12.0	944	-2.0	0.0	...
1020	AA	10423	AUS	15304	TPA	615	-10.0	930	-20.0	0.0	...

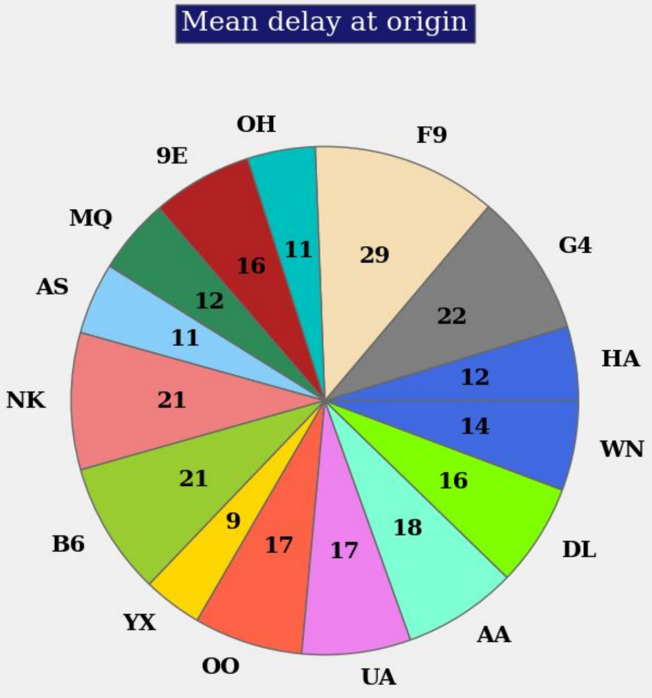
Data Sets

- US Airports Data Set
  - Open source data set of airports within the US
  - 23 features, of which we used 4 all related to geographic location
  - Data set used to create a dictionary for geolocation
- Open Travel Data Airport Time Zone Data Set
  - Open source data set of airports and their time zones around the globe
  - 5 features, of which we use 2
  - Data set used to append time zones to the airports and perform time corrections

