Will Your Flight Delay?

Using KNN to predict delay severity

QIAN ZHANG[1], SIMENG ZHU[1], FENG JIANG[1], AND HE QIN[1]

[1] *CSE Department in University of California - San Diego*

Compiled December 5, 2017

Using the Bureau of Transportation Statistics data, our study aimed at understanding the factors that have impact on flight delays and tried to predict flight delays using these factors.

1. INTRODUCTION

As the world becomes more and more connected, international students, businessmen, tourists are having huge demands for long distance travel. Usually, flights are the only way to facilitate travel across the huge geological boundaries. However, the flying experience is far from satisfactory. According to Peterson et al., flight delays in the U.S. alone are causing tens of billions of economic cost annually. Besides the economic cost, the cost of the psychological anxiety of millions of passengers is incalculable. Therefore, understanding the underlying factors for flight delays and making predictions using these factors are important for aiding passengers decision.

After formulating the prediction task as a multi-class classification problem, we applied KNN, SVM and Softmax Regression models. Based on the annual data from October 2016 to September 2017, our KNN model reaches the overall accuracy of 96.4402%.

2. DATASET

For most accurate information, we've chosen Airline On-Time Performance Data from Bureau of Transportation Statistics. Their table contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as departure and arrival delays, origin and destination airports, flight numbers, scheduled and actual departure and arrival times, canceled or diverted flights, taxi-out and taxi-in times, and non-stop distance[7].

For the purpose of this assignment, our group has chosen the data from the latest year (from Oct.2016 to Sep.2017), with the total amount of 247,250 lines of records. This amount of data should be able to support a solid learning process.

After collecting the data, we ran an exploratory analysis of the data to uncover the properties of the data set. We ended up with some intriguing insights.

A. Delay distribution and airline performance

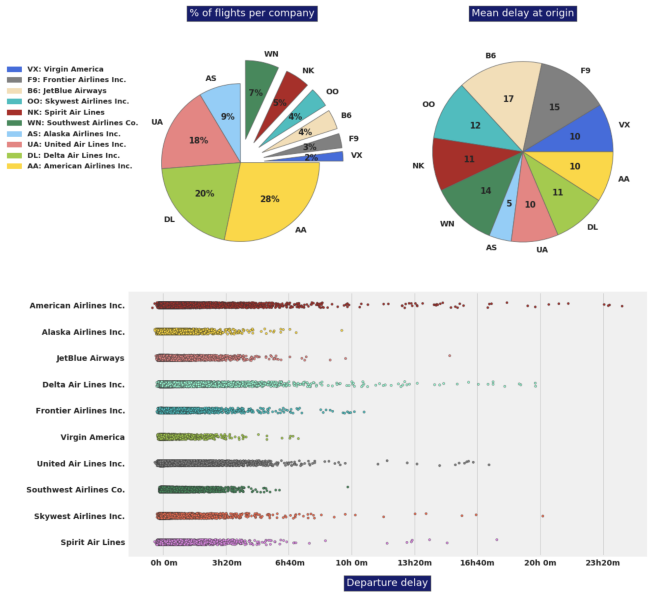

Fig. 1. Distribution of flight delay by three standards

Among all flights from Oct. 2016 to Sept. 2017, about half of them were handled by *Delta Airline* and *American Airline*, while the combination of the flights

