Optimizing Bus Schedules to Minimize Waiting Time

Steven Kornfeld, Wei Ma, Andrew Resnikoff
21-393 - Operations Research II

Abstract

A common complaint heard on the Carnegie Mellon University campus is that the Pittsburgh buses do not run efficiently. This analysis aimed to determine the optimal scheduling of buses for the Pittsburgh bus system. To do so, data from the Port Authority of Allegheny County was used to build and constrain a model.

The model indicated that, given our assumptions, the current bus scheduling by the Port Authority is non optimal. This suggests a need for further analysis, using a more extensive model to test whether or not the simplified model in this report gives an accurate result. If these same conclusions are reflected in the complex model, it would mean that the Pittsburgh buses are potentially wasting people's time and money unnecessarily.

Introduction

In Pittsburgh, like most cities, public transportation is an important aspect of life for residents. The bus system serves people who commute to work, students from the many universities and really anyone who needs to get from point A to point B cheaply and conveniently. While the public buses do a great service to people of Pittsburgh, there are many cases where the buses could run more efficiently. Many times the buses are not running on schedule and people can end up waiting a very long time for a bus.

Through our analysis, we aim to minimize the total waiting time for bus passengers. We have collected 65,973 observations of bus stops from the Port Authority. Each observation has the line, station, time arrived at each station and the number of people who get on the bus. From this data, we are able to determine the inter-person arrival time at each station. We also found the average speed of buses and the budgets for the Port Authority. Using this data, we will build a model that can determine the optimal number of buses to run during different time periods during the day for the 61 bus line (and from that we can get the inter-bus arrival time).

Many bus riders use the bus to commute to work. As a result, we are expecting the model to show a greater need for buses during rush hour times. Looking at the current bus schedules for the 61 line, we see that, in general, the buses run most frequently during these times. This gives us further reason to believe that the optimal solution produced will also run buses more frequently during these times.

Assumptions

In order to have a solvable model, we had to make some assumptions about the system that we are modeling. Many of these assumptions are, to the best of our knowledge, realistic approximations that do not take away from the relevance of our model. For example, we assume that at each stop, the same number of people are waiting to go in each direction, which is supported by the observation that how crowded buses are seems to depend on the time and location, not on which direction the bus is going. We also assume that 2% of all resources (money from the budget and hours of labor from the operators) can be allocated to our line, which is supported by the fact that we are considering four out of approximately 100 lines but are only considering half of the line.

Other realistic assumptions include that each year has 365 days, that the overall recorded average speed of buses applies to our line, that the data on number of people at each stop is accurate, that the cost of driving to and from the garage is insignificant, and that buses arrive at even intervals.

On the other hand, some of our assumptions simplify the system but were necessary for our model to be workable. For example, we assume that each bus has infinite capacity; that is, our model does not take into account the situation where a bus arrives but not all people waiting for that line can board. We also assume that all people's waiting times are equally important, whereas in

reality, the Allegheny County Port Authority may be concerned to varying degrees about providing service to different groups of people. For example, they may consider the waiting times of people commuting to work in the morning to be of high importance but not be concerned about the waiting times of Carnegie Mellon and University of Pittsburgh affiliates, who have no choice but to pay for bus passes each semester regardless of how often they actually use the bus. We have also made assumptions about the arrival times of the passengers, as described above.

Finally, we have made some assumptions that may be slightly inaccurate because of our limited access to data but which could be easily corrected given the correct data. For example, we assume that expenses may not exceed total revenue and operating assistance. In reality, this is not necessarily the case; in 2012, the Port Authority received a total of $294,143,986 but spent a total of $327,826,692. We have elected to use $294,143,986 as the amount that can be spent, but were a different number decided upon as being an acceptable amount to spend, our model could easily be rerun by simply changing that one number.

Similarly, we are assuming that the budget data from 2012 as reported by the Port Authority (http://www.portauthority.org/paac/portals/capital/budgetbooks/FY14BudgetBook.pdf) still applies, that the total number of employees listed includes only police officers and bus operators, that all of the budget is available for wages and gas, that money spent on salaries for the Port Authority Police is insignificant, that employees work 3.5 hours a day on average, that gas prices are the average for the last quarter of 2013 as reported by CONSOL (http://www.bloomberg.com/article/2014-01-31/a5QfKcg1w4ak.html), and that salaries for bus operators are the national average for May 2013 as reported by the Bureau of Labor Statistics (http://www.bls.gov/oes/current/oes533021.htm). Were any of these assumptions found to be inaccurate, they could easily be rectified by changing the corresponding number in the model.

