Data Analysis for US Airport Network, Part I

February 19, 2020

1 Basic Dataset Description and Problem Formulation

We utilize the flight information in 2008 for our two models. The entire data is encapsulated into Graphflow objects to describe graph-valued time series. In our visualization of a graphflow object(Figure 1), the locations of nodes(airports) in the network are based on their geographical locations. Each edge represents a time series, and the time series of the edges can be used to produce the time series information of nodes. The time series produced by edges may result on the changes of time series information of nodes.

1.1 Problem Setup

1.1.1 Definition of Delay Ratio

Asumme there are $n \in \mathbb{Z}^+$ airports from the dataset annually. Let $N_i \in \mathbb{Z}^+$ where $i = \{0, 1, ..., n\}$ as the total number of the flights in airport i. Let $N_i^*(t) \in \{0, 1, ..., N\}$ be the number of delayed flights in airport i at some time t.

We define delayed ratio as

$$p_i = \frac{N_i^*(t)}{N_i} \tag{1}$$

the proportion of the number delayed flights in terms of the total total number of flights of airport i(delay rate).

1.1.2 Prediction of Delay Ratio

Our goal is to predict delay ratio using past information from the dataset.

1.1.3 Test Set Description and Evaluation Method

For the test set using for evaluating our model, 300 time points are randomly chosen from the original dataset in 2008 without cheating. Each time point includes the information of flights of all the airports in the Graphflow object in this specific time.

MSE(Mean Squared Error) is used as the evaluation method to measure the errors produced by our models. The MSE formula is defined as following:

$$MSE = \frac{\sum_i \sum_t |\hat{p}_i(t) - p_i(t)|^2}{\sum_i \sum_t 1} \tag{2}$$

For clarity, the sample evaluation demonstrations are shown in Figure 2 and Figure 3. We choose LAX, LAS, OGG, and HNL to compare the real-world data and predicted data from 2007-02-01 01:00:00 to 2007-02-15 01:00:00, and the MSE value is 0.27906976744186046 in this case.

2 Naive Model Introduction

2.1 Formulation

We have two distinct concepts. For stochastic process $A_{ij}(t)$ $(i,j) \in E, t \in [t_0, t_1]$, t is continuous. E is the set of the directed edge of G. For time series $B_{ij}(t)$ $(i,j) \in E, t \in t_0 + \mathbb{N}dt$, t is discrete.

Let $d_{ij} = (D_{ij}(t), O_{ij}(t))$ be the vector valued time series, where $D_{ij}(t)$ is the number of departure-delayed flights from airport i to airport j between $[t, t+dt)$ and $O_{ij}(t)$ stands for the number of departure-on-time flights from airport i to airport j between $[t, t+dt)$.

$$d_i(t) = (D_i(t), O_i(t)) = (\sum_j D_{ij}(t), \sum_j O_{ij}(t))$$

Similarly, let $\alpha_{ij} = (D^{\alpha}_{ij}(t), O^{\alpha}_{ij}(t))$ be the vector valued time series, where $D^{\alpha}_{ij}(t)$ is the number of arrival-delayed flights from airport i to airport j between $[t, t+dt)$ and $O^{\alpha}_{ij}(t)$ stands for the number of arrival-on-time flights from airport i to airport j between $[t, t+dt)$.

$$d_i(t) = (D_i(t), O_i(t)) = (\sum_j D_{ij}(t), \sum_j O_{ij}(t))$$

So our problem becomes the following estimation:

$$p_i(t) = f(d_i(t) + \sum_k \alpha_{kj}(t)) \text{ with } f(x,y) = \frac{x}{x+y}.$$

We get estimator $\hat{\alpha}_{ij} = R_{E_{ij}}(d_{ij}(t)M_{ij})$. For each edge we associate a Markov transition matrix M_{ij}: Table 1. Notice that d_{ij} is the value from the original data, so the values of d_{ij} are precise, while $\hat{a}_{ij}$ can only be predicted using models.