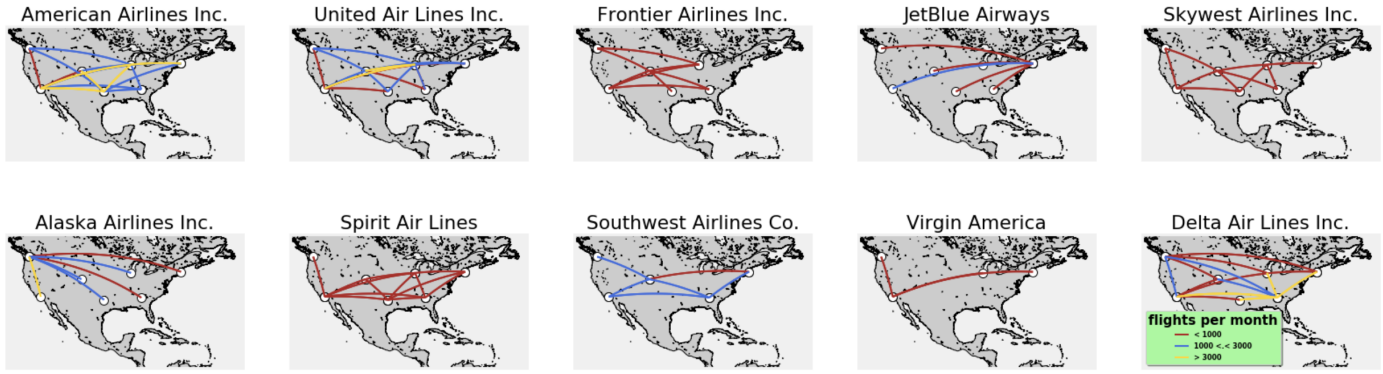

Fig. 2. Flight intensity by airlines

from 6 smaller companies (*SouthWest Airline, Spirit Airline, Skywest Airline, etc.*) consist of only 25% of all flights. It's easy to see that the distribution of flights according to airlines is very skewed.

Though the flights are skewed with respect to the operating airline, the mean delay time is much more balanced - averaging at 12.78 with a (small) standard deviation of 2.54. It's quite an interesting finding for two reasons: one, it seems like the whether your flight will delay has little to do with the airline your choose; two, it implies that on average you only need to wait less than 15 minutes for your delayed flight, which seems questionable - after all, all of us might remember how we waited for hours for the flight home the day before Christmas. One reasonable explanation could be that the one delayed flight gets etched in our memory much deeper than all its on-time counterparts.

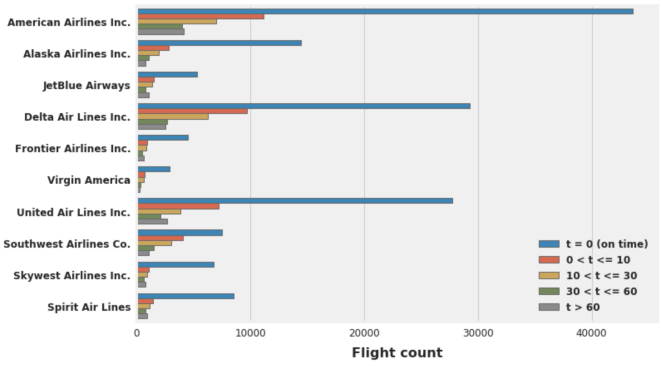

Fig. 3. Detailed look at flights by airline

Indeed, by looking at the scatter plot below the pie charts, plenty of the dots gather around the on-time zone, with dots getting sparser as the delay time lengthens. Extreme delays, like the one from American Airlines which took more than 23 hours, are extremely unlikely to happen. Yet delay can still happen to most of us, since many flights are shown to have a delay up to 3 hours.

Fig. 3 gives account of the delays subdivided into 5 categories. From the figure, we can see that extreme delays don't happen often, regardless of the airline. However, the proportion of different categories does relate to the airline. Specifically, Alaska Airlines has the lowest rate of delays over 1 hour. On the end of the spectrum, United Airlines has a high proportion of delays over 1 hour. In fact, it has more delays over 1 hour than Delta Airlines despite operating fewer flights than Delta Airlines in total.

B. Relationship between origin airport and airlines

Fig. 4. Origin airport and airlines' mutual effect on delays

To understand how origin airport and different airlines interact to affect the delays, we calculated the mean delay for each existing combination of origin airport and airline. For the most popular seven origin airports and airlines with reasonable scale, we were able to put our result into Fig. 3.

Examining Fig. 3 allows us to confirm several prior findings: First, airlines perform quite differently. Alaska Airlines is usually related to short delays, while JetBlue Airways more frequently find itself associated with long delays. Second, certain airports, for example, Boston Logan International Airport, does a better job than other fellow airports. With these two findings combined, we can see that the relationship between origin airport and airlines is highly unstable: good airline can delay long at a crowded airport, and vice versa. Thus, it inspires us to construct models that are specific to the origin airport and operating airline.