Model

From our data (http://www.portauthority.org/paac/portals/capital/budgetbooks/FY14BudgetBook.pdf), we have the following variables:

- Time intervals of T minutes (chosen to be meaningful based on periods of relatively consistent passenger arrivals)
- Inter-person arrival time $F_{i,k}$ minutes during interval i at stop k
- N buses available
- Gas budget G dollars per day
- x miles traveled during one time interval
- Fuel efficiency y miles per gallon
- Price of gas z dollars per gallon
- H hours of labor available each day

- Average speed v along line of length L

We then want to solve for our decision variables:

- Operate N_i buses on our line during time interval i
- Inter-bus arrival time B_i during time interval i

so that the total waiting time of all passengers at all stops during all time intervals is minimized.

Model 1

To solve the problem, we used two different models, differing by how we assumed that passengers would arrive at the stops. In model 1, we assume that passenger arrival is deterministic and occurs at even intervals. For example, if three people are going to arrive in a given hour, then we assume that the first person will arrive after 20 minutes, the second after 40 minutes, and the last after 60 minutes. Therefore, to calculate the total waiting time, we simply add each person's arrival time subtracted from the arrival time of the corresponding bus. Since buses arrive at even intervals, at each stop and each time interval, the total waiting time for each bus will be the same, so we can simply find the waiting time for the first bus and then multiply by $\frac{T}{B_i}$, the number of buses that will arrive during interval i. We then sum over all stops and all time intervals. Therefore, if we are considering S stops and days where buses operate for 1080 minutes, then the final expression for total waiting time in model 1 is

$$\sum_{k=1}^{S} \sum_{i=1}^{\frac{1080}{T}} \frac{T}{B_i} \sum_{j=1}^{\frac{B_i}{F_{i,k}}} (B_i - F_{i,k}j)$$

We minimize this obective function subject to the following constraints:

- Number of buses running at any given time: For all i,

$$N_i \leq N$$

- Gas budget:

$$\frac{xz}{y} \sum_{i=1}^{\frac{1080}{T}} N_i \leq G$$

- Hours of labor available:

$$\sum_{i=1}^{\frac{1080}{T}} \frac{N_i T}{60} \leq H$$

- For all i,

$$N_i = \frac{60L}{vB_i}$$

- All variables nonnegative
- All $\frac{T}{B_i}, N_i \in \mathbb{Z}$

Model 2

In model 2, we assume that rather than being deterministic, the arrival of passengers is a Poisson process with exponential inter-arrival times. As derived below, the objective function to minimize, subject to the same constraints, is now

$$\sum_{k=1}^{S} \sum_{i=1}^{\frac{1080}{T}} \frac{T}{B_i} \left(\frac{B_i^2}{2F_{i,k}} - B_i \right)$$

Data

Our data are the number of people getting on the bus at each station, all routes and departure times are included for each station. From these data, we pick up routes 61A, 61B, 61C and 61D and use the stations where these 4 routes shares. By summing up these data, we can assume these four routes are similar in selected stations.

So the final number of data we use are in total 65973 rows, and for each row we know the departure time of the bus, arrival time at each station and number of people getting on the bus. Then we visualize the data in Figure 1.

Each sub-figure indicates one single bus line, different colors indicate different bus stations, x-axis is time and y-axis is number of people getting on the bus. As can be seen, more people get on bus at "blue" station, and trends for all bus lines look the same. Besides, data are enough for us to estimate λ and other values we need.

