MTA Train Timetabling

David Bang, Jingya Bi, Tao Cui, and Jorge Solis

Overview - Hudson Line, Harlem Line, New Haven Line

Hudson Line	Harlem Line	New Haven Line
-running north from New York City along the east shore of the Hudson River -Annual ridership: 17 million people -29 Stations -74 Miles	-running north from New York City into eastern Dutchess County -Annual ridership: 27 million people -38 Stations -82 Miles	-running from New Haven alongside Long Island Sound -Annual ridership: 39 million people -30 stations -74 miles

Background

- Based on 2007 MTA origin-destination survey
- Considered factors:
  - total passenger volume
  - boarding stations
  - travel times
  - train capacity
- Focus on only AM peak hours

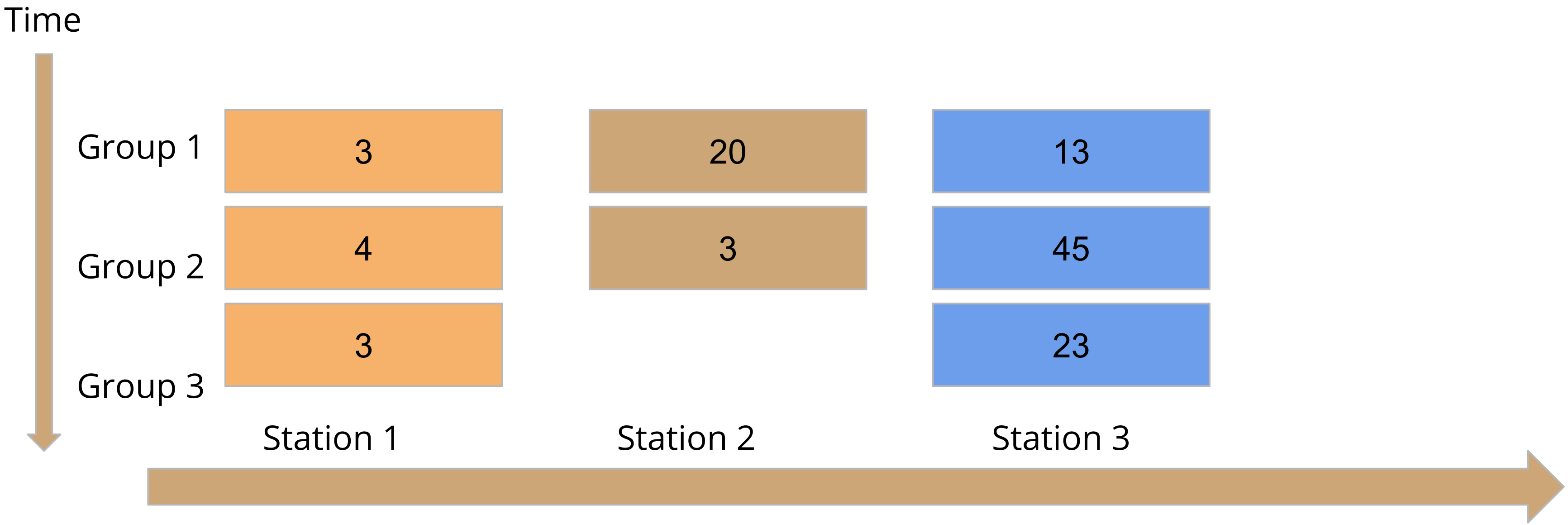
Assumptions and Problem Formulation

- Deterministic arrival times of passengers on each station
- No delays between trains
- Due dates = arrival time at a station + train travel time
- Waiting time to board the train = Tardiness of the passengers
- Minimize the cost of operation and tardiness

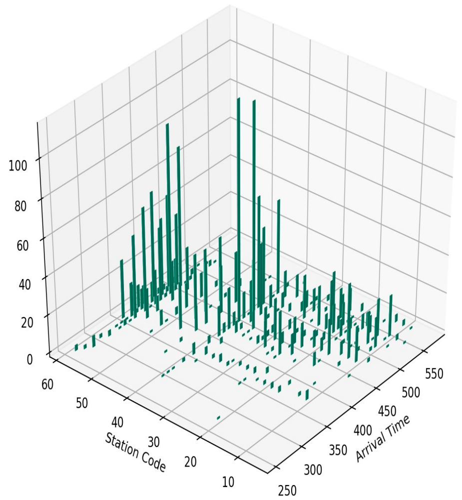
Group Formulation

-Given desired bucket length, we put arrival times into associated buckets

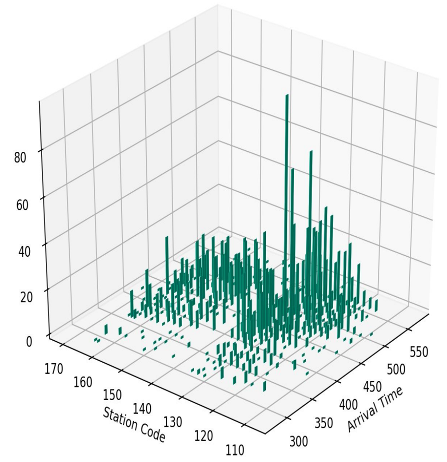
-Each bucket forms a group at each station