C. Effects of time on delays

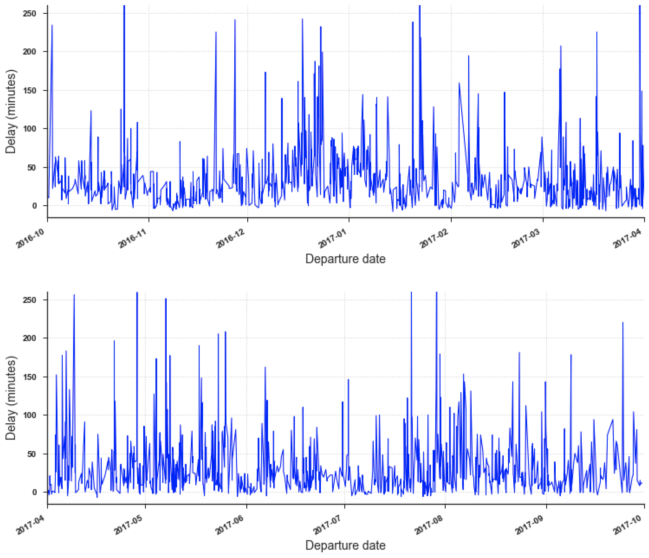

Fig. 5. Pattern of delays during the time of a year

Apart from the home airport's effect on delays, we've also turned our eyes to possible influences of the scheduled departure time has on how long a plane dawdles. On a greater scale, the date you choose for your itinerary has an effect on whether you'll have to re-watch an episode of Stranger Things at the terminal of your local airport. By inspecting Fig. 4, we notice that spikes of delays emerge at particular times of the year. Further scrutinizing the spikes seem to reveal the underlying pattern that long delays appear more often during national vacations: Christmas, Thanksgiving, Veteran's day... etc.

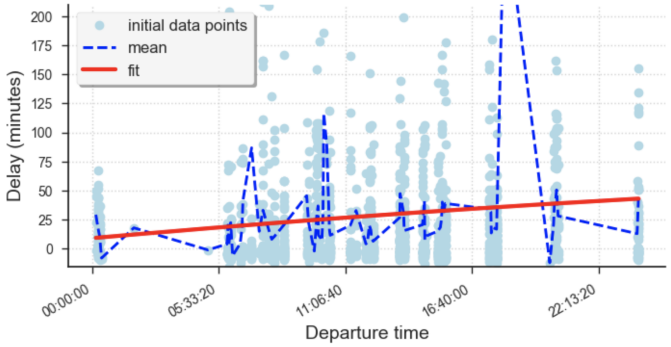

Fig. 6. Horometrical pattern of delays

On a smaller scale, the time of day your flight takes off also matters. Morning flights get the least delay, while the congestion reaches the maximum at around 6 pm. Staying up for the extra-early flight? Bad idea, as there's also a surge of delays around midnight.

D. Weather and on-time performance

Successful flight operation relies on supportive weather. We understand that it's an imperative to include weather in our model, thus we found weather data from Iowa Environmental Mesonet - an ever-growing archive of automated airport weather observations (commonly known as "ASOS sensor" or "METAR data") from around the world [6]. Among a total of 22 attributes, we chose to include Wind Speed in knots (sknt), Visibility in miles(vsby) and Present Weather Codes(presentwx), as they appear to be the most relevant.

3. PREDICTIVE TASK

A. Features selection

The purpose of our study is to predict flight delay given flight information. After the exploratory analysis of our dataset, we found that features, including the airline that the flight belongs to, the day and time when the flight takes place, the original and departure airports of the flight, and the weather condition, have a significant impact on the delay severity of a flight. Therefore, from the dataset, we chose the airline, the scheduled departure time, the month, the day of week, the origin and destination airport, and the flying distance as the features to represent each flight so as to include the predictive features from

Table 1. Features with encoding and dimension

Feature Name	Encoding	Dimension
airline	one-hot	10
scheduled_departure	one-hot	24
month	one-hot	12
day_of_month	one-hot	31
day_of_week	one-hot	7
origin_airport	one-hot	7
destination_airport	one-hot	7
distance	float	1
wind_speed	float	1
visibility_in_miles	float	1
sky_coverage	one-hot	5

the dataset and also ensure proper differentiation of each flight.

More specifically, the airline is the ten major airline companies: United, American, Frontier, JetBlue, Skywest, Alaska, Spirit, Southwest, Delta, Atlantic Southeast, and Virgin represented using the one-hot encoding. The time information is the scheduled departure time encoded using one-hot encoding for every one hour interval. For date information, we use month, day of month, day of week with one-hot encoding to compress 365 days into 49 features. Moreover, we also used the origin airport, destination airport with one-hot encoding and flight distance to encode the geological information of the flight. The weather factors were represented by the wind speed, visibility in miles. We also consider the sky coverage(cloud amount) from the Present Weather Codes, which include 5 level of Clear(CLR), Few(FEW), Scattered(SCT), Broken(BKN) and Overcast(OVC). The detailed feature summary is shown in Table 1.