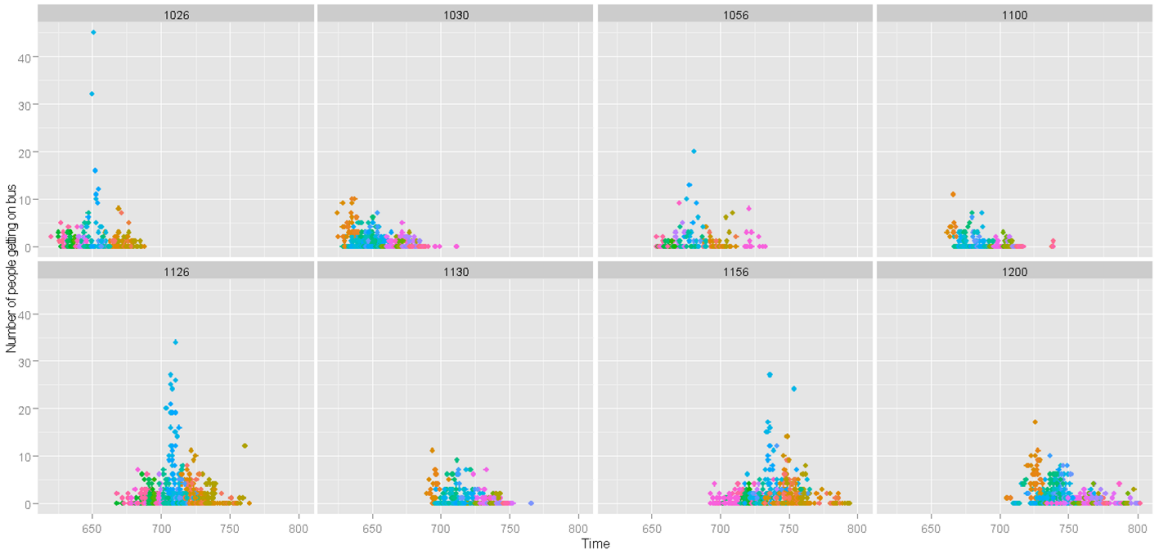

Figure 1: Number of people getting on the bus for different bus departure time at each station

Arrival Process

In the model, we need to estimate the inter-person arrival time in minutes. We assume that the arrival process to be a Poission process, so we can estimate the arrival rate λ from data. Poisson

process is a renewal process in which the inter-arrival interval have an exponential distribution function $f_X = \lambda \exp(-\lambda x)$.

Estimate λ

For each station and each time interval, the data tell us the number of people getting on the bus at some time. We assume that all the people get on the bus if bus comes, then we can use maximum likelihood estimator to derive $\hat{\lambda}$.

By the theory of Poission process, the number of people arrive in time period t is n, the probability density will be:

$$P_{N(t)(n)} = \frac{(\lambda n)^n \exp(-\lambda t)}{n!}$$

From the data, we get a series of n_i and time interval t_i, we can write the likelihood as:

$$\text{Likelihood} = \prod_{i=1}^{n} \frac{(\lambda t_i)^{n_i} \exp(-\lambda t_i)}{n_i!}$$

So the log-likelihood can be written as:

$$\text{Log-likelihood} = \sum_{i=1}^{n} n_i \log(\lambda t_i) - \lambda t_i - \log(n_i!)$$

We want to find the best λ to make the log-likelihood largest, so

$$\max_{\lambda} \text{Log-likelihood} = \max_{\lambda} \sum_{i=1}^{n} n_i \log(\lambda t_i) - \lambda t_i$$

The function is convex, so we can take the derivative and make it zero to maximize it.

$$\frac{\partial \text{Log-likelihood}}{\partial \lambda} = \sum_{i=1}^{n} \frac{t_i}{\lambda t_i} - t_i = 0$$

So

$$\hat{\lambda}_{MLE} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} t_i}$$

So we can estimate the λ at each station by this estimator.

Expected Number of People getting on the bus

For each station and each time interval, we want to estimate the expected number of people getting on the bus.

The density function we use is still $P_{N(t)(n)}$, then the expectation can be represented as:

$$\begin{aligned} E(N) &= \sum_{n=1}^{\infty} \frac{(\lambda t)^n \exp(-\lambda t)}{n!} n \\ &= (\sum_{n=1}^{\infty} \frac{(\lambda t)^n}{(n-1)!}) \exp(-\lambda t) \\ &= (\lambda t)(\sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!}) \exp(-\lambda t) \\ &= \lambda t \exp(\lambda t) \exp(-\lambda t) \\ &= \lambda t \end{aligned}$$

Note that $\sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} = exp(\lambda t)$ by Taylor expansion.

Expected Total Waiting Time

There two ways to calculate the expected total waiting time at each station for each bus interval.