And R_T is a right shift operator with $R_T(f(t)) = f(t+T)$, so to estimate $\hat{a}_{ij}$, we need to estimate E_{ij}, α_{ij}, β_{ij} as well. However, we remove seasonality of $\alpha_i(t)$ for simplicity by denoting $\bar{\alpha}_{ij} = \frac{\sum_{i=1}^{24} \alpha_{ij}(t+idt)}{24}$, where $t \in t_0 + \mathbb{N}(24dt)$.

The following data convinces us that this model might be applicable.

Total	ArrDelay	ArrOnTime
DepDelay	a_{ij}	$1 - a_{ij}$
DeoOntime	β_{ij}	$1 - \beta_{ij}$

Table 1: table:Markov transition matrix

From Figure 4(distribution of the elapsed time) and Figure 5(qq-plot of the distribution of the elapsed time and normal distribution), we notice that the distribution of the elapsed time is relatively similar to normal distribution with mean 323.2697 and variance 22.8742.

For E_{ij}, we fix two airports to observe the property of E_{ij} since if two airports are fixed, the time spend on air between two airports behaves almost like a constant value. We choose i to be LAX, j to be JFK in the following case.

From Figure 6, we can find that α_{ij} and β_{ij} share almost the same increasing and decreasing patterns while the upper bound of α_{ij}(if ignoring unusual peak values) is around 0.5, and the upper bound of β_{ij} is around 1.

2.2 Data Fitting: ARMA

For α_{ij}, we use ARMA(2,1) method to fit our data. From the PACF plot of α_{ij}(Figure 7), we can conclude that α_{ij} is slight correlated with time with the correlation value around 0.2. From the Q-Q plot (Figure 8) we can conclude that the the distribution of α_{ij} almost fit as normal distribution while from correlogram we can conclude that each value of α_{ij} is independent.

Similarly, for β_{ij}, we use ARMA(0,1) to fit our data. The PACF plot of β_{ij}(Figure 10), the Q-Q plot and correlogram(Figure 11 show the similar results as we saw from α_{ij}. These statistical plots indicate that the fitting works well.

However, from Figure 9 and Figure 12, we could see that the trending of the predicted data fit the real world data mostly.

2.3 Evaluation

Still needed to update.

3 Epidemic Model and its Stable Solutions

3.1 Formulation Review

Let δ_i be the recovery rate(the rate that an airport back to normal operation)and b_{ij} be the infection rate. If airport i is directly affected by airport j, then $b_{ij} \geq 0$; otherwise, $b_{ij} = 0$. As professor Heather's suggestion, b_{ij} can be considered as an adjacency matrix.

The dynamics of the epidemic process can be written as a differential equation:

$$\dot{p_i}(t) = -\delta_i p_i(t) + \sum_{j=1}^{n} b_{ij} p_j (1 - p_i)$$

3.2 Solution Analysis

By observing Figure 13, we notice that the solution is not linear, but under the initial conditions and experimental data, the solution would approach stable if we take time large enough. Hence, the solution can be considered as

$$\delta_i p_i = \sum_j b_{ij} p_j (1 - p_i). \tag{3}$$

This solution obviously cannot consistently characterize the patterns of airports in one year since the real-world values of delay ratio p_i are not stable. In summary, the real-world data of delay ratio p_i has the following characteristics: (i) Periodicity occurs in some specific time and seasons. (ii)Blowing up at the same time in space.

Nevertheless, $p = 0$ is also a solution for the model, though it might not be stable. Thus, some questions are raised: how many stable solutions for the epidemic model? Are they stable or not?

If we perform change of variable,

$$\begin{aligned} x_i &= p_i \\ A_{ij} &= b_{ij} - d_{ij}\delta_{ij} \\ B_{ijk} &= -b_{ij}\delta_{ik} \end{aligned} \tag{4}$$

written in summation we have

$$\dot{x}_i = A_{ij} x_j + B_{ijk} x_j x_k \tag{5}$$

In one dimension, the special case of this function as following:

$$\dot{x}_i = ax + bx^2. \tag{6}$$

This equation can be solved by $x = 0$ or $x = -\frac{a}{b}$; one solution is stable, and the other is unstable. If the flow around the origin approaches to 0, then $x = 0$ is stable.