B. Prediction classes

We aimed to build the model using the features described above to predict the severity of flight delay. The severity of flight delay was categorized into 6 classes, including no delay, delay(0,10] min, delay(10,30] min, delay(30,60] min, delay >60 min, and canceled. The class interval was so chosen both to ensure a relatively balanced classes instances and also to agree with people's psychological perception of delay severity.

C. Models and evaluation

Many supervised learning models are suitable for this prediction task. We wanted to compare the performance of several models and chose K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Softmax Regression, whose detailed discussion will be in the next section. For the baselines, we used the naive "all-on-time" predictor, that is to predict all flights on time. Since we are solving a multi-class prediction task, we used the 1)overall accuracy and 2)confusion matrix to tune and compare the performance of different models. The overall accuracy is calculated as the number of overall correctively labeled instances divided by the size of the test set.

4. MODEL

A. Models selection

Since predicting delay severity is a typical classification problem, we chose the above three models. The following contents discussed the assumptions we made to the dataset in order to select the models, together with brief descriptions of these models.

1. K-Nearest Neighbor (KNN)

   KNN is the target model in our project. It's natural to think that flights with similar weather conditions and flight information should let to similar delay decision by the flight crews. Thus, it is legitimate to predict new flights based on the known similar flights, which can be achieved by KNN.

   KNN comes with two advantages. First, it is a nonlinear classifier, which is good for predictions in complex feature space distribution. Second, it's training process is relatively fast. However, it is hard to explain the model because KNN does not provide contribution explanation to each input factor. Also, the prediction process is relatively slow.

2. Support Vector Machine (SVM)

   SVM method is another model worth trying to tackle the classification problem. In SVM, we partition in the high-dimensional feature space to classify instances based on the distribution of known training data. Analogous to the previous analysis in KNN part, we assumed that

instances that clustered in the feature space tend to hold similar delay condition.

In terms of advantages and disadvantages, SVM is similar to KNN. Besides, by adopting nonlinear kernels, SVM can generate more complicated classification boundary.

3. Softmax Regression

Softmax Regression can be viewed as the combination of two parts: 1) linear predictors, and 2)a Softmax classifier. Encoding the delay label in one-hot format, the linear predictor will generate the evidence for each class, as shown in the following equation.

$$evidence_i = \sum_j \theta_{i,j} x_j + b_i \quad \textbf{(1)}$$

Then, all evidences will be plugged into the Softmax function, which is the following:

$$softmax(evidence)_i = \frac{e^{evidence_i}}{\sum_j e^{evidence_i}} \quad \textbf{(2)}$$

The Softmax function will serve as a filter to amplify the evidences. It will also normalize the result, and make it the probabilities of our prediction over each class.

Actually, the Softmax regression transform the classification problem into linear regression over each class, and summarize the results with the Softmax amplifier. And it is reasonable to use linear regression on our features, with the assumption that all features are independently contributed to the final prediction.

Compared to KNN and SVM, Softmax Regression is better at explaining the contributions from each feature factor. Also because it is relatively simple, the train process is faster. However, as shown in the final comparison table, it was inferior to the other two in accuracy. One of the explanation is that it is deficient in predicting mixing boundaries and also is biased by the majority class.

B. Tuning and optimizations

1. K-Nearest Neighbor (KNN)

There are totally 247,250 instances of data in our flight delay dataset. When training the KNN model, we divided our dataset such that the first 70% were the training set and the rest were testing set. We tried different neighbors numbers from 3 to 25 for the KNN on the same training data and plot their overall accuracy as in Fig. 7. We also tried the distance weighted KNN since more similar flights should have greater contributions on the classification instead of equal voting. Indeed, we can see from the figure that the distance weighted KNN has better accuracy than the unweighted KNN. For the unweighted KNN, we can see that the accuracy increases from 3 to 15, peaks at 15 with overall accuracy of 0.6756, plateaus and slightly decreases afterward. While for the distance weighted KNN, the accuracy increases from 3 to 21 and peaks at 21 with accuracy of 0.8744 Therefore, the distance weighted KNN has larger range of increases when we vary the number of neighbors, and has a better overall performance.