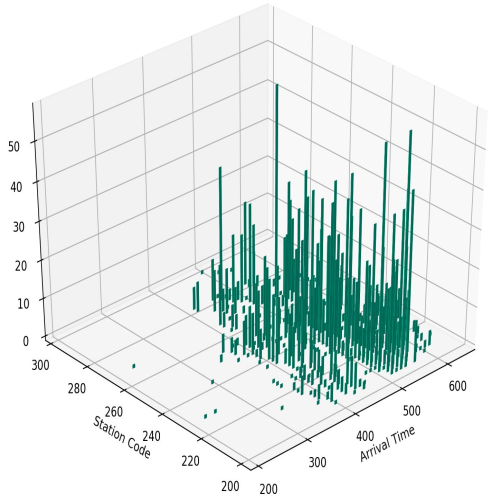
Passengers Arrival Plot



Hudson Line



Harlem Line



New Haven Line

Mixed Integer Programming Formulation

$$\min A \sum_{0}^{T} d_t + Z$$

$$\text{st} \sum_{t=a_{j,k}}^{T} x_{j,k,t} = 1 \ \forall j,k$$

$$x_{j,k,t} \le y_{j,k,t} \ \forall j,k,t$$

$$y_{j,k,t} \le d_t \ \forall j,k,t$$

$$y_{j,k,t} * (t + q_j - a_{j,k}) \le Z \ \forall j,k$$

$$\sum_{j,k} (x_{j,k,t} * p_{j,k}) \le C \ \forall t$$

$$d_m + d_{m+i} \le 1 \ \forall i \ 1 \le i \le \Delta t$$

$$x_{j,k,t} * (t + q_j - a_{j,k}) \ge 0$$

$$x_{j,k,t} \in [0,1]$$

$$y_{j,k,t} \in \{0,1\}$$

$$d_t \in \{0,1\}$$

The constants are:
$a_{j,k}$: the arrival time of group (j,k) at station j,
$p_{j,k}$: the size of the kth group of passengers that arrives at station j,
q_j: the interstation time between the first station to station j.
A: the cost parameter of running an additional train in comparison to a unit of additional max tardiness,
C: the capacity of each train.
The decision variables are:
d_t: the binary indicator of whether a train departs station 1 at time t,
$x_{j,k,t}$: the fraction of passenger group $p_{j,k}$ that boards the train that departs station 1 at time t,
$y_{j,k,t}$: the binary indicator of whether any passenger from group (j,k) boards the train that departs station 1 at time t.
Z is the tardiness of a group.

Need for Another Algorithm

- Only worked on examples with small number of variables
    - t<=30, #station<=10, #group <=20
- Actual data is too large to run
    - t ≈ 240 minutes, #station ≈ 25, #group ≈ 30 (for morning hours approximately)
- Requires baseline solutions and approximate solutions

Dynamic Programming Formulation

Time interval(Δt) is 3

$$f(10) = \min\{f(7)+A+\max(3a_7,T'_7),\ f(8)+\max(2a_8,T'_8),\ f(9)+\max(1a_9,T'_9)\}$$

a_t: Indicator of whether there are any people left over at time t
T'_t: Maximum tardiness of people who arrive between time t and time 10

Sample Data

HARLEM_dp_data

	stationCode	groupSize	arrivalTime
0	134	1	38
1	134	1	208
2	134	1	238
3	134	4	258
4	134	2	278
5	134	28	298
6	134	8	308
7	134	29	318
8	134	10	328
9	134	29	338
10	134	5	348
11	134	1	358
12	134	20	368
13	134	1	378
14	134	10	398
15	134	2	408
16	134	12	418
17	134	1	428
18	134	17	438
19	134	1	448
20	150	1	78
21	150	1	228

Run DP on Real Data

Stack stations, so that DP only takes in an array of number of people arriving, indexed by arrival time.