Similarly, to determine if 0 is a stable solution, we should check the eigenvalues of $A_{ij} = b_{ij} - d_i\delta_{ij}$. If one of the eigenvalues are positive, then 0 is unstable; if all of them are negative, then 0 is stable.

In summary, if the numbers of negative diagonal elements of $A_{ij} = b_{ij} - d_i\delta_{ij}$ are more than non-negative diagonal elements, then the airport has a faster pace to handle delayed flights. If b_{ij} is large, then the flights from other airports would cause congestion; if such congestion keep enlarging, then solution $x = 0$ is unstable.

Thus, for other solutions, we can utilize the same method to determine stability. Assume $x_i = x_i^* + \epsilon_i$. Assume x_i^* is a solution, we expand x_i around ϵ_i, then we have

$$\dot{\epsilon_i} = A_{ij}(x_j^* + \epsilon_j) + B_{ijk}(x_j^* + \epsilon_j)(x_k^* + \epsilon_k).$$

Since around 0, the non-linear term would goes out very fast, we can focuse on the liner terms. Thus, around the new solution x_i^*, we have

$$A_{ij}\epsilon_j + B_{ijk}x_j^*\epsilon_k + B_{ijk}x_k^*\epsilon_j = \bar{A}_{ij}\epsilon_j.$$

We can analyze the $\bar{A}_{ij}$ to determine the stability.

However, the predictability of δ_{dep} depends on the the relation between $\delta_{dep}, \delta_{arr}$ and t.