Before that, we take some notation to make our derivation easier.

i X_i is the time length between $(i-1)$th and ith arrival.

ii S_i is the time length between $t = 0$ and ith arrival

$$X_i = S_i - S_{i-1}$$

iii We want to calculate the total waiting time, but in this chapter, we focus on calculating $\sum_{i=1}^{N(t)} S_i$, we can derive total waiting time easily from it.

$$\text{Total waiting time} = N(t)t - \sum_{i=1}^{N(t)} S_i$$

By Intuition

The first idea comes from the basic intuition that if the expected inter-arrival time is $\frac{1}{\lambda}$, we can assume that for every $\frac{1}{\lambda}$ minutes, one person will come. By this assumption, the average arriving time will be $\frac{N(t)}{2}\frac{1}{\lambda}$, then we can calculate the $\sum_{i=1}^{N(t)} S_i$.

$$\sum_{i=1}^{N(t)} S_i = N(t)\frac{N(t)}{2}\frac{1}{\lambda} = \frac{\lambda t^2}{2}$$

By Taking Expectation

The assumption above is too naive, so we want to find a more general way to calculate the total waiting time.

From the Poission process theory, let $N(t) = n$, the marginal distribution of $\{X_i\}_{i=1}^n$ is:

$$p(X_1, \cdots, X_n) = \lambda^n \exp(-\lambda X_1 - \cdots - \lambda X_n)$$

Since

$$\sum_{i=1}^{N(t)} S_i = \sum_{i=1}^{n}(n+1-i)X_i$$

So

$$E(\sum_{i=1}^{N(t)} S_i) = \int_0^\infty \cdots \int_0^\infty \sum_{i=1}^{n}(n+1-i)X_i\lambda^n \exp(-\lambda X_1 - \cdots - \lambda X_n)dX_1 \cdots dX_n$$

This is a very complicated integral, it is a sum of n components, for X_i:

$$\begin{aligned}\int_0^\infty X_i \exp(-\lambda X_1 - \cdots - \lambda X_n)dX_i &= \frac{1}{\lambda^2}\exp(-\lambda X_1 - \cdots - \lambda X_n)|_0^\infty \\ &= \frac{1}{\lambda^2}\exp(-\lambda X_1 - \cdots - \lambda X_{i-1} - \lambda X_{i+1} - \cdots - \lambda X_n)\end{aligned}$$

Then we integrate other variable, for example X_1,

$$\begin{aligned}\int_0^\infty \frac{1}{\lambda^2} &\exp(-\lambda X_1 - \cdots - \lambda X_{i-1} - \lambda X_{i+1} - \cdots - \lambda X_n)dX_1 \\ &= \frac{1}{\lambda^2}(-\frac{1}{\lambda}\exp(-\lambda X_1 - \cdots - \lambda X_{i-1} - \lambda X_{i+1} - \cdots - \lambda X_n))|_0^\infty \\ &= \frac{1}{\lambda^3}\exp(-\lambda X_2 - \cdots - \lambda X_{i-1} - \lambda X_{i+1} - \cdots - \lambda X_n)\end{aligned}$$

We put the derivation into original formula:

$$\int_0^\infty \cdots \int_0^\infty \sum_{i=1}^{n}(n+1-i)X_i\lambda^n \exp(-\lambda X_1 - \cdots - \lambda X_n)dX_1 \cdots dX_n = \frac{1}{\lambda}\sum_{i=1}^{n}(n+1-i)$$

So

$$E(\sum_{i=1}^{N(t)} S_i) = \frac{\lambda t^2 +}{2} + t$$

This method are little bigger than the previous one by t.

Results

After running our non-linear model against our data, we found the following results:

Table 1: Optimal Solution for Model 1 (Deterministic Model), 1 hour intervals

Period	Number of Buses to Run	Inter-bus Arrival Time
5:30am-6:30am	1	20 minutes
6:30am-7:30am	1	20 minutes
7:30am-8:30am	2	10 minutes
8:30am-9:30am	2	10 minutes
9:30am-10:30am	2	10 minutes
10:30am-11:30am	2	10 minutes
11:30am-12:30pm	2	10 minutes
12:30pm-1:30pm	2	10 minutes
1:30pm-2:30pm	2	10 minutes
2:30pm-3:30pm	3	6.67 minutes
3:30pm-4:30pm	3	6.67 minutes
4:30pm-5:30pm	3	6.67 minutes
5:30pm-6:30pm	1	20 minutes
6:30pm-7:30pm	1	20 minutes
7:30pm-8:30pm	1	20 minutes
8:30pm-9:30pm	1	20 minutes
9:30pm-10:30pm	1	20 minutes
10:30pm-11:30pm	1	20 minutes

This allocation of buses gives us a final total waiting time of 254,584.638 minutes.