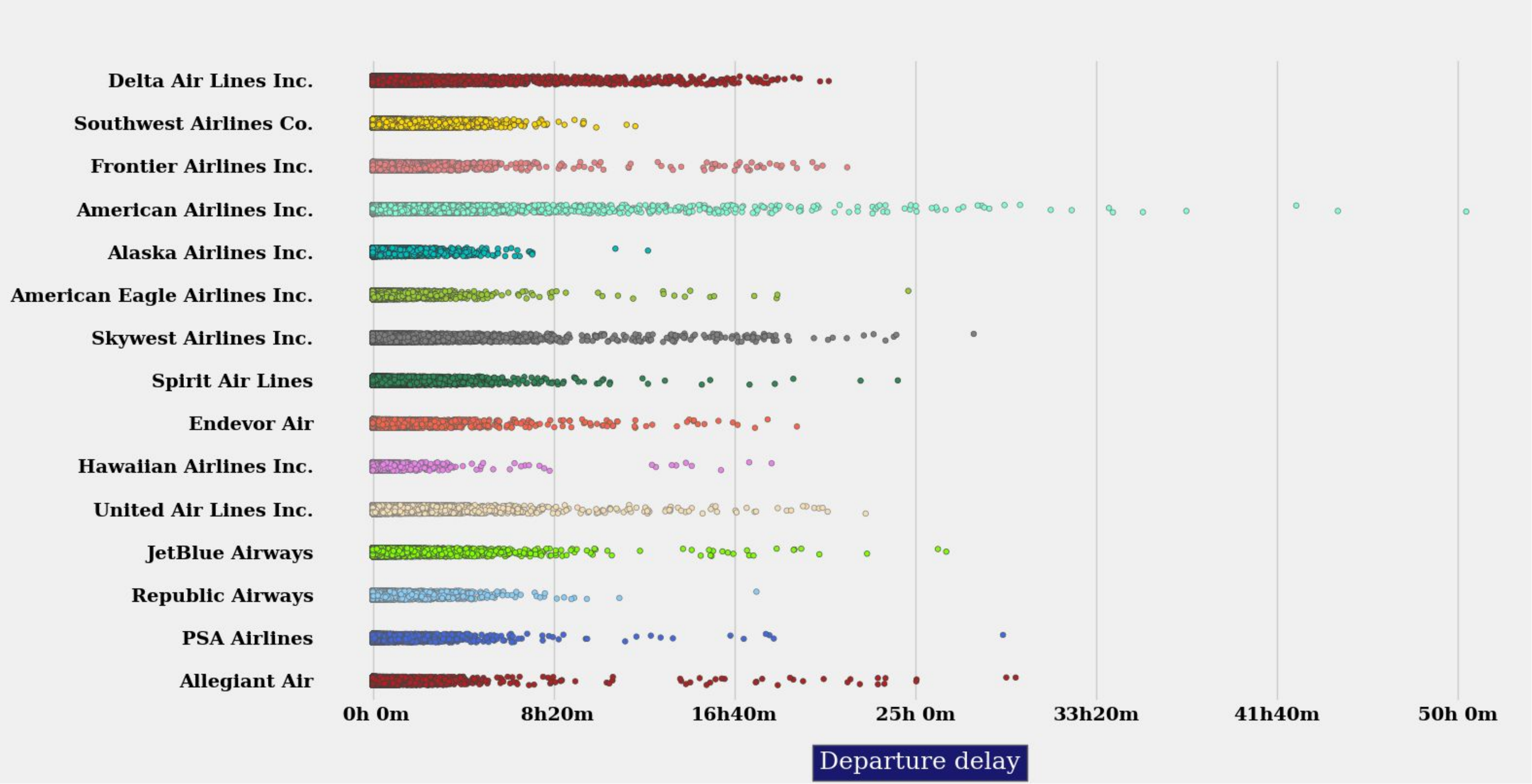
Data Exploration



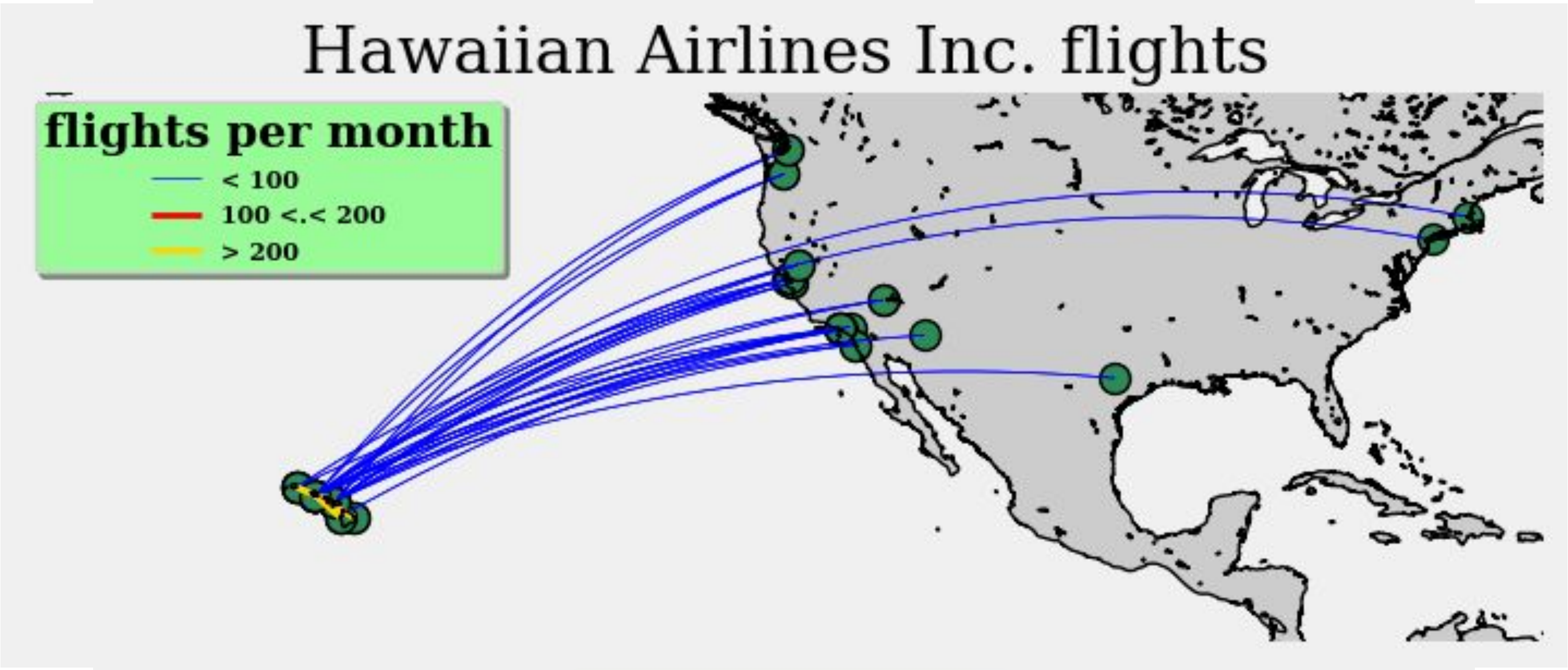




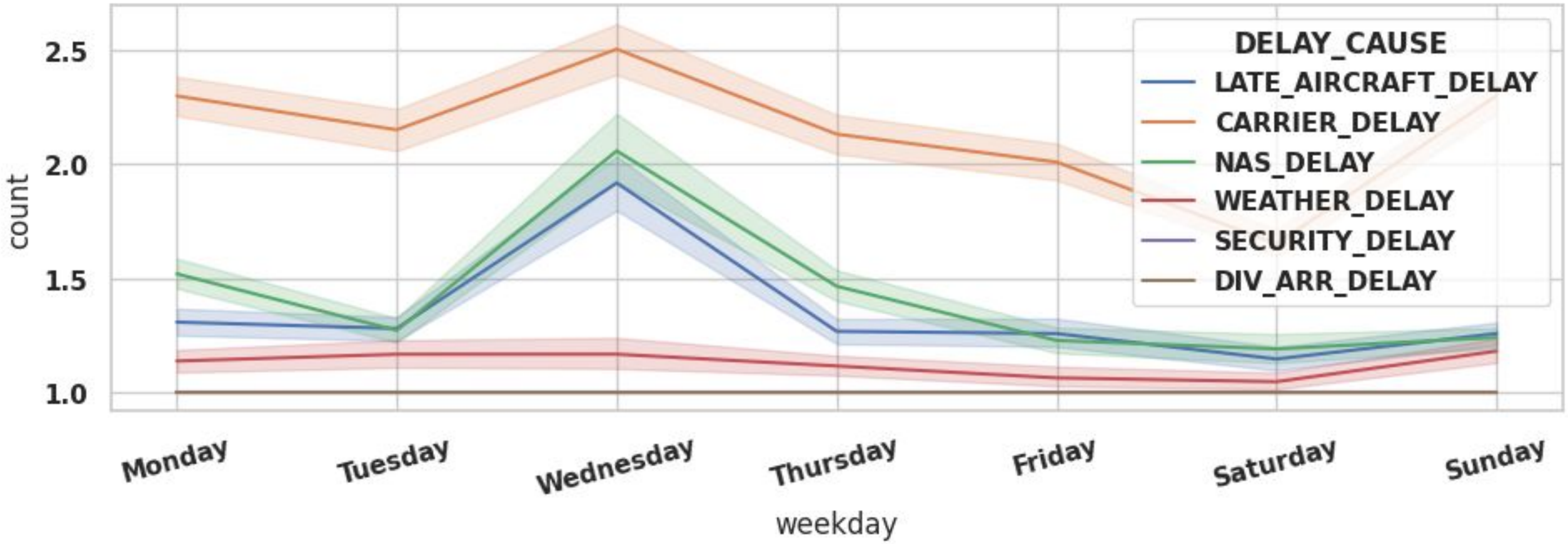
IATA_CODE	AIRLINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
YX	Republic Airways
HA	Hawaiian Airlines Inc.
MQ	American Eagle Airlines Inc.
9E	Endevor Air
OH	PSA Airlines
G4	Allegiant Air