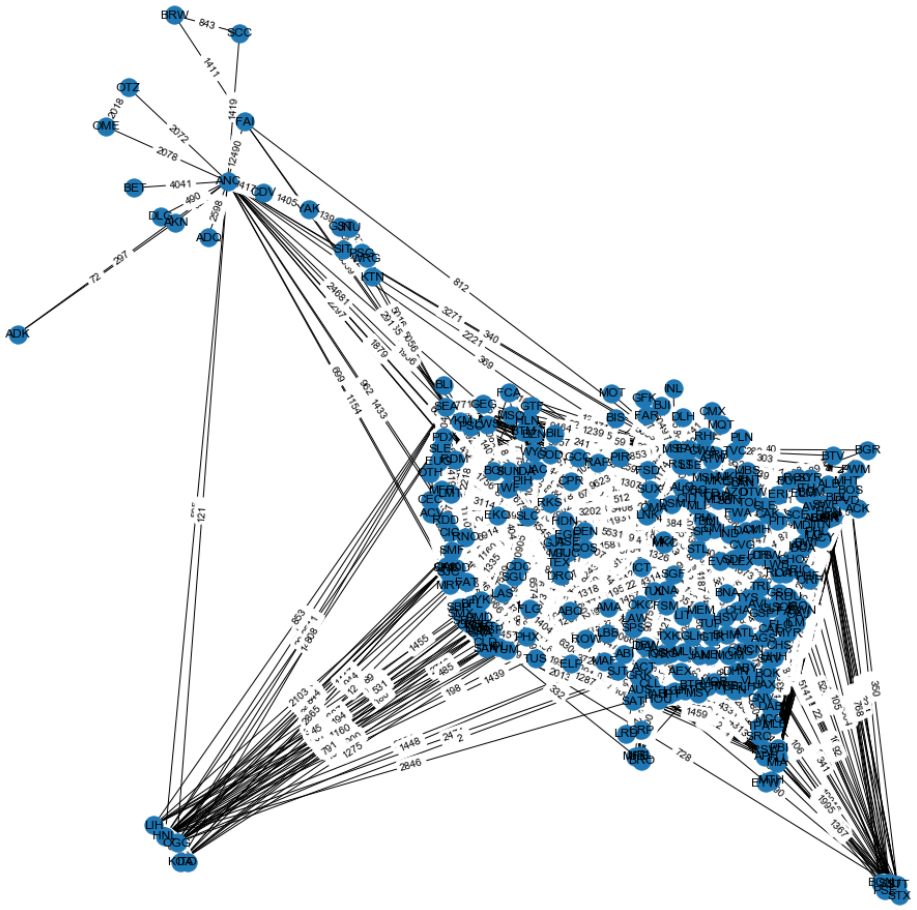

Figure 1: Network Visualization Example

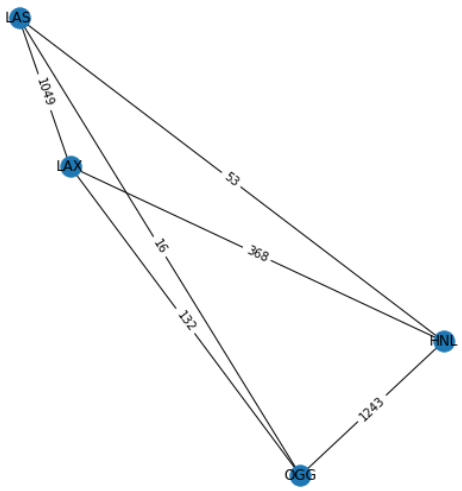

Figure 2: The SubGraph of LAX, LAS, OGG, HNL from Graphflow object

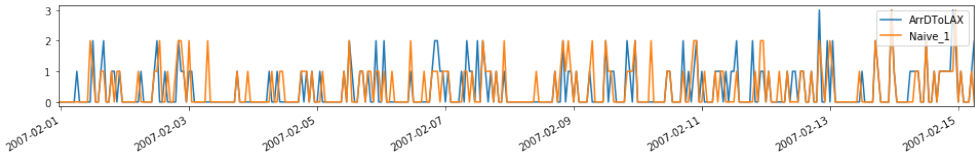

Figure 3: Real-world Data of Arrival Delayed Flights Vs. Predicted Data.

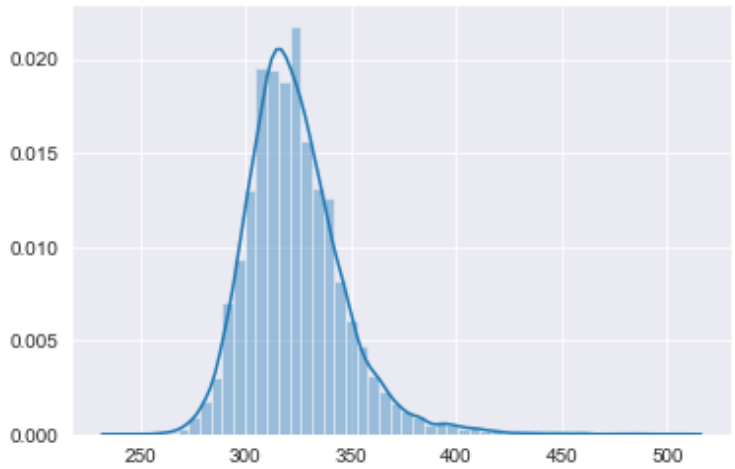

Figure 4: Distribution of the time speed on airplanes(elapsed time). x-axis:The time spend on airplanes(air time+taxi in time+taxi out time) vs. y-axis: density values.

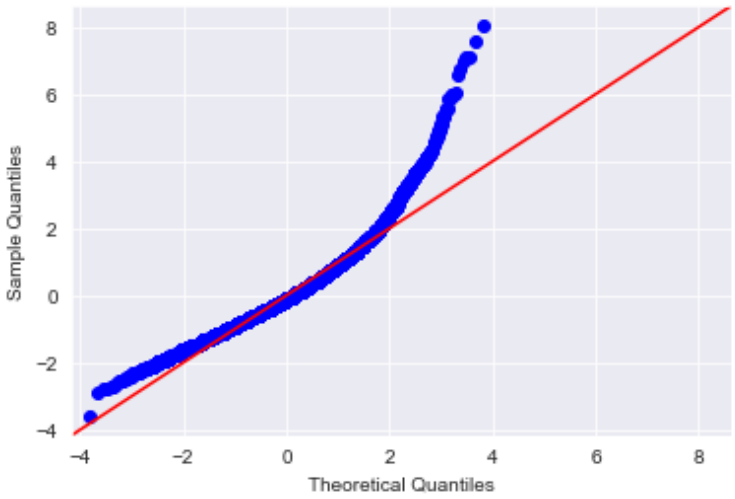

Figure 5: Quantile-Quantile Plot: comparing between distribution of the elapsed time and normal distribution.