Table 2: Optimal Solution for Model 1 (Deterministic Model), 2 hour intervals

Period	Number of Buses to Run	Inter-bus Arrival Time
5:30am-7:30am	1	20 minutes
7:30am-9:30am	2	10 minutes
9:30am-11:30am	2	10 minutes
11:30am-1:30pm	2	10 minutes
1:30pm-3:30pm	2	10 minutes
3:30pm-5:30pm	3	6.67 minutes
5:30pm-7:30pm	1	20 minutes
7:30pm-9:30pm	1	20 minutes
9:30pm-11:30pm	1	20 minutes

This allocation of buses gives us a final total waiting time of 291,667.6905 minutes.

Table 3: Optimal Solution for Model 2 (Poisson Model), 1 hour intervals

Period	Number of Buses to Run	Inter-bus Arrival Time
5:30am-6:30am	1	20 minutes
6:30am-7:30am	1	20 minutes
7:30am-8:30am	2	10 minutes
8:30am-9:30am	2	10 minutes
9:30am-10:30am	2	10 minutes
10:30am-11:30am	2	10 minutes
11:30am-12:30pm	2	10 minutes
12:30pm-1:30pm	2	10 minutes
1:30pm-2:30pm	2	10 minutes
2:30pm-3:30pm	3	6.67 minutes
3:30pm-4:30pm	3	6.67 minutes
4:30pm-5:30pm	3	6.67 minutes
5:30pm-6:30pm	1	20 minutes
6:30pm-7:30pm	1	20 minutes
7:30pm-8:30pm	1	20 minutes
8:30pm-9:30pm	1	20 minutes
9:30pm-10:30pm	1	20 minutes
10:30pm-11:30pm	1	20 minutes

This allocation of buses gives us a final total waiting time of 239,112.9521 minutes*

*Plus some constant t

Table 4: Optimal Solution for Model 2 (Poisson Model), 2 hour intervals

Period	Number of Buses to Run	Inter-bus Arrival Time
5:30am-7:30am	1	20 minutes
7:30am-9:30am	2	10 minutes
9:30am-11:30am	2	10 minutes
11:30am-1:30pm	2	10 minutes
1:30pm-3:30pm	2	10 minutes
3:30pm-5:30pm	3	6.67 minutes
5:30pm-7:30pm	1	20 minutes
7:30pm-9:30pm	1	20 minutes
9:30pm-11:30pm	1	20 minutes

This allocation of buses gives us a final total waiting time of 275,764.6547 minutes*

*Plus some constant t

Notice that Model 1 and Model 2 have the exact same allocation of buses for both the 1 hour intervals and the 2 hour intervals. The allocation of buses is also essentially the same between the models with 1 hour intervals and the models with 2 hour intervals (the only discrepancy occurs during the period 1:30pm-3:30pm which is the only 2 hour interval being looked at in the 2 hour intervals model where the allocation for the times disagree).

It is also important to realize when comparing the total waiting times, the times given in the poisson model (Model 2) are not necessarily accurate because those times need to be added to some constant t, as explained in the earlier section on the **Expected Total Waiting Time**. As a result, though it may initially appear that the poisson model has shorter expected total waiting times, the two models may in fact be much closer to returning the same result than it appears.

The consistency in our models serves as a check to the validity of both models. If the models would have disagreed, this would mean that the way people arrive (deterministic or random) has an affect on the optimal scheduling. This way, the optimal scheduling appears to be minimally affected by (if at all) the way that people arrive at a stop. We expect the models to give us the same (optimal) result.

While the Inter-bus arrival time may appear to be very frequent despite the low number of buses being run, keep in mind that the model is operating on a simplified route with less stops over a shorter distance. Relative to the average speed of the buses and the distance of the route, these numbers are what we would expect.

As far as the resulting scheduling is concerned, notice how in the middle of the day, more buses are needed (seemingly confirming our initial suspicions that more buses need to be run during rush hour). We did not predict, however, that in the afternoon there would be a need for more frequent buses than in the morning, but evidence for this is clearly shown in the solution. As expected, early in the morning and later in the night, there becomes less of a demand for buses to run as frequently as in the middle of the day.

Limitations

The most obvious limitation of our model is that the line we are considering does not exactly correspond to any of the Port Authority's actual bus lines but rather to the portion of lines 61A, 61B, 61C, and 61D that the four lines have in common (from Fifth & Wood to Forbes & Murray). Therefore, we are unable to determine how to split buses among the four lines or to take into account data from the other halves of the lines. However, this limitation could also be seen as an alternate transportation solution. The demand for buses is certainly very different at each end of these four lines, which run from Downtown Pittsburgh to the suburbs of Braddock or Homestead, so rather than adding more lines in the areas with higher demand, as is currently done, we have determined the optimal supply of buses for the busier half of the area covered and provided a new line that exactly meets that supply.