For example:

t	p_t
0	10
1	12
2	0
3	0
4	0
5	5

Hudson Line Schedule

Train	Departs POUGHKEEPSIE	Notes	Arrives GRAND CENTRAL	Notes	Travel Time In Minutes	Transfer(s)	Fares
802	4:12 AM		5:53 AM		101	THROUGH TRAIN	OFF-PEAK
806	4:42 AM		6:23 AM		101	THROUGH TRAIN	PEAK
810	5:12 AM		6:53 AM		101	THROUGH TRAIN	PEAK
814	5:32 AM		7:15 AM		103	THROUGH TRAIN	PEAK
816	5:56 AM		7:43 AM		107	THROUGH TRAIN	PEAK
818	6:13 AM		7:48 AM		95	THROUGH TRAIN	PEAK
830	6:21 AM		8:07 AM		106	THROUGH TRAIN	PEAK
832	6:39 AM		8:16 AM		97	THROUGH TRAIN	PEAK
834	6:46 AM		8:32 AM		106	THROUGH TRAIN	PEAK
836	7:02 AM		8:40 AM		98	THROUGH TRAIN	PEAK
838	7:09 AM		8:54 AM		105	THROUGH TRAIN	PEAK
840	7:32 AM		9:20 AM		108	THROUGH TRAIN	PEAK
842	8:01 AM		9:43 AM		102	THROUGH TRAIN	PEAK
844	8:28 AM		10:15 AM		107	THROUGH TRAIN	OFF-PEAK
846	8:55 AM		10:47 AM		112	THROUGH TRAIN	OFF-PEAK
850	9:45 AM		11:37 AM		112	THROUGH TRAIN	OFF-PEAK

DP solution:

4:14
4:25
4:44
4:56
5:07
5:18
5:29
5:40
5:51
6:02
6:13
6:24
6:44
7:04
7:24
7:44
8:04
8:23
8:43

Objective: 200
Tardiness: 10

Interval: 10 mins
Capacity: 110
Train Cost: 10

Harlem Line

*If **bold** letter appears in Note column, click on it for details.*

Train	Departs DOVER PLAINS	Notes	Arrives GRAND CENTRAL	Notes	Travel Time In Minutes	Transfer(s)	Fares
908	5:18 AM		7:13 AM		115	THROUGH TRAIN	PEAK
916	5:53 AM		7:53 AM		120	THROUGH TRAIN	PEAK
920	6:36 AM		8:30 AM		114	THROUGH TRAIN	PEAK
926	6:57 AM		8:56 AM		119	THROUGH TRAIN	PEAK

DP Solution:

4:14	5:18	6:13	7:44
4:25	5:29	6:24	8:04
4:44	5:40	6:44	8:23
4:56	5:51	7:04	8:43
5:07	6:02	7:24	

Tardiness: 10
Objective: 143

Interval: 10 mins
Capacity: 110
Train Cost: 10

Comparison of MIP and DP on Simulated Data

Normalize passenger arrivals:

1. Offset arrivals backwards in time by cumulative travel time.
2. For each point in time, sum passengers across all stations in line.

- Three stations.
- Traveling between two consecutive stations takes 2 units of time.

t	p_{t1}	p_{t2}	p_{t3}
0	8	0	2
1	0	1	1
2	2	3	0
3	0	5	0
4	1	1	1
5	1	3	0

→

p_{t1}'	p_{t2}'	p_{t3}'
8	3	1
0	5	0
2	1	0
0	3	0
1	0	0
1	0	0

→

p_t^*
12
5
3
3
1
1

Comparison of MIP and DP on Simulated Data

Cost: 3
Capacity: 10
Min Departure Interval: 2
Interstation: 2

$$\min A \sum_{0}^{T} d_t + Z$$

MIP:
Objective value: 18
Max tardiness: 9
Train departures: 1, 3, 5

DP:
Objective value: 21
Max tardiness: 12
Train departures: 1, 4, 7

Conclusion:
Our DP algorithm provides us with an approximate solution.

Limitations

- Obsolete data: MTA conducted survey 12 years ago.
- Cost-to-tardiness ratio subject to various factors, internal and external.
- Current technology limits computation of MIP solution.
- DP solution relies on a recursive substructure of problem that may not exist.

Next Steps

- Rigorous numerical simulation of datasets to derive benchmarks for MIP and DP solutions.
- Estimation of solutions to perturbations of dataset to derive sensitivity to input.
- Development of more efficient models to improve computational tractability.

References

[1] http://web.mta.info/mta/network.htm
[2] http://web.mta.info/mta/news/books/docs/2017%20MNR%20Ridership%20Appendix.pdf
[3] http://web.mta.info/mta/planning/data/MTA_ MNR-Survey-Final-Report.pdf