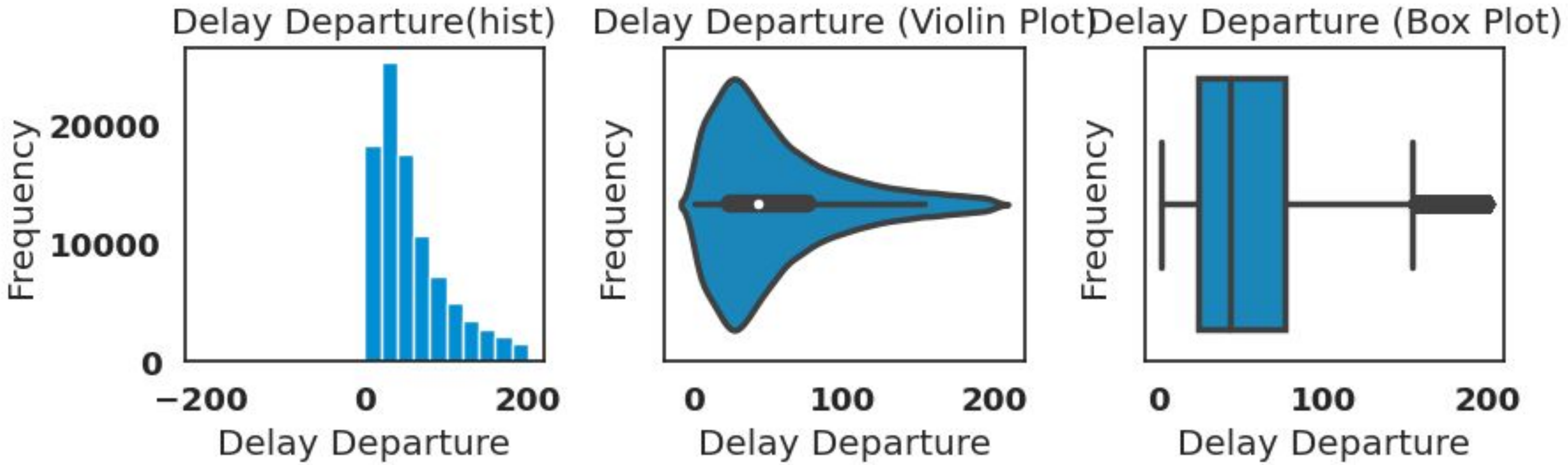

Delta Air Lines Inc.
Southwest Airlines Co.
Frontier Airlines Inc.
American Airlines Inc.
Alaska Airlines Inc.
American Eagle Airlines Inc.
Skywest Airlines Inc.
Spirit Air Lines
Endevor Air
Hawaiian Airlines Inc.
United Air Lines Inc.
JetBlue Airways
Republic Airways
PSA Airlines
Allegiant Air
0h 0m
8h20m
16h40m
25h 0m
33h20m
41h40m
50h 0m
Departure delay

Data Exploration



Data Exploration




Delay Departure(hist)
Frequency
20000
10000
0
−200
0
200
Delay Departure
Delay Departure (Violin Plot)
Frequency
0
100
200
Delay Departure
Delay Departure (Box Plot)
Frequency
0
100
200
Delay Departure