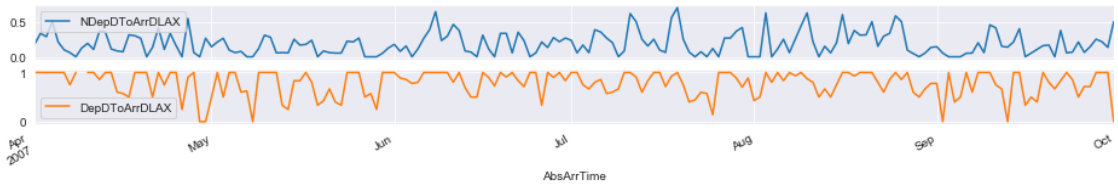

Figure 6: The time series plot of α_{ij}(the above) and β_{ij}(the below) from Apr 2007 to Oct 2007

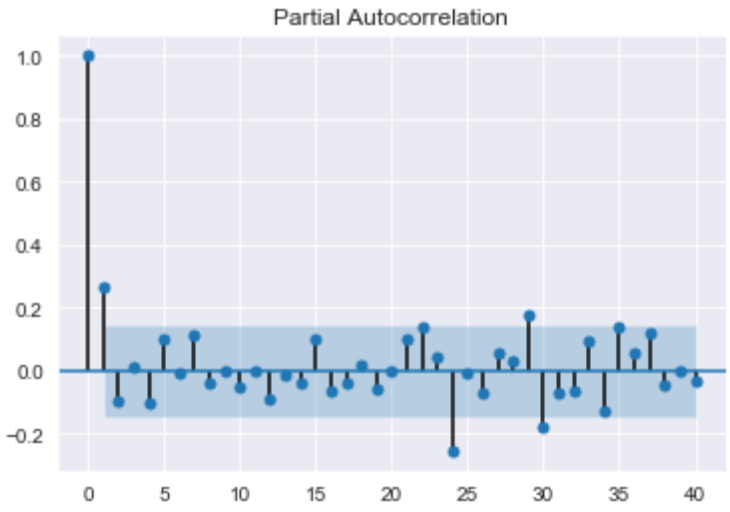

Figure 7: PACF of α_{ij}

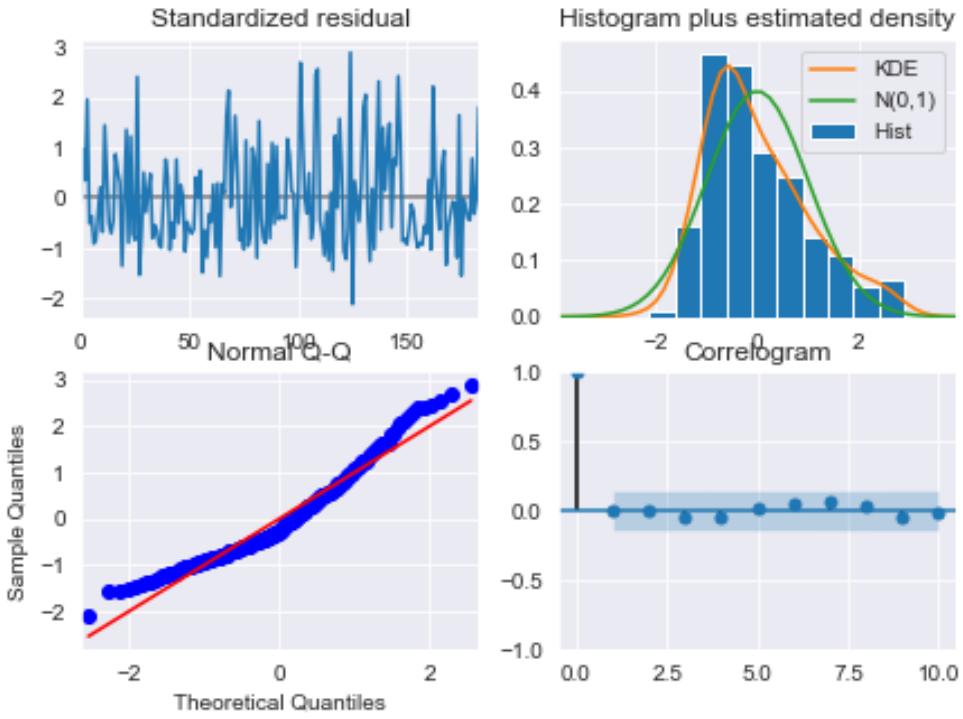

Figure 8: Standardized residual, Histogram, Q-Q plot between α_{ij} distribution and normal distribution, correlogram of α_{ij}