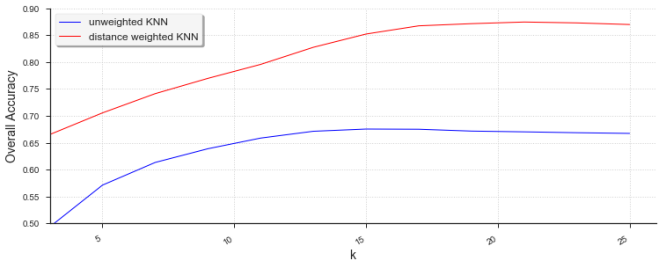

Fig. 7. Tuning the number of neighbors (k) of KNN

2. Softmax Regression and Support Vector Machine (SVM)

Softmax regression and SVM was also trained using the same partition of training set and testing set as KNN. We tuned the penalty terms of SVM and Softmax in the range of 0.1 to 30, and saw a similar pattern of increasing and decreasing accuracy, though the variation in accuracy was smaller than we expect and was within the range of 0.01 to 0.05. We also tried a linear kernel for SVM, but its performance was much worse than the Gaussian (RBF) kernel with an overall accuracy of 0.6347, indicating that our task is not linearly separable in high dimensional space.

3. Tackling Scalability

The original dataset contains 318 airports, which will greatly increase our feature dimension. Therefore, we downsized to 7 airports so that to include the busiest airports and also ensure a

balanced geological distribution of the airports, which can help us to avoid the dimension disaster due to large one-hot encoding format. On one hand, one-hot encoding ensures the equality over different attribute values. On the other hand, because our features are heavily exploiting one-hot encoding, our training matrix becomes very sparse. This naturally lead us to consider the PCA method to first reduce the dimension and then train on it.

We tried the PCA, and found that we can save our current 106-dimension feature by 21, and still get a very close accuracy on Softmax Regression model. Therefore, we can expect that PCA will be a crucial pre-processing step in the future, if we need to expand attributes like airports and departure time period.

5. LITERATURE

A. About datasets

Our data is the existing dataset collected by the Bureau of Transportation Statistics since 1995. This data is easily accessible in the database directory of Bureau of Transportation Statistics, with the name of "Airline On-Time Performance Data". Most common use of this dataset is to rank airlines on their on-time performances. Many websites, including that of Bureau of Transportation Statistics itself, had given out such rankings.

Weather data came from Iowa Environmental Mesonet hosted by Iowa State University. Its website allows subtraction of data using filters such as station location, data unit, time range of dataset..., etc.

More specifically, we downloaded both the "On-Time Performance Data" and "ASOS Weather Data"(7 airports independently) from October 2016 to September 2017. Then we carefully labeled the flight data with weather features, by matching the flight departure time to the closest half-hour weather data(using the airport local time zone). The "ASOS Weather Data" collects weather data with a roughly 5-minute period. However, there are some missing data due to sensor failure. Therefore, we mannually rounded the weather to half-hour period. And for the very rare situations that the whole half-hour weather data is missing, we discard the corresponding flight data.

Table 2. Model comparison in overall accuracy

Models	Overall Accuracy
Naive all-on-time predictor	0.6586
KNN (distance weighted)	0.8744
SVM (RBF kernel)	0.8616
Softmax Regression	0.8462

B. Other similar datasets

There are very few similar datasets for flights delay. Most of them are composed of delay messages only, this kind of datasets only care about the result of delay rather than the reasons, and are mainly used for analyzing the economic impact of flight delays using Statistical Analysis and Probabilistic Models.

For example, calculate the economic impact of flight delays on airlines and passengers, the cost of lost demand, and the collective impact of these costs on the U.S. economy with 2007 data [1]. How delay impacts an airline's schedule and thus provide an input to airline and FAA decisions [2]. Developed a probabilistic model based on expectation-maximization combined with genetic algorithms to predict the distribution of departure delay at Denver International Airport [3].

C. State-of-the-art Methods

Besides the research area of Statistical Analysis mentioned before, operational research and data management also are the most applied in the past. In terms of models selection, neural networks, SVM, fuzzy logic, and random forests are commonly used for prediction. Among them, network method is currently employed to study this type of data. For example, in 2016, Wu built a Bayesian network to model delay propagation [4]. Baspinar built a network-epidemic process using historical flight-track data of Europe to create a novel delay propagation model [5].

6. RESULTS AND CONCLUSIONS

Table 2 shows the performances of different models compared with the baseline. The naive all-on-time predictor baseline always predicted no delay which is the majority class in the dataset and had accuracy of 0.6586. As we can see, all the models outperform the baselines that we chose and on average

achieved an accuracy of 85%. In our experiments, the three models that we chose had similar overall accuracy, while the fine-tuned distance weighted KNN out-performed the other two models with 0.8744 accuracy.