Graph Structure

Operational Graph

- Graphs generated at 15 minute time steps
- Graph takes the form of : $G^{(t)} = (V, E^{(t)})$
  - Where V are the airports and E are the edges
  - Edges logically vary with the time step
  - Edges are created between airports with an existing flight between them and, in the case for the input into the adapted Operational GCN, with a haversine distance of less than the parameter rho (ϱ)
    - We set rho to 165 miles (~265.542 km)
  - The set of neighbors for the general GCN and Operational GCN respectively are defined as: $N_f(v) = \{u\}$and $N_f(v) = \{u | d(z_u, z_v) \leq \rho\}$

Graph Structure

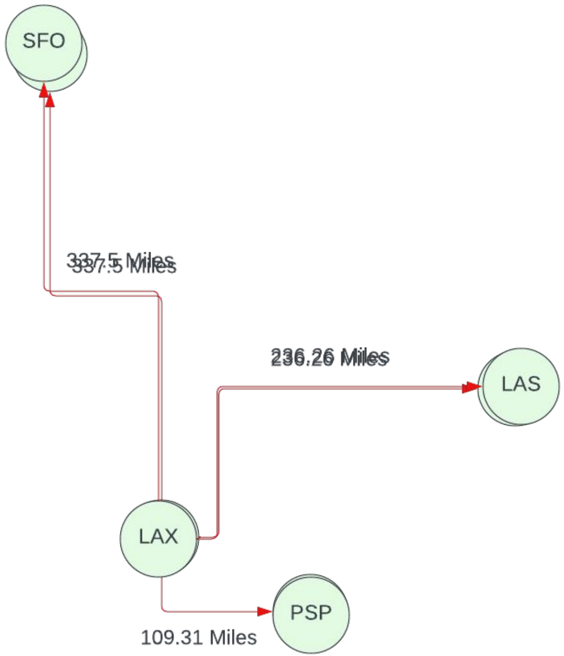
	Lat	Long
LAX	33.94	-118.40
SFO	37.61	-122.37
LAS	36.08	-115.15
PSP	33.82	-116.50

Example of Node Location Information

Graph Structure

	Origin	Dest	Dist	Delay
Flight 1	LAX	SFO	337.5	2
Flight 2	LAX	LAS	236.26	4
Flight 3	LAX	PSP	109.31	0