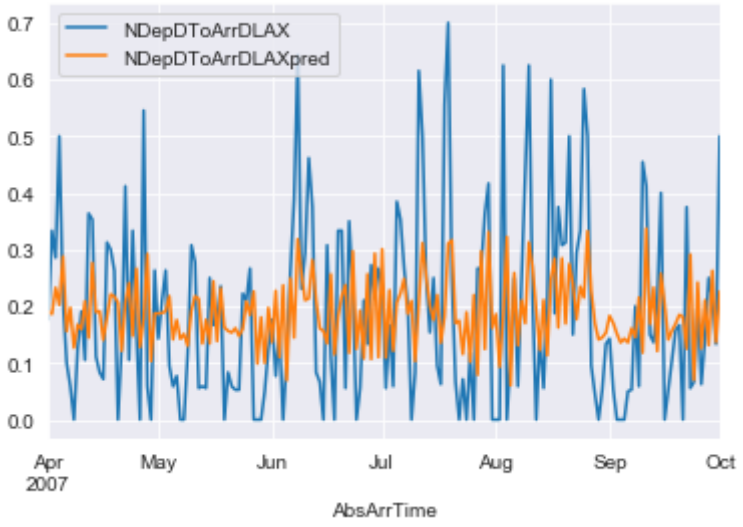

Figure 9: The real α_{ij} and predicted data

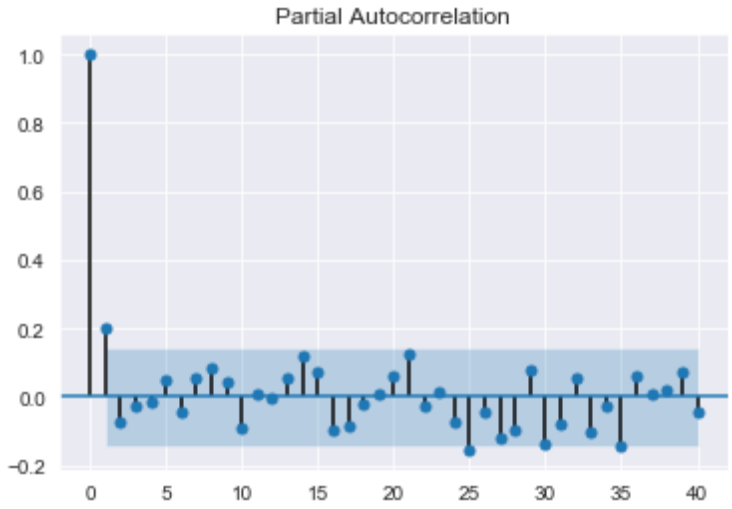

Figure 10: PACF of β_{ij}

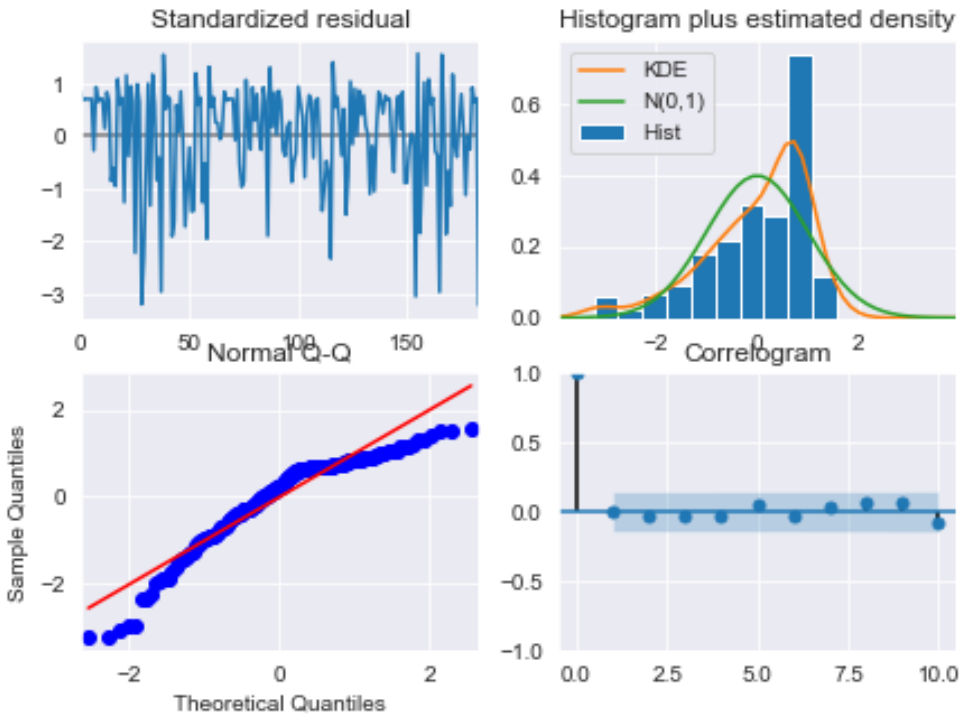

Figure 11: Standardized residual, Histogram, Q-Q plot between