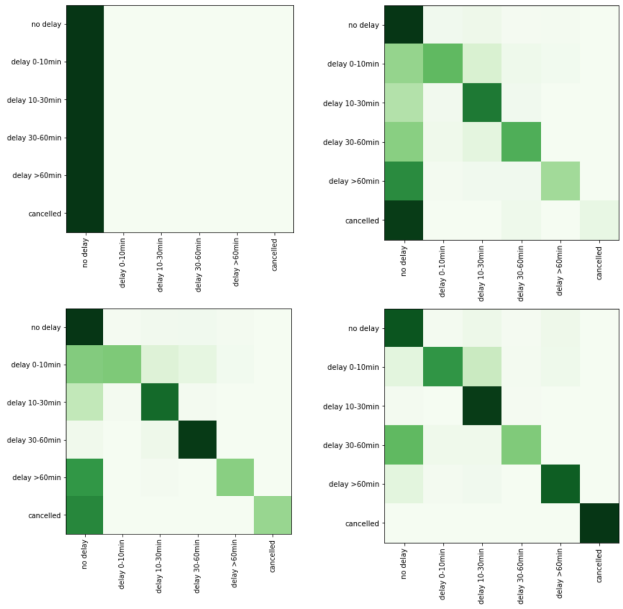

Fig. 8. Comparison of different models' confusion matrix

We can use the confusion matrix in Fig. 9 to make further comparison. In each graph, the vertical axis labels the actual class certain instance belongs to, while the horizontal axis is the predicted class by each model. So each square is a combination of actual class and predicted class, and the darker the square the more instances belonging to its combination. Therefore, the darker the diagonal square, the better performance the predictor has on all the classes. For the four sub-graphs in Fig. 9, the top left is the baseline which just predicts no delay for all instances. The top right is Softmax Regression, where many instances are mistakenly labeled as no delay probably due to the majority class bias. SVM with Gaussian (RBF) kernel can draw non-linear decision boundary for classification, and it has a relatively uniform performance on all the classes in the bottom left, but it also has difficulty predicting delay >60 and canceled flight. Since Softmax Regression pushes the classification boundary for all the training examples as far as possible, while our dataset classes are imbalanced and the majority class of our data set was no delay (0.6586), so Softmax Regression might be biased by the majority class and don't have a good performance. And we hypothesize that the performance of SVM should be better since it just minimizes the classification mistake so it's less influenced by the majority class. And the results testify our hypothesis, the SVM with RBF kernel has over all accuracy of 0.8616, which out-compete Softmax Regression's overall accuracy 0.8462. Lastly, In our experiment, distance weighted KNN has the best performance as shown in the bottom right. Flights with similar flight information under similar weather condition might result in the same delay decision by the flight crews. Since KNN lets the k most similar training case to vote on the classification of the test case, it might be the most suitable among the models that we chose to compare. In conclusion, our results indicate that the flight delay prediction is overall a tractable task if we combine the weather condition and the flight information and KNN provides the best prediction among SVM and Softmax regression when tackling the flight delay prediction task.

REFERENCES

1. Ball, Michael, et al. *Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States*. 2010.
2. Beatty, Roger, et al *Preliminary evaluation of flight delay propagation through an airline schedule.. Air Traffic Control Quarterly* 7.4 (1999): 259-270.
3. Tu, Yufeng, Michael O. Ball, and Wolfgang S. Jank *Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern.. Journal of the American Statistical Association* 103.481 (2008): 112-125.
4. WU, Wei-wei, Ting-ting MENG, and Hao-yu ZHANG *Flight Plan Optimization Based on Airport Delay Prediction.. Journal of Transportation Systems Engineering and Information Technology* 6 (2016): 029.
5. Baspinar, B., and E. Koyuncu *A Data-Driven Air Transportation Delay Propagation Model Using Epidemic Process Models.. International Journal of Aerospace Engineering* 2016 (2016).
6. herzmann, D. Iowa Environmental Mesonet. IEM :: Download ASOSAWOSMETAR Data. Retrieved from https //mesonet.agron.iastate.edu/request/download.phtml?network=CO_ASOS/. 2017 (2017).
7. Bureau of Transportation Statistics. OST_R | BTS | TranstatsṘetrieved from https //www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time 2017 (2017).