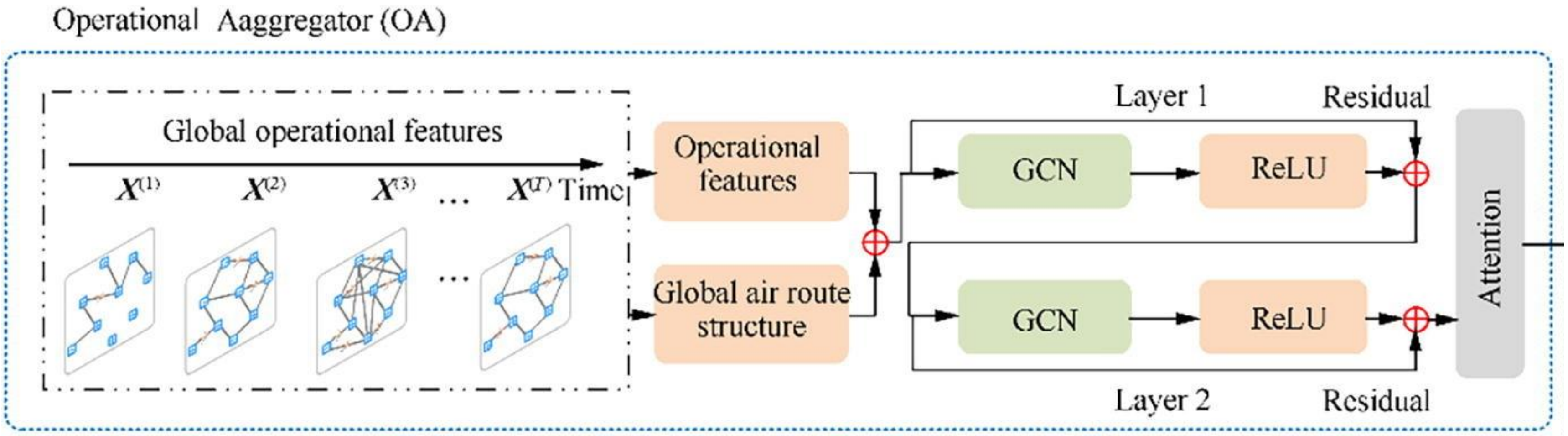
Example of Flight Information



Data Preprocessing and Graph Generation

- Subset to only include airports and flights to and from airports of a medium to large size
- Bucket flight times into T groups of 15 minutes
- Remove flights with important missing information
- Summarize flight information to find average flight delay for an airport within it’s T group
- Take the flight information’s origin and destination columns to create all edge pairings
- Use additional flight information columns and node information to create edge embeddings
- Sort and remove duplicate edges
- Loop through the T groups to build and load subsequent graphs

Methodology



Methodology

Operational Aggregator

- Input is the result of the graph generation is X, which is a T x N x M tensor
- 2 GCN layers each with a ReLU activation function
- GAT layer used to learn the weight (C_vu) of neighbor u and overall assign importance of nodes in the neighborhood

$$\boldsymbol{X} = \left\{ \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(T)} \right\} \in \mathbb{R}^{T \times N \times M}$$

$$\boldsymbol{h}_v^{(0)} = \boldsymbol{X}$$

$$\boldsymbol{h}_v^{(l)} = \mathrm{OA}\left(\left\{ \boldsymbol{h}_u^{(l-1)} \right\}_{u \in \widetilde{N}_{\mathrm{f}}(v)} \right)$$

$$= \sum_{u \in \widetilde{N}_{\mathrm{f}}(v)} C_{vu} \boldsymbol{h}_u^{(l-1)} \boldsymbol{W}^{(l)}$$

$$C_{vu} = \left(\widetilde{\deg}(v) \cdot \widetilde{\deg}(u) \right)^{-\frac{1}{2}}$$

Experiments

- 70% / 30% Train-test split
- MSE loss function
- Adam optimizer used with a learning rate of 0.001
- Train for 10 epochs

Results

	Base GCN(whole graph)	Operational GCN(whole graph)	Base GCN(subset graph)	Operational GCN(subset graph)
Train RMSE	16.039	16.032	19.956	19.927
Test RMSE	15.770	15.767	14.789	14.807

Method	MAE	RMSE
RF	9.032±0.011	12.375±0.006
GCN	8.213±0.036	10.001±0.009
GAT	8.167±0.042	9.953±0.015
GraphSAGE	8.337±0.047	10.205±0.016
Geom-GCN	8.132±0.039	9.914±0.007
BGCN	7.938±0.033	9.862±0.007
MSTAGCN	8.012±0.068	9.901±0.021
GSNet	8.025±0.046	9.912±0.012
GOGCN	7.742±0.030	9.619±0.006

Method	MAE	RMSE
GOGCN-NO	8.052	9.941
GOGCN-NG	7.851	9.785
GOGCN	7.742	9.619

Problem may be over engineered:

Linear Regression model:

RMSE: 5.64

Random Forest Regressor

RMSE: 1.479

Conclusions and Future Work

- Subsetting according to rho appears to improve performance slightly
- Findings in the paper regarding the performance degradation of the Operational GCN to that of a GNN appear to be accurate
- Interest in examining a number of features that weren’t directly considered in the work
  - Passenger volume
  - Cargo volume
  - Aircraft sizes
  - Airport tarmac complexity
  - Explicit geographic relationship feature definitions