Similarly to our assumptions, some of our limitations could easily be overcome given updated data. For example, our model would be invalid on exceptional days, such as holidays, when demand for buses would be significantly different than on a typical day, or when there is a detour along the line. However, as our model was purposefully designed to be very generalizable, given updated data on passenger arrival or line length, the model could easily be rerun. Similarly, the constraints

that were estimated from the budget data that we were given could be replaced with more accurate figures.

Conclusions

After running our model against the data, we found that the afternoon period (2:30pm-5:30pm) demands the most buses. As we expected, more buses need to be run during rush hour times to accommodate the many commuters who use the buses to go to work and less buses need to be run in the morning and evening when people are still home.

It is interesting to note that the actual Pittsburgh bus schedule runs buses less frequently in the period between the rush hours and also does not run buses more frequently in the afternoon rush hour than the morning rush hour (http://www.portauthority.org/rt/61a.pdf). This is not supported in our findings, which suggests that either the Pittsburgh Port Authority is scheduling buses in a suboptimal way or that some of our simplifying assumptions actually do affect the resulting allocation. In a future analysis, we could use a much less simplified model to determine which is the case.

Appendix

Code

Visualize Data

```
#wei ma

library(ggplot2)

sheet1<-read.csv("OR2.csv", header=FALSE)
hours<-sheet1[,3]
minutes<-sheet1[,4]
a.time<-hours*60 + minutes
num.gettingOn<-sheet1[,6]
frame.plot <- data.frame(a.time, num.gettingOn, dept = sheet1$V1, station = sheet1$V2 )
frame.plot <- subset (frame.plot, dept < 1220)
frame.plot <- subset (frame.plot, dept > 1000)
attach(frame.plot)
g <- ggplot(data = frame.plot)
g <- g + geom_point(aes(x=a.time, y=num.gettingOn, color = station), scale=0.1) + theme(legend.position="
g <- g + facet_wrap(~dept, nrow=2) + xlab("Time") + ylab("Number of people getting on bus")
```

Calculate λ

```
#Wei Ma
#11/21/2014

library(foreign)
sheet1<-read.csv("OR2.csv", header=F)
hours<-sheet1[,3]
minutes<-sheet1[,4]
a.time<-hours*60 + minutes
num.gettingOn<-sheet1[,6]

sep.interval <- 240 #minute
station.name <- unique(sheet1[,2])
station.num <- length(station.name)

start.hour <- 5
start.min <- 30
start.time <- start.hour * 60 + start.min

end.hour <- 23
end.min <- 59
end.time <- end.hour * 60 +end.min

interval.num <- floor((end.time-start.time)/sep.interval)
lambda.raw <- rep(NA, station.num * interval.num)

for (i in 1:station.num)
{
  station.set <- subset(sheet1,V2==station.name[i])
  for (j in 1: interval.num)
  {
    time.lower <- start.time + sep.interval *(j-1)
    time.upper <- start.time + sep.interval * j
    station.time.set <- subset(station.set, V3*60+V4 <= time.upper)
    station.time.set <- subset(station.time.set, V3*60+V4 >= time.lower)
    #start to calculate lambda
    arrival.num <- length(station.time.set$V1)
    lambda=0
    if(arrival.num >2)
    {
    arrival.time <- station.time.set$V3*60 + station.time.set$V4
    station.time.set$arrival <- arrival.time
```

```
      station.time.set.sort ← station.time.set[order(arrival.time),]
      arrival.time.sort ← station.time.set.sort$arrival
      arrival.interval ← rep(NA, arrival.num−1)
      for(k in 2:arrival.num)
      {
        arrival.interval[k−1] ← arrival.time.sort[k] − arrival.time.sort[k−1]
      }
      arrival.number.sort ← (station.time.set.sort$V6)[2:arrival.num−1]
      lambda ← sum(arrival.number.sort)/sum(arrival.interval)
      }
      print(lambda)
      lambda.raw[j+(i−1)*interval.num] ← lambda
    }
}

#process results
lambda.matrix ← matrix(lambda.raw, nrow = station.num, ncol=interval.num, byrow = TRUE)
lambda ← data.frame(lambda.matrix)
row.names(lambda) ← station.name
write.csv(lambda, file = "a.csv")
```