α_{ij} distribution and normal distribution, correlogram of β_{ij}

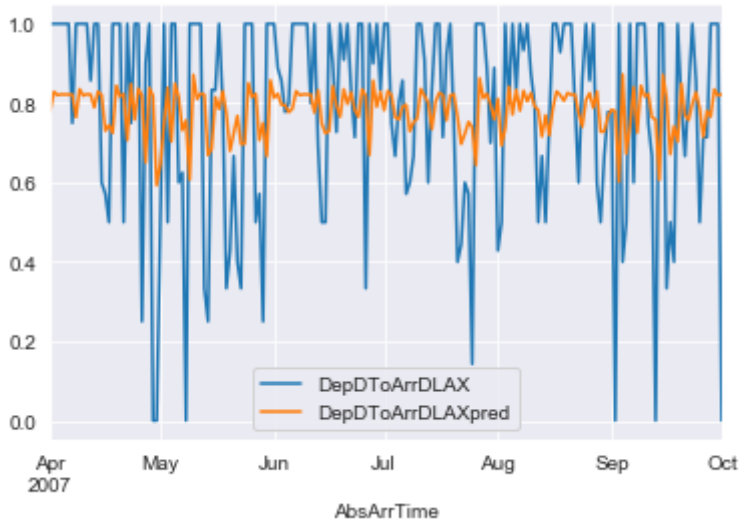

Figure 12: The real β_{ij} and predicted data

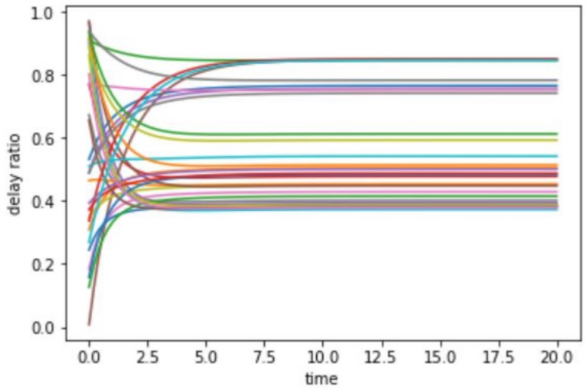

Figure 13: Delay Ratio(solutions of the epidemic model) vs.Time